The Role of Context in Image Annotation and Recommendation

Philip James McParlane

School of Computing

University of Glasgow

This dissertation is submitted for the degree of

Doctor of Philosophy (Ph.D)

University of Glasgow 2016

To my parents.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work, under the supervision of Joemon M. Jose & Simon Rogers, and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. Permission to copy without fee all or part of this thesis is granted provided that the copies are not made or distributed for commercial purposes, and that the name of the author, the title of the thesis and date of submission are clearly visible on the copy.

Philip James McParlane
2016

Acknowledgements

I have been especially looking forward to write this part of my thesis as (a) it is the last page I will write, and (b) it is the only page many people will read (i.e. my family). I would firstly like to thank my colleagues, in particular, my supervisor Joemon Jose for your endless help, supervision and for giving me this amazing opportunity. I would also like to thank my second supervisor, Simon Rogers, for your feedback with respect to my yearly reports and vivas. Secondly I would like to thank those in my office for making the experience so enjoyable: Adam, Fajie, Felix, James, Jesus, Kim, Rami & Stewart. Thanks for the laughs and making submission deadlines bearable! I would like to say a *huge thanks* to Yashar for teaching me how to write an academic paper and for all your guidance - I really could never have finished this without you. As for my viva, I would like to thank my examiners, Frank Hopfgartner and Paul Clough, for their extremely detailed revisions which hugely contributed to the final version of this thesis. Finally, I would also like to thank my funding body, the LiMoSINe project[1], for allowing me to undertake this life long passion.

Most importantly, I would also like to thank my friends and family for their continued support over the years. In particular, I want to say a huge "thank you" to my parents for giving me such a great start in life; I am forever grateful for this and hope to repay the favour at some point in the near future. In addition to my birth parents, I'd like to thank my now parents-in-law (!), Anne and Alan, for their support through this long process. Also I would like to thank my brothers (Andrew and Lewis) and send a special commemoration to my Gran who unfortunately didn't see me finish this PhD, but would have been pleased to get a mention in an academic publication. For keeping me sane through extensive Fifa therapy, I'd like to thank Calum & Kieran. Finally, and most importantly, I would like to say a huge thank you to my ~~fiancé~~ wife, Juliet, for your endless love and support in everything I do. Particularly, thank you for being so open minded, especially with respect to my travels and internships; I really really appreciate it - **we** did this.

[1]This research in this PhD was supported by the the European Community's FP7 Programme under grant agreements nr 288024 (LiMoSINe)

Abstract

With the rise of smart phones, lifelogging devices (e.g. Google Glass) and popularity of image sharing websites (e.g. Flickr), users are capturing and sharing every aspect of their life online producing a wealth of visual content. Of these uploaded images, the majority are poorly annotated or exist in complete semantic isolation making the process of building *retrieval systems* difficult as one must firstly understand the *meaning* of an image in order to *retrieve* it. To alleviate this problem, many image sharing websites offer manual annotation tools which allow the user to "tag" their photos, however, these techniques are laborious and as a result have been poorly adopted; Sigurbjörnsson and van Zwol (2008) showed that 64% of images uploaded to Flickr are annotated with < 4 tags. Due to this, an entire body of research has focused on the automatic annotation of images (Hanbury, 2008; Smeulders et al., 2000; Zhang et al., 2012a) where one attempts to bridge the semantic gap between an image's appearance and meaning e.g. the objects present. Despite two decades of research the semantic gap still largely exists and as a result automatic annotation models often offer unsatisfactory performance for industrial implementation. Further, these techniques can only annotate what they *see*, thus ignoring the "bigger picture" surrounding an image (e.g. its location, the event, the people present *etc*). Much work has therefore focused on building photo tag recommendation (PTR) methods which aid the user in the annotation process by suggesting tags related to those already present. These works have mainly focused on computing relationships between tags based on historical images e.g. that `NY` and `timessquare` co-exist in many images and are therefore highly correlated. However, tags are inherently noisy, sparse and ill-defined often resulting in poor PTR accuracy e.g. does `NY` refer to *New York* or *New Year*? This thesis proposes the exploitation of an image's *context* which, unlike textual evidences, is always present, in order to alleviate this ambiguity in the tag recommendation process. Specifically we exploit the "what, who, where, when and how" of the image capture process in order to complement textual evidences in various photo tag recommendation and retrieval scenarios.

In part II, we combine text, *content-based* (e.g. # of faces present) and *contextual* (e.g. day-of-the-week taken) signals for tag recommendation purposes, achieving up to a 75% improvement to precision@5 in comparison to a text-only TF-IDF baseline. We then consider *external* knowledge sources (i.e. Wikipedia & Twitter) as an alternative to (slower moving) Flickr in order to build recommendation models on, showing that similar accuracy could be achieved on these faster-moving, yet entirely textual, data-

sets. In part II, we also highlight the merits of *diversifying* tag recommendation lists before discussing at length various problems with existing automatic image annotation and photo tag recommendation evaluation collections.

In part III, we propose three new image retrieval scenarios, namely *"visual event summarisation"*, *"image popularity prediction"* and *"lifelogg summarisation"*. In the first scenario, we attempt to produce a rank of relevant and diverse images for various news events by (i) removing irrelevant images such memes and visual duplicates (ii) before semantically clustering images based on the tweets in which they were originally posted. Using this approach, we were able to achieve over 50% precision for images in the top 5 ranks. In the second retrieval scenario, we show that by combining contextual and content-based features from images, we are able to predict if it will become "popular" (or not) with 74% accuracy, using an SVM classifier. Finally, in chapter 9 we employ blur detection and perceptual-hash clustering in order to remove noisy images from lifelogs, before combining visual and geo-temporal signals in order to capture a user's "key moments" within their day. We believe that the results of this thesis show an important step towards building effective image retrieval models when there lacks sufficient textual content (i.e. a cold start).

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Introduction

1.1 Introduction

For millennia man has sought organisation; from libraries to phone books, organising content in order to improve retrievability has been an inherent motivation of man. In the modern age, this motivation has shifted to the organisation of the internet. Since the first internet search engine, JumpStation, was devised at Stirling University (Scotland), an entire research area focusing on *web search systems* has reshaped the way we categorise and find information online.

In the 1970s, early information retrieval research (van Rijsbergen, 1979) focused primarily on building off-line systems which could search for relevant *textual* documents in small digital libraries (Cleverdon, 1967) and by 1992 the US Government began funding the Text Retrieval Conference (TREC) (Harman, 1992) which aimed to further advance text retrieval technologies. More recently, research focus has shifted to building large scale retrieval systems which can be used to find relevant web pages on the internet for a textual query (Baeza-Yates et al., 1999; Wang et al., 2010a). On this, much success has been achieved where search engines, such as Google, achieve almost prefect accuracy for popular queries (Vaughan, 2004).

However, the internet is a heterogeneous network which contains more than text-based web documents; in particular, the number of *images* uploaded everyday has increased dramatically in recent years with the rise in popularity of smart phones and photo sharing websites, such as Flickr. Further, the emergence of new lifelogging devices, such as Google glass, suggests a continuing trend. Building retrieval systems for user images is very difficult, however, as we must first understand the *semantics* of photographs (i.e. the subjects, or objects, within them and their context) before we can match them against a user's search intent. In order to be able to draw some meaning from images, an entire field of research has focused on their automatic annotation (Hanbury, 2008; Smeulders et al., 2000; Zhang et al., 2012a) with many researchers taking part in image annotation benchmarking tracks such as ImageCLEF (Villegas et al., 2015). In general, these works attempt to infer meaning from high level visual features (e.g. colour, texture, shape *etc*) in order to assign images with textual repre-

sentations[1]. Specifically, researchers have attempted to bridge the so called *"semantic gap"*:

> *...between the information that one can extract from the visual data and the interpretation that the same data has for a user in a given situation.*
> - Smeulders et al. (2000), page 5.

Bridging this semantic gap is the overall goal of all automatic image annotation (AIA) approaches, however, creating a system which can identify objects within photographs, much like a human eye/brain, is extremely difficult for various reasons. This is in large part due to the fact that text is man's creation. What we write is easy to learn; what we see is nearly impossible to teach. For example, understanding a book is relatively trivial as there exists some external resource which maps each phrase to its meaning (i.e. a dictionary). On the contrary, there exists no such resource for mapping visual representations to their meaning. In fact, creating such a resource is nearly impossible due to the complexity of images and visual diversity of objects e.g. consider the many different makes, models and colours of a *car*. Further complications arise when the object in question is observed in different conditions (e.g. lighting, weather, angle, orientation *etc*). This lack of definition makes photographs difficult to understand and therefore retrieve.

Fig. 1.1 Automatic image annotation limitations: automatic model may be able to determine `music festival` from the visual appearance of this image, but they will be unable to identify the event i.e. `titp2014`

[1]These textual representations are uni-gram terms (i.e. single words, or concatenated phrases) e.g. `fish`, `musicfestival` *etc*. In this thesis we interchangeably refer to these textual representations as *tags*, *keywords* or *annotations*

Despite almost two decades of research, the semantic gap still largely exists meaning that fully automatic annotation methods are often unreliable for real life applications. Additionally, these methods can only identify objects/scenes within an image, thus ignoring any contextual details which may be important for retrieval purposes. For example, an automatic image annotation method may be able to understand (from the pixels) that Figure 1.1 is of a music festival, but it will unlikely be able to determine the exact festival (i.e. T in the Park 2014); a deeper understanding of the *context* an image is taken in is required to understand this. In our example, we hypothesise that the tag which describes the event, `titp2014`, will be of more interest to a user in a retrieval scenario due to its discriminative power over high level concepts such as `musicfestival`.

In fact the classification of tags themselves, and how users annotate images has been an area of much research in recent years. For example, Hollink et al. (2004) developed a framework, based on the output on an empirical study, in order to categorise image annotations into various high level categories: (i) nonvisual metadata (ii) perceptual descriptions, and (iii) conceptual descriptions. It should be noted that it is only the last classification category which can rely solely on visual appearance; therefore, non-visual metadata and perceptual descriptions will be missed by state-of-the-art annotation models. In a similar work, Jaimes and fu Chang (2000) proposed a 10-level pyramid scheme defining various types of image indexes (e.g. image type, specific concepts contained *etc*) detailing their descriptive power for retrieval applications. For example, the authors define "specific object" tags, such as `Mount Everest`, to have higher *value* for annotation purposes than "generic object" tags, such as `mountain`. These works build upon theoretical studies on "iconology" taken out many years before multimedia retrieval existed, such as those by Panofsky (1972) and Shatford (1986). Panofsky's book concerns the definition of "meaning" within Renaissance artwork; the author states that there exist both factual and expressional meaning within these paintings, for which the latter category varies from person to person based on their perception. This finding echos those in our own work in that there is a distinction between what an AIA model *can*, and *cannot*, identify within an image and thus the semantic gap may never be fully bridged.

Due to this limitation of AIA models, as well as their unreliable performance & computational complexity, image sharing websites have relied on user tagging[2] in order to organise their content. This has many *advantages* over automatic approaches: (i) humans are more accurate than computers in recognising objects within images (ii) humans can identify concepts which a computer cannot (e.g. the people present, the location *etc*) (iii) additionally, crowdsourcing this task can be viewed as a cheaper option for image sharing websites as less servers and staff are required. However, relying on user tagging also introduces a number of *disadvantages*: (i) due to the laboriousness of the task, photographs are often insufficiently tagged, as highlighted by Sigurbjörnsson and van Zwol (2008) where the majority of images on Flickr were shown to be

[2]Allowing the user to annotate images with keywords which best describe the content

annotated with less than four tags (ii) allowing full tagging freedom can lead to problems where users "inject" (irrelevant) popular tags, such as `girl`, in order to "trick" retrieval models into promoting their content in search results, much like the problem of "keyword stuffing" faced by search engines in the early days of the internet (Nathenson, 1998) (iii) as observed in Figure 1.2, humans refer to the same entity or concept using different languages & vocabularies (i.e. `cat`, `kitty`, `kitten`, `kočka`, `chat` or `gatto`?) (iv) also, humans often tag images with emotional or opinionated tags (e.g. `cool`) which are useless, or even decremental, for retrieval purposes as the annotations are subjective and therefore incorrect for some users. These problems make building image retrieval models difficult as photographs will often be poorly annotated and therefore may be omitted from search results, or returned for irrelevant queries.

Photo tag recommendation (PTR) has offered a semi-automatic alternative where new, additional tags are offered based on those already assigned to an image. Adopting these approaches offers an effective compromise where: (i) accuracy is maximised where the user ultimately decides if a tag is relevant or not (ii) relevant *non-visual* concepts, such as the location or event, can be suggested and annotated for an image (iii) annotation labour is minimised and speed maximised (iv) money & resources are saved by image sharing websites as the computational complexity of PTR is less than that of AIA (v) the recommendation vocabulary can be constrained in order to reduce the probability of synonymity & opinionated/emotional tags (vi) finally, users are suggested with tags which they may not have considered adding. These benefits have increased the popularity of PTR systems in recent years with all of the major multimedia sharing websites (e.g. YouTube, Flickr *etc*) offering tag suggestions to the user.

Fig. 1.2 An example Flickr user annotated image (Liu et al., 2010). The author has given us permission to use this image.

Existing work in the area of photo tag recommendation has considered only tags added by the user, however, in order to make new recommendations. This method attempts to build relationships between annotations based on historically tagged images. For example, many images tagged with `ny` will also be tagged with `timessquare`. Therefore, using this strategy, for any new image tagged with `ny`, `timessquare` is suggested to the user. This strategy can often fail, however, due to the ambiguity of tags. For example, does `ny` refer to *New York City* or *New Year*? If the image is taken at `00:15` on `1/1/2014` & a friend's recent image upload contains the annotation `2014`, then perhaps `ny` refers to *new year* and `party` is a more suitable tag recommendation. Further, if the image is taken in Edinburgh (Scotland), then perhaps `Hogmanay` (the Scottish celebration of new year) is a more suitable suggestion. From an information retrieval perspective, we do not even need to have a deep understanding of these temporal, geographic and social contexts (e.g. know the meaning of Hogmanay), but instead we must accept that they exist and exploit their relations from within the underlying dataset.

This thesis attempts to exploit these types of context for both the annotation and retrieval of web images. Specifically we focus on *when*, *where* and *how* photographs are captured and why these aspects can be used to improve photo retrieval applications. In particular we highlight the benefit of exploiting image context in a cold start scenario, where we know nothing about an image, for two annotation tasks and three retrieval scenarios. In the following sections, we further detail these tasks chapter by chapter.

1.2 Research Overview

In the following subsections, we firstly define our high-level research questions (HL-RQ) before briefly discussing the methodology used, results achieved and publications output during this thesis.

1.2.1 High Level Research Questions

In this thesis we aim to address a number of high level research questions (HL-RQ), which are further broken down in the various chapters. These questions are as follows:

HL-RQ1. Can the *context* an image is taken in be exploited for photo tag recommendation purposes in order to complement existing textual evidences? Which contexts are most effective for this task?

HL-RQ2. Can the *context* an image is taken in be exploited for image recommendation and retrieval purposes? How can this context be used to alleviate the problems associated with retrieval on un-annotated images?

1.2.2 Methodology Overview

In the following we briefly summarise at a high-level our methodology used in each chapter.

Part II: Photo Annotation Firstly, we introduce the methodologies used within the first major part of this thesis on *Photo Annotation*.

Chapter 3. In our first work we extract novel features from an image's context (e.g. device type), visual appearance (e.g. dominant colour) and user's context (e.g. gender), before linearly combining all features in order to improve a TF-IDF tag recommendation model which considers only textual evidences.

Chapter 4. We then consider external data-sets, namely Wikipedia and Twitter, as an alternative to Flickr (which is often "out-of'date" with users uploading their images many days after capture) in order to build TF-IDF tag recommendation models.

Chapter 5. We then attempt to overcome the problem of synonymous tag recommendations as made by existing models; in these works, the models often suggest many redundant tags in the top ranks (e.g. `ny`, `newyork`, `newyorkcity` *etc*). To alleviate this problem, we adapt the popular Maximal Marginal Relevance (MMR) diversification model (Carbonell and Goldstein, 1998) for tag recommendation purposes, evaluating on a new crowdsourced image collection where the ground truths are clustered into *topics*, or aspects, for each image.

Chapter 6. After taking out various photo tag recommendation experiments, we then consider the entire evaluation process for AIA and PTR models; specifically, we identify many problems which could give misleading results or make comparative experiments difficult. For example, existing evaluation collections do not make any efforts to remove synonyms within the ground truths meaning that a model could be *penalized* for suggesting `sea` instead of `ocean`.

Part III: Photo Recommendation Secondly, we introduce the methodologies used within the second major part of this thesis on *Photo Recommendation*.

Chapter 7. In the first work of part III, we propose the task of *visual event summarisation* where we attempt to select the most relevant images from social media for various news stories. We achieve this by firstly removing unsuitable content (e.g. memes, screenshots *etc*) using support vector machine (SVM) classifiers trained upon colour and edge histograms, before removing near-duplicate images using a popular image hashing technique. Finally, we rank images by selecting the most popular images from the largest semantic tweet clusters.

Chapter 8. In chapter 8 we attempt to identify images which will become popular in the future based on many context, content and user based features. Specifically,

we attempt to classify, using an SVM classifier, whether an image has high/low comments or page views.

Chapter 9. In our final work, we attempt to reduce the information overload problem associated with lifelogging devices by summarising a user's day in photographs. We achieve this by firstly removing noisy photographs (e.g. blur detection) before selecting the sharpest image from the largest clusters (which are grouped based on GIST (Oliva and Torralba, 2006) and geo-temporal features).

1.2.3 Results Overview

In the following section we briefly summarise our high-level results for each of the discussed chapters.

Part II: Photo Annotation Firstly, we discuss the high level results and contributions from the chapters in the first major part of this thesis on *Photo Annotation*.

Chapter 3. We show that by considering 17 new features we are able to improve tag recommendation performance by up to 75%, for precision@5, in comparison to a text-only TF-IDF approach. We show that time, orientation, high level scene (e.g. city, party, home, food or sports) and dominant colour are the most effective features for tag recommendation purposes and that they are able to reduce the *query ambiguity* faced by recommendation approaches. Additionally we highlight the merit of these features in a cold start scenario where no textual evidences exist.

Chapter 4. We firstly motivate this work by highlighting the problem of images being uploaded to Flickr many days, or even weeks, after they are captured, meaning that models built on these data sets are often "out-of-date" for new and fast moving events. We also highlight that images are mostly annotated with nouns and entities, indicating that recommendation models should try to reflect this in order to achieve highest accuracy. Finally, we show that by combining recommendations from Twitter and Wikipedia we are able to achieve comparable performance to the industry standard recommendation approach used on the actual Flickr website.

Chapter 5. In this chapter we firstly quantify the level of tag synonymity present within Flickr annotations (which posses a problem for both recommendation and evaluation purposes) in a crowdsourced experiment. The results of this experiment indicate that more than half of image annotations are redundant; alternatively, for half of all annotations there exists at least one other synonym in the tag set (e.g. `newyorkcity` and `newyork`). Based on this, we adapt the MRR diversification approach for tag recommendation purposes, achieving a 6+% improvement for the α-nDCG metric in the top 5 & 10 ranks.

Chapter 6. From various analyses we highlight a number of problems with existing AIA & PTR evaluation methodologies. For example, we show that tag synonymity can result in misleading annotation evaluation performance where models may under-perform by up to 15% for the IAPR collection (Grubinger et al., 2006). We also show that models can "overperform" by exploiting the long tail distribution of tags and propose that annotation models should also consider evaluating on a *normalised ground truth* in order to optimize for *visual* annotation performance.

Part III: Photo Recommendation Secondly, we discuss the high level results and contributions from the chapters in the first major part of this thesis on *Photo Recommendation*.

Chapter 7. As this a novel work within a new area, there unfortunately exists no baseline to compare against. Therefore we instead propose and compare the performance of a range of strategies, namely: two image selection approaches, six ranking methods and three presentation systems. From our experiments we are able to retrieve "relevant" images in more than half of the top positions. Furthermore we show that best performance is achieved when images are selected from both the tweets, and the URLs within the tweets, highlighting that these sources are complementary and that they can improve topic coverage in combination. Finally we show that visual event summarisation is most suitable for events which happen *"in the public domain"* (e.g. sports), opposed to those which happen behind closed doors (e.g. politics).

Chapter 8. From our experiments we show that popularity can be most effectively predicted by combining context, content and user based features, where upto 76% accuracy is achieved. We also show that highly viewed images tend to be images of people, especially women, (i.e. `girl, portait, woman`) and that `nature` photographs tend to be the least viewed but most discussed (i.e. # of comments).

Chapter 9. We firstly highlight many problems with images collected using lifelogging devices in a crowdsourced experiment. The results of this show that over 16% of images are "very blurry" with 60% considered noise (i.e. blurry or a visual duplicate). Using our selection and ranking approaches, we are able to improve sharpness, interestingness and visual diversity by 12%, 13% and 40% respectively, in the top ranks.

1.2.4 Publications

During the four year course of this Ph.D. a number of publications were submitted, accepted and presented at a range of multimedia and information retrieval conferences. Below details these accepted submissions, specifically: (i) five full papers (ii) three short papers (iii) one poster, and (iv) one demonstration paper.

Full Papers

1. Philip J. McParlane, Joemon M. Jose (2014); "Picture the scene..." Visually Summarising Social Media Events, *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM) 2014, Shanghai, China.* ACM, New York, NY, USA, pp1459-1468.

2. Philip J. McParlane, Yashar Moshfeghi, Joemon M. Jose (2014); "Nobody comes here anymore, it's too crowded"; Predicting Image Popularity on Flickr, *Proceedings of ACM International Conference on Multimedia Retrieval (ICMR) 2014, Glasgow, Scotland (UK).* ACM New York, NY, USA, pp385-392.

3. Philip J. McParlane, Yashar Moshfeghi, Joemon M. Jose (2014); Collections for Automatic Image Annotation and Photo Tag Recommendation, *Proceedings of ACM International Conference on MultiMedia Modeling (MMM) 2014, Dublin, Ireland.* ACM New York, NY, USA, pp133-145.

4. Philip J. McParlane, Joemon M. Jose (2013); Exploiting Time in Automatic Image Tagging, *Proceedings of European Conference on Information Retrieval (ECIR) 2013, Moscow, Russia.* Springer-Verlag Berlin, Heidelberg, pp520-531.

5. Philip J. McParlane, Stewart Whiting, Joemon M. Jose (2013); Improving Automatic Image Tagging Using Temporal Tag Co-occurrence, *Proceedings of ACM International Conference on MultiMedia Modeling (MMM) 2013, Huangshan, China.* Volume 7733 of the series Lecture Notes in Computer Science, pp251-262.

Short Papers

1. Philip J. McParlane, Joemon M. Jose (2014); Exploiting Twitter and Wikipedia for the Annotation of Event Images, *Proceedings of ACM Special Interest Group on Information Retrieval (SIGIR) 2014, Gold Coast, Australia.* ACM New York, NY, USA, pp1175-1178.

2. Philip J. McParlane, Yashar Moshfeghi, Joemon M. Jose (2013); On contextual photo tag recommendation, *Proceedings of ACM Special Interest Group on Information Retrieval (SIGIR) 2013, Dublin, Ireland.* ACM New York, NY, USA, pp965-968.

3. Soumyadeb Chowdhury, Philip J. McParlane, Md. Sadek Ferdous, Joemon M. Jose (2015); "My Day in Review": Visually Summarising Noisy Lifelog Data, *Proceedings of ACM International Conference on Multimedia Retrieval (ICMR) 2015, Shanghai, China.* ACM New York, NY, USA, pp607-610.

Posters

1. Philip J. McParlane, Yelena Mejova, Ingmar Weber (2013); Detecting Friday Night Party Photos: Semantics for Tag Recommendation, *Proceedings of European Conference on Information Retrieval (ECIR) 2013, Moscow, Russia.* Springer-Verlag Berlin, Heidelberg, pp756-759.

Demonstrations

1. Philip J. McParlane, Joemon M. Jose (2014); A Novel System for the Semi Automatic Annotation of Event Images, *Proceedings of ACM Special Interest Group on Information Retrieval (SIGIR) 2014, Gold Coast, Australia.* ACM New York, NY, USA, pp1269-1270.

1.3 Outline

This thesis is split into four parts and 13 chapters as follows:

Part I **Background and motivation:** Chapter 2 details a background on image annotation and retrieval. Firstly, we discuss the more general area of information retrieval before describing models and evaluation techniques for image annotation with a focus on photo tag recommendation. Focus then shifts to the task of image retrieval, where we firstly detail content based image retrieval (CBIR) approaches from the early 1990's to web scale retrieval of the modern age. Based on a general overview of these areas and focus on prominent works in image annotation and retrieval we motivate the need for exploiting context for our purposes.

Part II **Photo annotation:** This section discusses our works in *photo tag recommendation*. In Chapter 3, we propose the exploitation of *internal* evidences, gathered from the image itself (e.g. time taken, orientation, camera type *etc*), in the photo tag recommendation task. In Chapter 4, we propose the exploitation of *external* evidences, gathered from some resource related to an image (e.g. social media *etc*), for PTR. In Chapter 5 we propose the diversification of tag recommendation lists due to the significant number of synonyms suggested by existing state-of-the-art recommendation models. Finally, after spending much time executing the work discussed in the previous chapters, we identified a number of evaluation problems in PTR and AIA as highlighted in Chapter 6.

Part III **Photo recommendation:** This section discusses our works in *photo recommendation*. In Chapter 7 we propose the task of visual event summarisation which attempts to gather the most relevant images related to an event for summation purposes. In Chapter 8, we attempt to predict if an image will become popular based on the features proposed in Part II. Finally, in Chapter 9 we consider the

growing area of lifelogging by proposing an approach which is able to offer the user a succinct visual summary of their day.

Part IV **Conclusion:** In Chapter 10 we summarise the key results and insights identified by the work taken out over the course of this thesis before proposing future directions in Chapter 11.

Chapter 2

Background and Motivation

2.1 Introduction

This chapter gives a background and introduction to the high level research areas covered in this thesis, namely: (i) automatic image annotation (ii) photo tag recommendation (iii) photo recommendation and retrieval. This thesis goes far beyond these high level topics, however, also covering many sub-areas such as: (i) image annotation evaluation (ii) diversification (iii) visual event summarisation (iv) photo popularity prediction (v) lifelog summarisation. Due to the specific nature of these sub-topics, we cover these in detail in their respective chapter. This chapter instead gives a high level introduction to the areas of information retrieval and multimedia retrieval which are relevant to the various works described in this thesis.

Firstly, we detail a high level background on information retrieval (IR) in Section 2.2 before discussing works in the automatic, and semi-automatic, annotation of images in Section 2.3. Section 2.4 introduces the various image retrieval & recommendation paradigms proposed over the past two decades, from content based image retrieval (CBIR) to object recognition. In order to build this background literature chapter, we predominantly used Google Scholar[1] to search for prominent works due to its superior size as well as retrieval & filtering capabilities in comparison to other academic databases (Falagas et al., 2008). Specifically, for each section we made various generic queries (e.g. "automatic image annotation") and used both the time filtering functionalities to find the most recent works as well as the "cited by" links under *prominent works* in order to find relevant related papers. By undertaking this process, we believe we have created a complete and unbias background literature chapter for the various topics relevant to this thesis. Finally, in section 2.6 we summarise this chapter.

[1] https://scholar.google.co.uk/ - last accessed on 18th July 2016.

2.2 Information Retrieval

Firstly we discuss the problem of information retrieval (IR) detailing some of the popular concepts which are used throughout this thesis. In traditional information retrieval, the goal is to return a ranked list (ordered by descending relevance) of documents given some textual query. In order to achieve this, one must be able to measure the similarity, or distance, of a given query and document. In the following, we detail document & query representation as well as result ranking & presentation before discussing recommendation techniques and evaluation methodology in IR.

2.2.1 Document Representation

In order to build effective retrieval systems, a retrieval model must firstly be able to effectively and compactly capture the semantic meaning of a document. Since the early days of information retrieval many researchers have employed a high dimensional *vector based representation* for this purpose (van Rijsbergen, 1979). The model represents a text document (or any object) as a vector of identifiers where each dimension corresponds to some term:

$$d_j = \Big[f(t_1, j) \, , f(t_2, j) \, \ldots \, f(t_y, j) \Big] \tag{2.1}$$

where d_j is the vectorial representation of the j-th document in the collection and $f(t_i, j)$ is a function which computes some "score" for the i-th term t_i in a vocabulary of size y. This approach is commonly referred to as the Bag-of-Words (BOW) model as it represents each document as a bag of terms, ignoring any term ordering semantics. In the simplest case of the vector space model (i.e. the boolean model), $f(t_i, j)$ simply denotes the presence (i.e. 1) or absence (i.e. 0) for term t_i in the document d_j. A more powerful model instead counts the term frequency (TF) of a given word in a document.

Using a model which simply captures the presence or frequency of a given term, however, fails to consider the varying semantic power of words and phrases. For example, consider the varying *semantic values* of the terms `olympics` and `is` for retrieval purposes. If a document contains the term `olympics`, we can infer the document is likely relevant (or at least partially) for the topic: Olympic Games. On the contrary, due to the popularity of `is` in the English language (and subsequently web documents), we can infer no semantics for such a document containing this term. Therefore, we should treat the presence of `olympics` with a higher "importance", or weighting, than `is`. This varying discriminatory function of terms has resulted an entire body of research on the topic of weighting schemes, which attempt to compute how "important" a word is to a given collection. Arguably most popular of these works is the *inverse document frequency* (IDF) model:

$$IDF(t_i) = log(m/m^{(t_i)}) \tag{2.2}$$

where m is the number of documents in the collection and $m^{(t_i)}$ is the number of documents containing term t_i. By computing each term's *IDF* score, we can promote, or demote, the significance of a term in describing a document (e.g. *stop words* such as `is` generate a low IDF score due to the the large value for $m^{(t_i)}$). This value is multiplied with the term's document frequency (i.e. TF) to give the TF-IDF vectorial model of document representation, as used throughout information retrieval literature and this thesis.

2.2.2 Query Representation

Aside from capturing the semantics of a document, one must also capture the query intent of the user in order to facilitate retrieval. Vectors can also be employed for this purpose. One of the main differences between documents and queries, however, is their size: documents tend to contain hundreds, if not thousands, of terms whereas the average query length is less than 3 keywords (Lau and Horvitz, 1999). Aside from this length mismatch, it is assumed that users do not formulate queries using the most effective terms; therefore, in order to gain more insight into what the user is actually searching for, many works have proposed *query expansion & reformulation* techniques (Spink et al., 2001) where the aim is to elaborate, or better define, the query by adding, or modifying, search terms. Using these techniques many different modifications are automatically made, such as: (i) adding pluralised forms of root keywords (e.g. `dog` + `dogs`) (ii) adding synonyms (e.g. `uk` + `united kingdom`) (iii) expanding acronyms (e.g. `NASA` + `National Aeronautics and Space Administration`) (iv) correcting spelling errors (e.g. `Sweeden` becomes `Sweden`) *etc*. (v) adding additional terms from relevant documents using a technique called pseduo relevance feedback (Lee et al., 2008).

By automatically reformulating the search terms without the user's knowledge the query is more elaborately defined thus reducing the ambiguity and ultimately increasing the retrieval accuracy; increasing query length has been shown to correlate with better retrieval performance (Kekäläinen and Järvelin, 1998). Therefore, query expansion is a crucial step for retrieval purposes as the majority of queries contain only a few terms (Jansen et al., 1998). As before, an overview of query formulation and expansion is far beyond the scope of this thesis; we refer to the book by Manning et al. (2008) for more information. The crucial message is that documents, which can take many forms (e.g. webpages, images, tags *etc*), can be matched using a vector space model against some query, which can *also* take many forms.

2.2.3 Ranking Results

Based on the notion that we are able to represent a document & user query in vector space (i.e. a bag of words), we are now able to produce ranked retrieval lists based on the user's *information need* (Belkin and Croft, 1992). Two of the major approaches of

achieving this are through: (i) *vector space* matching (ii) *probabilistic based* ranking.

In the *vector space model*, documents are ranked according to their *similarity* to the query using "matching" methods, such as cosine similarity. Therefore, given a query and a set of documents, search results can be ranked based on their ascending distance from the query. This notion that documents similar to queries will match user information needs, however, does not always hold as highly *similar* documents may be highly *irrelevant*; as a result probabilistic methods have been proposed to alleviate this issue.

Probabilistic models, as proposed in the 1970's (Robertson and Jones, 1976), instead predict the *relevance* of a document given a set of query terms q i.e. $P(d_j \mid q)$. Therefore, this formalisation presents a more accurate representation of the user's information need (i.e. retrieving relevant documents) and is often referred to as the *Probability Ranking Principle* (PRP):

> If the retrieved documents are ranked decreasingly on their probability of relevance (w.r.t a query), then the effectiveness of the system will be the best that is obtainable - van Rijsbergen (1979), page 88.

In order to probabilistically rank documents for a query, many ranking functions have been proposed. The first and simplest approach, called the "Binary Independence Model" (Robertson and Sparck Jones, 1988), uses boolean weighted vectors denoting the presence of terms in documents and applies Naive Bayes theory in order to compute the probability of a document's relevance for a given query; the *naive* assumption of the Naive Bayes classifier treats each term independently and therefore no term associations are modelled. More complex models, such as Okapi BM25, employ more elaborate IDF type weighting schemes as well as consider additional components such as "the average document length" *etc*. A full review of vector space and probabilistic approaches is far beyond this thesis and therefore we refer the reader to more complete sources, such as those by Manning et al. (2008) and Baeza-Yates et al. (1999), for more information.

Ranking results is a complex, and ongoing research problem (Chapelle and Chang, 2011; Duhan et al., 2009), however, which must consider various complicated issues such as: (i) language ambiguity and synonymity (e.g. does `Jaguar` refer to the car or the animal?) (ii) user intent (i.e. what is the user trying to do? Gain information on a topic? Make a purchase? *etc*) (iii) query context (e.g. is the query location sensitive?) *etc*.

In order to alleviate the problems of language ambiguity, research has focussed on promoting *diverse* results lists which have a wider coverage of the query sub-topics. For example, `Java` is an ambiguous query since it has different interpretations e.g. the programming language, the island, and the coffee; therefore, ranking methods now attempt to maximise the diversity (as well as relevance) of the documents in the top ranks e.g. returning documents on the all three aspects related to `Java`. In order to gain better inside into the the user's intent, modern search engines build personalised

models (Salton and Buckley, 1997) of the user based on their historical interactions which take two forms: (i) *explicit feedback:* interactions which are knowingly given by the user and often require some level of effort e.g. queries (ii) *implicit feedback*: knowledge which is inferred from user behaviour e.g. mouse clicks. Retrieval approaches may consider a wide range of aspects (Bai et al., 2007) in order to model the query's context, such as: (i) the time (e.g. time of the day, season *etc*) (ii) the user's location (iii) device (e.g. laptop vs mobile) *etc*.

In reality, search engines consider a (weighted) combination of all of these factors, plus many more in order to create a ranked results list. Even after ranking has completed, search engines often re-rank the results using *pseudo relevance feedback* techniques (Lee et al., 2008). This approach aims to improve the results list by assuming that the top k ranks contain relevant documents allowing for further query refinement *etc*. In this thesis we instead focus on ranking *tags* for query images in a tag recommendation paradigm as well as ranking *images* against various user and textual query types.

2.2.4 Result Presentation

Finally, once documents have been ranked in order of relevance by some ranking model with respect to a given query, the search engine, or information retrieval application, has to present these to the user in a concise and helpful manner. For traditional web search, almost all search engines contain certain aspects within the results presentation, such as: (i) *ranked documents:* obviously the most important aspect of the results page is the ranked list of documents. (ii) Paid advertisements: aside from the ranked list of "organic" web pages (i.e. those documents which have risen to the top of the ranks based on their relevance to the query), search engines also include relevant paid adverts which are usually visually separated by the organic list (iii) *text snippets:* below the title of each result exists a query biased textual snippet (Tombros and Sanderson, 1998; White et al., 2003); a sentence or 2 from the document which has been selected based on the similarity to the user's query. Additionally, keywords (or synonyms) from the user's query are often highlighted (e.g. bold text) in order to highlight the document's potential relevance to the query (iv) *aggregated Search:* based on the query type, search engines often include tailored results for domain-specific collections (Arguello et al., 2009; Zhou et al., 2012), called verticals e.g. images, weather, news, maps, entity information *etc*. For some queries, this often enhances the search experience by including high precision query specific information (e.g. showing cinema times for the search query `cinema glasgow`); for some other queries, aggregated search can help alleviate the problem of search ambiguity by displaying results from multiple verticals (e.g. showing a map of Java beside books on the Java programming language) (v) Website structure: aside from displaying a link to the website's main URL, search engines often also include links to pages within the website (e.g. for

the query `celtic fc`[2]: displaying links to *Fixtures*, *Tickets*, *News* etc), allowing for quicker access to relevant information.

Aside from displaying the results list, search engines also offer a number of tools to aid the user in their search process, such as: (i) *related search queries:* allowing the user to select queries with additional relevant terms (e.g. for the query `celtic fc`: (1) `celtic fc tickets` (2) `celtic fc fixtures` *etc*) (ii) *ranked vertical links:* ranking categorical links (e.g. News, Shopping *etc*) based on their relevance to the query (e.g. the shopping vertical should be ranked highly for the query: `nike trainers`).

Commercial search engines constantly change aspects of their interface, however, with web logs and mouse movement constantly analysed in order to give insight into user behaviour and improve the information retrieval process. Aside from web logs, users are constantly taking part in A/B testing experiments without their knowledge in order for the company to gauge the effectiveness of certain features. More recently, experiments monitoring emotion (Arapakis et al., 2008; Moshfeghi and Jose, 2013b), eye movement (Cutrell and Guan, 2007) and brain activity (Moshfeghi and Jose, 2013a) have all been employed in order to give a more detailed insight of the information retrieval process.

2.2.5 Recommender Systems

More recently, much attention has focused on building effective recommender systems where content is instead *suggested* to the user (e.g. product suggestions on shopping websites), without any explicit user interaction (e.g. querying). This focus has been further fuelled by the growth in social media which creates many new scenarios for recommendation purposes. Existing recommendation approaches generally fall into three different categories: (i) Content-based recommendation (Pazzani and Billsus, 2007) (ii) Collaborative filtering (Su and Khoshgoftaar, 2009), and (iii) Hybrid approaches (Burke, 2007).

In *content-based recommendation*, new items are suggested based on a description (or modelled description) of an item and the user's preference. For example, a content-based video recommendation system may suggest the film *Apocalypse Now* for a user if they have already watched many other Marlon Brando films in the past. Firstly, in order to facilitate content based recommendation, models must be able to effectively capture the semantics of an item based on some set of *features* (e.g. the movie year, the movie genre *etc*). Additionally a user profile must also be created capturing their interests (e.g. films they have previous watched or rated). Therefore, given a particular user, items which are *semantically similar* to those which they have a preference for, can be suggested. In this domain, many different approaches have been used to model a user's preference for a given item, such as: decision trees (Cho et al., 2002), relevance feedback (Ahn et al., 2007), probabilistic methods (Semeraro et al., 2009) *etc*. For a

[2] A Scottish football team from Glasgow

full summary of content-based recommendation methods, please refer to the extensive survey by Pazzani and Billsus (2007).

In *collaborative filtering*, new items are instead suggested if they are popular amongst *similar users*. The advantage of this approach is that items do not necessarily need to be analysed (i.e. understood) in order to be recommended to the user. This is particularly useful for situations where it is difficult to automatically extract semantics from the item e.g. videos. Collaborative filtering recommendation systems instead focus on finding similar users, using approaches such as k-nearest neighbour search (Sarwar et al., 2001), based on a number of implicit (e.g. watching a video) and explicit (e.g. "liking" a video) interactions. Therefore, given a list of similar users and the items which they have interacted with (positive or negatively), the collaborative filtering system is able to infer the likelihood of the user finding a new item interesting. One drawback of collaborative filtering, however, is that they often require extensive amounts of knowledge regarding a user (which isn't present in a cold start) in order to make reliable recommendations (Rubens et al., 2011). Collaborative filtering models can be mostly grouped into two different categories: (i) *neighborhood approach*: where methods focus on building matrices which model relationships between items, or users, such as the work by Koren (2009) (ii) *latent-factor models*: where methods instead attempt to project both items and users onto the same latent-factor space, such as the work by Chen et al. (2009). For a full summary of collaborative filtering methods and techniques, please refer to the extensive survey by Su and Khoshgoftaar (2009).

Finally, *hybrid recommendation* models combine both collaborative and content-based methods in order to overcome the limitations of each. Specifically, hybrid approaches can mostly be further categorised into three sub-types: (i) *the combination of predictions made by two separate collaborative and content-based methods*: for example, many early works in this area used linear combination (Claypool et al., 1999) or a voting scheme (Pazzani and Billsus, 2007) to combine the predictions made by these two different models (ii) *the exploitation of content-based features in collaborative models*: for example, Melville et al. (2002) used content-based features to "smooth" their collaborative filtering ratings matrix by filling in its missing values, producing a "pseudo ratings" matrix (iii) *the exploitation of collaborative filtering techniques in content-based models*: for example, Nicholas and Nicholas (1999) use dimensionality reduction techniques to build a collaborative view of a collection of users, where profiles are represented by term vectors related to their interests. For a full summary of hybrid recommendation methods, please refer to the extensive survey by Burke (2007).

We should finally note that recommender systems is a *multidisciplinary field* encompassing techniques not only from information retrieval, but also from machine learning and human-computer interaction (Ricci et al., 2011). Due to the high number of recommender system papers submitted to major information retrieval conferences, such as the ACM Special Interest Group on Information Retrieval (SIGIR)[3], we argue

[3]According to http://dl.acm.org/, there exist 113 SIGIR papers which contain exactly the phrase "recommender system" - last accessed on 18th July 2016.

this field is a subclass of information retrieval. Despite this, with the rise in popularity of the field there now exist many conference dedicated solely to recommender systems, such as the ACM Conference on Recommender Systems (RecSys)[4], thus highlighting the blurred distinction between the two fields (Burke, 2007).

2.2.6 Evaluation Methodology in IR

In order to gauge the effectiveness of a retrieval model, one must be able to benchmark against the information need of the user. In the following subsections, we detail two different benchmarking methodologies used throughout the literature (Clough and Sanderson, 2013): (i) *system-orientated evaluation*, and (ii) *user-orientated evaluation*.

System-orientated Evaluation Traditionally, retrieval models are evaluated using a *system-orientated* approach where a model is assessed based on its effectiveness i.e. how well a system can retrieve relevant documents for a given query. Since the early days of information retrieval (Cleverdon, 1967; Cleverdon et al., 1966) there has been a strong focus on this type of system-orientated evaluation. The key components of this "Cranfield evaluation" experiment are a document collection and a set of topics (or queries) with relevance assessments. These assessments usually denotes the binary relevance of a document to a query, however, some works have also considered using a graded judgement (Kekäläinen and Järvelin, 2002).

Given a set of queries, a model must return a set of ranked documents, ordered in descending relevance, for each query. These documents are then compared against the relevant documents (i.e. also known as *ground truth*) allowing for the computation of various evaluation measures, which attempt to gauge the performance of the model based on the number of "relevant" documents returned in the top ranks. One of the major advantages of this "Cranfield" type evaluation methodology is that its results are reproducible and comparable against other baselines; this evaluation methodology has been the favoured approach for many decades (Clough and Sanderson, 2013) and as a result many evaluation campaigns, such as TREC (Harman, 1992) & NTCIR (Kando et al., 1999), have attempted to create unified test collections in order to allow for comparative system-orientated studies to be undertaken.

In order to measure the effectiveness of models, various evaluation metrics have also been proposed which score the documents returned against those deemed "relevant" by the assessors. In the early days of information retrieval evaluation, *set-based* measures were proposed which simply compared those returned against the relevant documents, thus ignoring the ordering of results. These evaluation metrics are also often measured at some cut-off (N) e.g. P@5 denotes the fraction of documents relevant to a query in the top five positions:

[4]https://recsys.acm.org/ - last accessed on 18th July 2016.

- *Precision (P@N):* denotes the fraction of relevant documents which are retrieved.
- *Recall (R@N):* denotes the fraction of documents which are relevant to the query which are successfully retrieved.

More recently research has focused on *rank-based* metrics which also consider the ordering of results lists, opposed to just considering relevance.

- *Average Precision (AP):* considers the order in which documents are returned by averaging the precision at every cut off, until all documents are retrieved.
- *Mean Average Precision (MAP):* simply takes an average of AP for all queries.
- *Mean Reciprocal Rank (MRR):* is measured as $1/r$, where r is the position of the first relevant document in the retrieved list.

Exhausting the list of metrics is far beyond the scope of this thesis, however; we therefore again refer the author to the book by Manning et al. (2008) for more information. In our work, we adapt these metrics for the purposes of photo tag recommendation and for summarisation evaluation purposes.

User-orientated Evaluation Despite the advantages, some researchers have highlighted problems with system-orientated evaluation experiments, such as its abstraction from reality (Ingwersen and Järvelin, 2005), and as a result the topic of information retrieval evaluation is still an ongoing research area (Harman, 2011; Robertson, 2008). Additionally, with increased research in works which look beyond simple retrieval effectiveness there has also been a focus on *user-orientated* experiments where the end user judges the effectiveness of a given system. There exist many situations where system-orientated experiments are either insufficient or impossible, such as those which consider interface appearance (Pak and Price, 2008), the wider user context (Kelly, 2009) or multi-session evaluations (Kanoulas et al., 2011).

In this line, a technique called A/B testing (also known as split-run testing) has been adopted by many different user-orientated evaluations (Kohavi, 2015). In these tests, users are split into two different segments (i.e. A vs B) where each set is evaluated on a slightly different variant of the experiment (e.g. two different interface styles). Based on this, researchers can evaluate various aspects of their experiment by analysing the different metrics achieved for each user set. For example, Valieri and Marin (2012) employ A/B testing in order to optimise click-through rates in transactional emails.

Unfortunately user-orientated experiments are difficult to reproduce making comparative studies difficult to undertaken. Furthermore, these lab based experiments are often more expensive (in both time and money) than their system-orientated counterparts; crowdsourcing evaluations have in some part been able to alleviate these problems however. Crowdsourcing draws upon large networks of online workers in order

to undertake specific tasks as defined by the user; given its affordability and availability through services such as Amazon MTurk[5] and Crowdflower[6], crowdsourced evaluations have grown in popularity in recent years (Eickhoff and Vries, 2013; Hirth et al., 2010; Zuccon et al., 2013) with system-orientated test collections also obtaining relevance assessments in this way (Carvalho et al., 2011).

In this thesis, we adopt both system-orientated evaluations as well as user-orientated evaluations for tasks where either: (i) there exists no ground truth, or (ii) where the relevance of a document is perceptual e.g. judging the "interestingness" of an image.

2.3 Photo Annotation

Secondly, we discuss the problem of *photo annotation* by giving an overview and history of both fully automated (i.e. AIA) and semi-automated approaches (i.e. PTR). Specifically, we formalise each problem, discuss various features & models used as well as evaluation techniques & baselines adopted in this thesis. Finally, due to a number of highlighted problem, we motivate our focus on photo tag recommendation.

2.3.1 Automatic Image Annotation (AIA)

Automatic image annotation has been a widely researched area over the last decade with a large number of works attempting to bridge the *semantic gap* between low level image features and high level concepts (Duygulu et al., 2002; Jeon et al., 2003; Krizhevsky et al., 2012; Lavrenko et al., 2002; Makadia et al., 2010).

Problem Formulation Firstly we define the problem of automatic image annotation. Let d_j be the set of tags annotated for the j-th image in our collection. The overall goal in automatic image annotation is therefore to recommend a set of tags, p_j, based on extensive analysis of the image's visual appearance, so that it maximizes $p_j \cap d_j$. Alternatively, an automatic image annotation model predicts the annotations of an image based on its appearance.

Image Features At a digital level, images are a matrix of values, with each value representing some colour at a given location. At a semantic level, images are a collection of objects, concepts and contexts. The overall goal in automatic image annotation is therefore to attempt to map between this *matrix* and its relevant *objects* & *contexts*. Researchers have proposed to achieve this by first making some *sense* of the matrix by capturing an image's high level visual appearance in a process called *feature extraction*.

[5]https://www.mturk.com/ - last accessed on 18th July 2016.
[6]https://www.crowdflower.com/ - last accessed on 18th July 2016.

In feature extraction, the aim is to succinctly capture/represent some aspect of an image's visual appearance e.g. colour, texture, keypoints *etc*. In the last two decades a range of image features which create a vectorial representation of real values, capturing one of these visual aspects, have been proposed. These features generally fall into two high level categories: (i) *global features*, where one attempts to represent the visual aspect of an entire image and (i) *local features*, where one attempts to represent an image as a group of various distinctive local "patches". In the following we summarise some major works in this area:

Firstly, many global *colour* features have been proposed in recent years (Deselaers et al., 2008) which attempt to *invariantly* (i.e. under varying illumination and shading conditions) and *discriminately* capture a scene's colour. The most basic of these features are colour histograms which represent the binned distribution of colours present in an image. Histograms have been used effectively in a number of works to capture an image's high level colour and have been built in a range of additive (i.e. RGB), perceptual (i.e. HSV) and opponent (i.e. LAB) colour models (Schettini et al., 2001). Colour histograms, however, fail to capture the spatial information of pixels, resulting in similar colour distributions for visually diverse images. The colour coherent vector (CCV) (Pass et al., 1996) feature attempted to alleviate this problem by classifying each pixel as "coherent" or "incoherent" based on whether its neighbourhood is similarly coloured i.e. the pixel is part of a larger, similarly coloured region. By doing so, the feature represents the relationship between the number of coherent and incoherent pixels for each colour bin. Deng et al. (2001) proposed the dominant colour descriptor (DCD), which attempted to capture the distribution of dominant colours in an image. In 2001, the MPEG-7 standard (Manjunath, 2002) formalised a range of low level features for image retrieval purposes. This standard included those methods already discussed, as well as other colour features such as the scalable colour descriptor (SCD), colour structure descriptor (CSD) and colour layout descriptor (CLD). More recently, research in this domain has focus on incorporating colour information into more powerful local features (van de Sande et al., 2010).

A range of features which attempt to describe an image's *texture* have also been proposed. Image texture can be defined as the "spatial arrangement of color or intensities in an image" (Stockman and Shapiro, 2001). The MPEG-7 standard also defines a number of texture descriptors, such as the: texture browsing descriptor (TSD), homogeneous texture descriptor (HTD) and the local edge histogram descriptor (EHD) (Manjunath, 2002). The texture browsing descriptor is a compact feature which captures an image's regularity, directionality and coarseness based on a multi-resolution decomposition using Gabor wavelets. The homogeneous texture descriptor computes a quantitative characterisation of texture by computing the mean frequency and standard deviation of an image, after orientation and scale filtering. Finally, the local edge histogram descriptor produces a binned distribution of various edge types (e.g. horizontal, vertical *etc*). In recent years, more success has been achieved with using local features for image retrieval (Deselaers et al., 2008) and annotation (van de Sande et al.,

2010) tasks.

Differing from global colour and texture features which produce a vectorial representation of the *entire* image, local features describe small patches, or points, of an image which differ significantly from their immediate neighbourhood. In order to "extract" these local features, intensity, colour and texture are commonly considered. Local features are powerful tools for the annotation and retrieval of images as they are often "invariant" i.e. unchanged under a number of visual transformations such as lighting and orientation variations. Originally proposed for object recognition (Lowe, 2004), where one wants to retrieve all the images of a given object within a database, a number of invariant local features have been proposed in recent years. Arguably the most prominent work in this area was that of Lowe (2004). In their work, Lowe (2004) proposed the scale-invariant feature transform (SIFT) which attempted to identify the most descriptive local features in an image which were invariant to a number of transformations. This was achieved by identifying local maxima and minima based on a difference of Gaussians over various image sizes. Based on this, the authors were able to match keypoints using a nearest neighbour approach in a k-dimensional tree. Since this paper was published, many researchers have tried to improve on the idea in a number of different ways. For example, Bay et al. (2008) proposed the speeded-up robust feature (SURF) descriptor which, as the name suggests, attempted to improve the performance of the SIFT descriptor. Similarly, Ke and Sukthankar (2004) improved upon SIFT by employing principal component analysis in the keypoint detection phase whilst van de Sande et al. (2010) proposed C-SIFT which encoded colour information within the keypoint descriptor. Due to the discriminative power of these local features, we extract and match SIFT keypoints for image retrieval purposes in Chapter 7.

Finally and most recently, mid-level features have been proposed (Boureau et al., 2010) which transform low-level features (e.g. SIFT descriptors) using *coding* and *pooling* steps; the first coding stage makes a pointwise transformation of these low-level features into a new representation more suited for the task (e.g. increased compactness) before summarising the coded features for larger neighbourhoods (e.g. the entire image for bags of features). For example, Boureau et al. (2010) compare various different coding and pooling techniques in order to automatically identify concepts in the popular *Caltech-101* (Fei-Fei et al., 2007) and *Scenes* (Lazebnik et al., 2006) datasets. Similarly, Oquab et al. (2014) employ convolutional neural networks in order to learn and transfer mid-level features from limited amounts of training data using the ImageNet (Deng et al., 2009) & Pascal VOC (Everingham et al., 2010) annotation tasks to validate their methods. Singh et al. (2012) instead focus on extracting *frequent* yet *discriminate* "patches" as mid-level features. Specifically, the authors propose an iterative process which uses k-means clustering based on HOG features to identify patches before training linear SVM classifiers, where patches *inside* a cluster are deemed as a "positive" example and all others considered "negative". Despite the value of mid-level features, we have instead focused on exploiting low and high level features in this thesis as they also been shown to achieve high annotation accuracy (Li

et al., 2010b; Makadia et al., 2010); we therefore leave the exploitation of mid-level features for photo tag recommendation purposes as future work.

Annotation Models Following over two decades of research, many different automatic image annotation approaches have been proposed. The first works in image annotation used machine translation techniques. Duygulu et al. (2002) attempted to translate between visual "blobs" and high level textual concepts in images, testing on the Corel 5k collection. This work was then extended by Virga and Duygulu (2005) whom compared multiple statistical machine translation models as well as language modelling techniques in order to capture both the visual appearance of images and the semantic relations between annotations for annotation purposes. Jeon et al. (2003) adopted the cross lingual language model of Lavrenko et al. (2002), Cross-Media Relevance Models (CMRM), to predict the probability of generating a word given blobs in an image in the training set. The model assumed that regions in an image can be described by a small vocabulary of blobs, which were created from clustered image features in order to build a probabilistic joint distribution of blobs and tags. Lavrenko et al. (2003) proposed the Continuous-space Relevance Model (CRM) which generalised the previous CMRM to model high dimensional continuous features without clustering and quantization. Feng et al. (2004) improved on the CRM in their Multiple Bernoulli Relevance Model (MBRM) by computing features from segmented image regions. The model computed a joint probability distribution between words and images using a multiple Bernoulli model. Monay and Gatica-Perez (2004) proposed an unsupervised probabilistic latent space image annotation model which offered a novel approach to the construction of the text and visual latent space representations. Feng and Lapata (2010) employed topic modelling techniques, particularly the latent Dirichlet allocation model (Blei et al., 2003), in order to annotate images with keywords under the assumption that images and their co-occurring textual annotations are a mixture of latent topics. Other related matrix factorisation techniques, such as the latent semantic analysis model (Dumais, 2005), have also been adopted for image annotation purposes (Pham et al., 2007).

Athanasakos et al. (2010) showed that the SML (Carneiro et al., 2007) and Multiple Bernoulli Relevance Models (Feng et al., 2004) could be out-performed by using a simple Support Vector Machine (SVM) approach trained on MPEG-7 global features. The authors also demonstrated that the high performance reported by these models was more to do with the test collections used than the approach itself. Yang et al. (2006) attempted region based image annotation using an Asymmetrical Support Vector Machine-based Multiple-Instance learning algorithm, an extension of the traditional Support Vector Machine model which employed loss functions in order to identify false positives & negatives. Goh et al. (2005) employed and compared single-class, two-class and multiple class SVMs trained on low level colour and texture image features in order to identify high level semantic concepts. Additionally the authors employ a confidence based classification hierarchy in order to assess the quality of anno-

tations. Carneiro et al. (2007) proposed a Gaussian mixture model, named SML, using a bag of visual words approach for class conditional dependencies. Fan et al. (2007) proposed hierarchical image annotation classification which attempted to bridge the semantic gap more effectively by attempting to bridge four smaller gaps. The authors learned contextual relationships between image concepts and their co-occurrences in a multi-modal boosting framework before hierarchical image classification was taken out.

Makadia et al. (2010) showed that five of the previously stated models could be outperformed by adopting a K-nearest neighbour approach trained on colour, Gabor and HAAR image features. Du et al. (2009) proposed a non-parametric Bayesian model for clustering images into coherent classes by representing images as an aggregation of distinct, segmented objects, for which annotations are assigned. Wang et al. (2008) aimed to tackle the mismatch between semantic and visual spaces stating that "visual similarity does not guarantee semantic similarity". The authors approached this by computing higher level semantic spaces by clustering correlated tags into topics based on their neighbours and using these topics as annotation lexicon. Zhang et al. (2010) considered feature selection and exploited properties of image features, such as their sparsity, for the automatic annotation of the Corel 5k and IAPR evaluation collections.

More recently, *deep learning methods* which build upon artificial neural networks (ANN) research, originally proposed in 1943 (McCulloch and Pitts, 1943), have adopted much interest for AIA purposes. These neural network classification models, which at their core are multi-layered weighted graphs, have been shown to mimic neural processing functionality in the human brain (Shrager and Johnson, 1995; Utgoff and Stracuzzi, 2002) tailoring their application for image annotation purposes. One of the advantages of deep learning methods is that they are able to replace handcrafted visual features, such as those discussed in the last section, by instead automatically learning patterns from raw data (e.g. pixel intensity) in large image collections. In one of the most significant deep learning works in recent years, Krizhevsky et al. (2012) classified images by building a 7-layer convolutional neural network, containing 650k neurons and 60M connections. This method achieved the highest annotation performance when annotating for 1,000 topics in the large scale ImageNet (Deng et al., 2009) collection. Park et al. (2004) proposed image classification using a neural network which first segmented the largest region at the centre of an image before extracting texture features in order to build a 3-layer neural network for annotation purposes. Similarly, Kim et al. (2004) employed a 3-layer ANN to classify images as "object" or "non-object" by considering the centre 25% of an image. Using this approach, high classification accuracy was achieved on 900 test images, outperforming Naïve Bayes & Decision tree baselines. Most recently, Chan et al. (2015) proposed a new simple deep learning baseline for image classification, called "PCANet", which employs principal component analysis, binary hashing and block-wise histograms in its data processing stage in order to learn classes for various different tasks such as: face recognition, digit recognition & texture classification. In 2016, it is these deep learning methods which are now con-

sidered "state-of-the-art" for image annotation purposes due to their high classification performance on various collections and tracks (Gilbert et al., 2015; Russakovsky et al., 2015).

Related to that of image annotation is the task of scene classification where one attempts to identify a high level scene (e.g. indoor vs outdoor) within an image opposed to building a generic model which annotates for hundreds of visual concepts. Due to the specific nature of scene classifiers, where image features can be tailored for a given purpose, and the low number of classes (usually 2-5), high classification accuracy is achievable. Oliva and Torralba (2001) proposed one of the most popular scene classification features of recent years which attempted to capture the so called "gist" of an image by considering its naturalness, openness, roughness, expansion and ruggedness. Experimenting on 1,500 images, classification accuracy of up to 90% was achieved when testing for four different scenes (i.e. coast, country, forest, mountain). In our work, we adopt this feature in Chapter 3 in order to classify images for two visual scenes.

An extensive overview of automatic image annotation techniques is far beyond the scope of this thesis and many papers/journals have already attempted to undertake this task (Datta et al., 2008; Hanbury, 2008; Smeulders et al., 2000; Zhang et al., 2012a). We refer the reader to these papers for a wider overview of the models and techniques adopted in recent years.

As these works annotate images considering only their visual appearance, their performance is often unsatisfactory (due to the presence of the semantic gap) and computationally expensive. In addition to this, automatic image annotation is unable to identify contextual tags such as the event or location an image is taken at e.g. `titp2014`. This is firstly due to the fact that identifying a specific event or location from a photograph is almost impossible, even for a human; further, the visually similarity of events (e.g. TITP vs Glastonbury festivals) and locations (e.g. Sahara vs Gobi desert) makes this task even more difficult. Further, automatic approaches may find other non-visual concepts difficult to identify (e.g. `anger`, `religion`) as these are often human interpretations of complex visual scenes. In order to overcome these problems, much research has shifted to the area of semi-automatic image annotation, in particular, photo tag recommendation which is therefore the focus of this thesis. In the following sections, we formulate photo tag recommendation before discussing prominent works and evaluation procedures.

2.3.2 Photo Tag Recommendation (PTR)

In recent years, many tag recommendation works have been proposed for a range of web 2.0 applications and scenarios, such as web bookmarking (Chirita et al., 2007; Krestel et al., 2009) and folksonomies (Jäschke et al., 2007). Recently, photo tag recommendation systems have also been proposed, which offer additional tags based on those already added by the user. In the following sections we cover a number of as-

pects regarding photo tag recommendation: firstly, we concretely define the problem of photo tag recommendation before covering prominent works in the area. We then discuss evaluation methodology before describing the metrics, collections and baselines used in this thesis.

Problem Formulation Firstly we define the problem of photo tag recommendation. Considering that d_j is the set of tags annotated for the j-th image in our collection. The overall goal in tag recommendation is therefore to recommend a set of tags, p_j, given a subset of tags q_j (from d_j) so that it maximizes $p_j \cap (d_j - q_j)$. Photo tag recommendation systems differ from automatic image annotation models in that they benefit from the knowledge of 1 or more existing annotations within the given image. PTR models are also expected to identify non-visual concepts (e.g. context) however; an aspect which is generally ignored in AIA evaluation (as the goal is to identify concepts from their *visual appearance* only).

Recommendation Models In order to address the photo tag recommendation task, various different approaches have been proposed in recent years; at the core of almost all of these models, however, lies a tag co-occurrence matrix in which inferences about the relationships of tags can be measured (e.g. that images are often tagged together with `ice` and `snow` *etc*).

Most predominately, Sigurbjörnsson and van Zwol (2008) proposed a tag recommendation strategy to support users annotating photos on Flickr. In particular, the authors highlighted the problem of employing raw co-occurrence for recommendation purposes i.e. popular tags will naturally co-exist highly with other popular tags despite having no meaningful relationship e.g. `blue` & `dog`. Therefore, the authors proposed using various *normalized*, symmetric (i.e. Jaccard) and asymmetric, co-occurrence measures for this purpose. The authors compared tag recommendations using these symmetric and asymmetric measures and noted that the former was more effective at identifying synonyms, whereas the latter offered a more diverse list of tags, thus suiting photo tag recommendation. In this work, the authors evaluated their method on a small test set of images where the ground truth was defined by crowdsourced workers. We believe, however, that a crowdsourced worker will not be able to effectively annotate an image taken by another person as they will not understand the non-visual context for which it was take in (e.g. for an image taken at a football game, the crowdsourced worker will unlikely have knowledge of the team, location, stadium *etc*). In chapter 6 we offer a new evaluation framework for photo tag recommendation which attempts to overcome this problem.

Liu et al. (2009) instead focused on the random ordering of tags annotated by users and proposed a re-ranking method which attempted to determine the relevance of a tag given an image, using a Kernel Density Estimation (KDE) approach. Instead of relying on input tags, however, the authors proposed a recommendation strategy which propagated *prominent* tags from visually similar images, using a k-nearest neighbour

approach. The authors later extended this paper (Liu et al., 2010), highlighting that visual similarity does not always denote semantic similarity; based on this notion, neighbouring images were also selected based on their semantic similarity. The authors also proposed employing an external knowledge base in order to constrain the tagging vocabulary to only visual concepts in the recommendation phase. We believe, however, that a PTR should help a user annotate their image with *contextual* tags (e.g. `London`), as well as visual concepts. Additionally, using a k-nearest neighbour approach poses scalability issues, as for each new image, the visual and semantic similarity must be compared to every other image in the collection. Due to this problem, the authors test their approach on a relatively small images collection (<50,000 images).

Weinberger et al. (2008) proposed a tag recommendation approach which helped users to *better describe their content* by focusing on recommending tags which reduced the overall *ambiguity* within the annotation set. For example, an image tagged with `Cambridge` could relate to both the city in Massachusetts, or the town in the UK. Therefore, Weinberger's approach tried to offer an additional keyword which would increase the Kullback-Leibler divergence of tag distributions within the annotations of Flickr images (i.e. `Massachusetts` or `UK` in our example). Crucially, allowing users to annotate images with *free text* naturally results in ambiguous and ill defined annotation sets, for which the authors suggested exploiting *semantic* evidences in order to alleviate this problem. In this thesis, we extend this notion by exploring additional evidences, such as visual & contextual cues, in order to further reduce this tag ambiguity.

Chen et al. (2008) proposed a system called SheepDog which firstly classified images for 62 different visual concepts, before gathering *popular* tags from related Flickr "groups" (i.e. sub categories of images for a particular concept e.g. dogs $\supseteq$ small dogs $\supseteq$ Cocker Spaniels). Although the users combine both visual and semantic evidences, we hypothesise a number of problems with their approach: (i) the authors assume that all images will be able to be categorised into one of the 62 concepts (ii) as tags are selected based on their *popularity* from related *popular* images, the suggestions will also likely have a bias towards popular tags, which do not offer enough granularity in order to reduce tag ambiguity (e.g. `tinthepark` offers more annotation value than `festival`).

Krestel et al. (2009) viewed the tag recommendation problem from a topic modelling perspective which exploited latent relationships between annotations in historical images. Using the popular Latent Dirichlet Allocation (Blei et al., 2003) model, images were described as a mixture of topics and tags suggested based on their probability of existing in each of the given topics. One drawback of this model however, is that the number of latent topics has to be predefined, which is a non-trivial task for a large scale social image collection. Ma et al. (2010) built a graph using visual and semantic cues in order to capture tag-to-tag relationships for recommendation purposes. Shen and Fan (2010) proposed a classification framework, using support vector machines trained on positive and negative instances for a given tag, in order to make

suggestions. In our work, we instead propose the use of a tag co-occurrence matrix, which beyond the computationally expensive training phase, allows for quick and computationally inexpensive recommendations to be made.

Popular image sharing website, Flickr, also offer their own tag recommendation approach[7], which is described as "returning a list of tags *related* to a given tag, based on clustered usage analysis". As with the other works previously discussed, however, suggestions are only made based on textual features, thus ignoring other signals which can be exploited for recommendation purposes. In this thesis we consider and compare many new evidences for tag recommendation; in particular we focus on exploiting an image's context in order to improve tag recommendations.

We hypothesise that images which are *contextually* similarly will contain similar tags, or will at least have some bias towards certain tags. For example, consider images taken at 3pm on a Saturday afternoon in the Stretford area of Manchester, UK. This is the location of the largest club football stadium in the UK (i.e. Old Trafford) and 3pm on a Saturday is the traditional time for Premier league matches to take place. Thus, ignoring visual cues, images taken at this location are more likely to contain football related tags. If we also have knowledge that the photographer is male, and that the image contains hundreds of faces, we could further infer that the image is probably taken *within* the stadium based on the high number of faces and predominantly male demographic of football fans[8].

Much work has considered the *context* in which a document is created within in other areas of information retrieval (Jones and Brown, 2004; Mylonas et al., 2008; Shen et al., 2005). In particular, researchers have considered the exploitation of the *time* in which a document is created for various purposes. For example, Kleinberg (2002) developed a framework for modelling periodic bursts of keywords in a corpus with hierarchical structure using an infinite-state automaton. More recently, Leskovec et al. (2009) performed a large-scale study of "memes" diffusing throughout news media as a result of temporal rhythms. Context, specifically time and location, have also been studied in some multimedia retrieval papers. For example, Rattenbury et al. (2007) automatically extracted place and event semantics from geo-tagged Flickr images. Specifically, the authors modelled an event as a tag set which was seen to burst for a specific time and location. Similarly, Zhang et al. (2012b) attempted to cluster tags based on geolocation and temporal trends allowing for the construction of tag cluster visualisations. Shane (2006) developed ZoneTag, a mobile phone application which automatically supplied location meta-data for photographs uploaded to Flickr as well as suggested tags from nearby entities, a user's interaction log and their social network. Silva and Martins (2011) exploited the location of geo-referenced images for tag recommendation by suggesting highly occurring tags from geographically nearby

[7]https://www.flickr.com/services/api/flickr.tags.getRelated.html - last accessed on 18th July 2016.

[8]http://opendorse.com/blog/2013-sports-fan-demographics/ - last accessed on 18th July 2016.

and visually similar images. These works, however, either employ context for a different purpose other than annotation purposes, or only focus on a very small part of an image's context i.e. its time and location.

These works have repeatedly highlighted the value of time and location information for a number of tasks; additionally, many other works have even attempted to predict these missing contexts further highlighting their value e.g. time (Thomee et al., 2014) and location (Murdock, 2014) prediction. With regards to time, previous works have generally used a time series model, attempting to identify bursts within these series. In our work, we instead focus on identifying the relationship between tags and: (i) the time of the day (e.g. morning) (ii) the day of the week (e.g. saturday) (iii) the season (e.g. summer). As suggested by Dubinko et al. (2007) interactive Flickr demonstration, we hypothesise that many image tags pertain to *daily*, *weekly* and *yearly* cycles. Using our previous example, we would expect the tag `football` to peak during: (i) the afternoon (ii) the weekend, and (iii) various seasons. The season in which football league is played depends also on the country (e.g. in Scotland the season is predominantly played in the winter months, whereas in Russia the season is played in the summer). Therefore, we also consider the *country* in which an image is taken in as we hypothesise this signal will be effective in many tagging scenarios.

Aside from geospatial features, we hypothesise that there are a multitude of other contexts which can be exploited for our purposes. For example, the gender of the user may introduce tagging biases; in Chapter 3 we compare tagging trends between male and female Flickr users. In related literature, recent work has instead focused on personalised tag recommendation. For example, Garg and Weber (2008) offered personalised tag recommendations in their approach which looked to combine suggestions made from personalised and global tag co-occurrence matrices. In particular, the authors weighted the influence of suggestions made from the personalised tag co-occurrence matrix based on the number of images the users had previously uploaded to Flickr. Rae et al. (2010) exploited a user's social context in the recommendation process by combining different contexts on Flickr, such as a user's tagging history, their social circles and user groups. The authors demonstrate the power of personalised recommendations (i.e. computed from the authors previous uploads), in comparison to making suggestions based on all images or those in their social network. One of the problems with using these user profiling approaches is that they fail in a cold start scenario (i.e. when the user logs into the system for the very first time). In this thesis, we instead focus on exploiting tagging trends of *similar users*, thus overcoming the problem of a cold start, based on a number of easily computable, lightweight features.

Additionally, we also consider *social media* evidences from Twitter for photo tag recommendation purposes as we hypothesise that the textual conversation on microblogging websites will also be of value for tagging purposes for many large scale events. For example, consider a music festival such as *T in the Park*: there will exist much instantaneous social commentary regarding this event (e.g. bands playing, the weather *etc*) on Twitter which we hypothesise could be value for the training of photo tag rec-

ommendation models. In particular, we hypothesise that the speed and coverage of data on microblogging websites (i.e. Twitter) will be far greater than in comparison to image sharing websites (i.e. Flickr) for a number of reasons: (i) due to restrictive mobile data usage plans, many users will wait until they are connected to a WiFi network (e.g. at home) in order to post a high quality image; in comparison, posting a textual message (i.e. a tweet) is of many magnitude smaller in size and therefore cost (ii) many photographers will often go home to visually edit/adjust their their image on a computer before uploading, adding an additional time lag (iii) Twitter[9] also has many more users than Flickr[10]. This exploitation of social media content has been explored in a number of different IR contents. For example, Picault et al. (2013) presented a framework for the indexing and retrieval of video segments by employing text mining and topic modelling techniques in order to collate related tweets. Shamma et al. (2010) exploited microblog posts for the segmentation and summarisation of broadcast media events e.g. 2008 presidential debate. Despite the number of works exploiting social media for multimedia applications, the potential of social media data for photo tag recommendation purposes has not yet been explored. In this thesis, we propose a photo tag recommendation model which draws evidence from both social media streams as well as Wikipedia, presenting preliminary results for this application.

In the following sections we discuss the evaluation of photo tag recommendation models covering the metrics, collections and baselines used within this thesis. First we discuss the "standard" methodology used for PTR evaluation purposes.

Evaluation Procedure As previously discussed, in photo tag recommendation the goal is to offer the user a set of *relevant* tags for a photograph based on some prior textual information. These recommendations should also attempt to maximise concept *diversity* in order to offer the user with a wide range of annotations. The prior textual information takes the form of tags which have already been added by the user in order to annotate their image. In our work, we assume that an image contains at least 1 tag in order to suggest additional tags. Due to the lack of annotations added by users, however, we focus on using other (non-textual) evidences as "input" to photo tag recommendation models.

The standard evaluation procedure takes N tags as input, suggesting K tags to the user. In our work, we set $K = 5$ tags, as adopted by previous work (Garg and Weber, 2008). These K suggestions are then evaluated against those *other* tags (i.e. ignoring the N tags used for input) already added by the user. In order to ensure there exist sufficient ground truth to evaluate against, we only select images which contain at least 4 annotations for testing purpose.

[9]http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ - last accessed 25 July 2015

[10]http://www.statista.com/statistics/252566/number-of-unique-us-visitors-to-flickrcom/ - last accessed 25 July 2015

Relevance based PTR Evaluation Metrics To evaluate our photo tag recommendation methods, we use the following popular metrics, comparing those *suggested* tags against those annotated by the user. All of the discussed metrics are commonly used and have been adopted by previous work in the field of photo tag recommendation (Garg and Weber, 2008; Sigurbjörnsson and van Zwol, 2008):

- *Precision at One (P@1)*: The percentage of runs where the top tag is relevant (equal to S@1).
- *Precision at Five (P@5)*: The percentage of relevant tags amongst the top five, averaged over all runs.
- *Success at Five (S@5)*: The percentage of runs, where there exists at least one relevant tag amongst the top five returned.
- *Mean Reciprocal Rank (MRR)*: Computed as $1/r$ where r is the rank of the first relevant tag returned, averaged over all runs.

Diversity based PTR Evaluation Metrics In 1998, Carbonell and Goldstein (1998) presented a poster which proposed the *diversification* of information retrieval results, creating an entirely new IR evaluation framework in the process. Given a query, traditional IR systems ranked documents according to their estimated relevance, however, searchers' queries are often ambiguous or have multiple facets (Spärck-Jones et al., 2007). For example, *Java* is an ambiguous query since it has different interpretations e.g. the programming language, the island, and the coffee. Further, *java programming language* is a multi-faceted query since it has several aspects, e.g. development kit download, language specifications, tutorials, courses, and books. Ambiguous queries are an issue for search engines and were not originally addressed by early retrieval models. Therefore, research now focusses on promoting *diverse* results lists which have a wider coverage of the query sub-topics, thus offering a solution to search ambiguity (Agrawal et al., 2009; Drosou and Pitoura, 2010; Santos et al., 2011).

In Chapter 5 we also attempt to *diversify* the lists of recommended tags in order to promote a wider range of topics within the suggestions and reduce the number of synonyms offered. To evaluate these methods, we use *intent aware* evaluation metrics which consider the diversity of a suggestion lists. In our application, we translate an *intent*, or sub-topic, to refer to a particular aspect of an image. For example, the tags `sky` and `cloud` would be considered a single sub-topic, with `london` and `uk` considered another. In chapter 5 we reward those recommendation models which maximise these subtopics e.g. a system which suggests the tags [`sky`, `london`] is preferred over one which suggests [`sky`, `clouds`] for an image tagged as [`sky`, `cloud`, `london`, `uk`]. In order to measure this "coverage", we adopt two official *intent aware* metrics from web search evaluation:

- *α-normalised discounted cumulative gain (α-nDCG)*: This metric computes the usefulness, or gain, of a tag based on its position in the recommendation list. We believe that by considering the *information gain* (i.e. beyond relevance) achieved for a given tag, we will more effectively evaluate the *usefulness* of a photo tag recommendation list, thus promoting suggestion lists which offer more benefit (for annotation purposes) to the user (i.e. cover more of an image's aspects). In this metric, the parameter α in α-nDCG balances the importance of relevance and diversity (Clarke et al., 2009). Tuning the value of α changes the rewarding strategy of the metrics for diversity and relevance; particularly, diversity is rewarded over relevance as α increased. In the case where $\alpha = 0$ this metric is equivalent to the traditional nDCG (Järvelin and Kekäläinen, 2002).

- *Intent-aware Expected Reciprocal Rank (ERR-IA)*: This metric computes the contribution of each tag based on the relevance of tags ranked above it, by computing the Expected Reciprocal Rank (ERR) for each sub-topic, with a weighted average computed over subtopics (Clarke et al., 2009). As with α-nDCG, we hypothesise that by considering the *contribution* of a tag in evaluation metrics, and not just its relevance for a given image, we will be able to build better PTR models by proposing those which rank tags based on their information gain.

Both intent aware metrics are reported at two different rank cut-offs: 5 & 10. These cut-offs focus on the evaluation at early ranks, which is important in the task of diversification (Jansen et al., 1998).

PTR Evaluation collections In this thesis we evaluate our photo tag recommendation methods on various public and private image collections. Specifically, we benchmark our main PTR work, which considers image's context (i.e Chapter 3), on the popular & publicly available MIR FLICKR collection which contains 1M images (Huiskes et al., 2010). For evaluating our work which exploits the external context of an image (i.e. Chapter 4, we build a testbed containing images from Flickr, as well as microblog posts from Twitter, related to the Austin City Limits 2012 music festival. Finally, based on a number of problems we identify with the evaluation of existing PTR and AIA works, we propose a new collection which is briefly discussed in this section and fully presented in Chapter 6.

- *MIR-FLICKR 1M:* In our work we evaluate our photo tag recommendation methods, which exploit various internal evidences, against a popular public image test set called MIR-FLICKR 1M (Huiskes et al., 2010) allowing for any future comparative studies to take place. This collection has been widely adopted in various photo annotation works (Srivastava and Salakhutdinov, 2012; Wang et al., 2012) and tasks (Nowak and Dunker, 2010; Nowak et al., 2011; Thomee and Popescu, 2012). The collection contains 865,833 images (at the time of download) under the creative commons license, taken by 40k users, collected from Flickr. These images are annotated

with 853k distinct tags by their photographer with each image annotated with 11.4 tags on average.

Previous works in PTR have evaluated against both crowdsourced (Sigurbjörnsson and van Zwol, 2008) and user (Garg and Weber, 2008) tags as ground truth, however, both introduce a number of biases in the evaluation process. Firstly, crowd-sourced workers cannot be expected to know the full context of an image (e.g. its location, time, people *etc*) and therefore may miss many key annotations for an image (i.e. non-visual aspects). On the contrary, the photographer of an image would be expected to have a more knowledgeable understanding of an image's content and context; therefore we choose to evaluate against user annotations.

Flickr's batch tagging functionality, however, presents a potentially biased evaluation scenario where users are able to tag multiple images with identical tag sets. Using two images from the same "batch" in the test and training sets would therefore create an easier evaluation framework where test data is essentially included within the training set. Therefore, any prior work which uses Flickr data for evaluation purposes and does not take suitable measures to avoid this situation may unintentionally over-estimate the performance of their model. In order to avoid this and present a more difficult test scenario, we only select images from unique users (i.e. users with only one image in the MIR-FLICKR 1M collection) in our test and validation sets. Additionally, in order to have sufficient ground truth to test upon, we only select images containing at least 4 tags. In total, 1,000 *test* images and 500 *validation* images (used for parameter setting purposes) are randomly selected from our collection which fulfil this criteria. We use the remaining 864,333 images for training purposes.

Media Type	**Number**	**Users**	**Number of unique terms**
Tweets	1,507	1,309	14,570
Flickr Images	2,750	68	732
Wikipedia article	-	-	949

Table 2.1 ACL2012 collection by media type.

- *ACL2012:* In chapter 4 we exploit the potential of social media and encyclopaedic data for photo tag recommendation purposes as we hypothesise that there are many situations where there will exist a wealth of online content, related to an image, which could be exploited for the training of recommendation models. In this work we attempt to suggest tags for images taken at a large scale social event. For our experiments we collect tweets, Flickr images and Wikipedia content related to the *Austin City Limits (ACL) 2012 music festival*; an overview of this collection is shown in Table 2.1. We choose to evaluate on a major social event for a number of reasons: (i) there exists much related Flickr, Twitter and Wikipedia content (ii) there exist many sub-events within this overall event e.g. bands playing, parties *etc*, and (iii) the

event contains temporally and geographical diverse content. In order to build the collection we undertook the following processes for each data type:

(a) *Images:* firstly, we searched Flickr using the standard search API[11] for ACL 2012 images annotated with one of the event tags[12] and taken between 11-15 Oct 2012. Although some of the tags are fairly generic and could refer to other entities or events (e.g. ACL is also the acronym for the "Association for Computational Linguistics" conference), we believe that by filtering based on *when* images are taken, we will remove much, if not all, of the irrelevant images. In total we collected 2,750 images taken by 68 users, annotated with 732 different tags, with each image containing 10.6 tags on average. As no additional filters or constraints were used, we consider our collection to be a significant (if not full) set of the *annotated* Flickr ACL 2012 Festival photographs.

(b) *Tweets:* secondly, we collect tweets related to the ACL 2012 music festival from a public Twitter event detection collection (McMinn et al., 2013). Specifically, we select those tweets containing one of the predefined hashtags (as before) which were also posted between 11-15 Oct 2012. We use this collection, which contains three months of a random 1% Twitter stream sample, as unfortunately the Twitter API does not allow for historical tweet search, making the process of gathering tweets for a historical event difficult/impossible. As with our image collection, we apply strict time filters (i.e. tweets posted on 5 days) in order to remove irrelevant tweets - although we would need to manually inspect all of the tweets in our collection to fully support this claim[13]. Using our approach, we collected 1,507 tweets, containing around 14,570 different terms, posted by 1,309 users. The small number of tweets is a direct result of the Twitter streaming API only allowing a 1% sample to be crawled. In order to collate all the related tweets, we would have to download historical tweets, containing specific hashtags and posted between specific dates, which is currently not possible through the official API. As we are only able to collect a 1% sample of tweets, we hypothesise that there exist over 150K relevant tweets (i.e. 1,507 $\times$ 100) related to the ACL music festival 2012. Therefore, considering only a subset of tweets may undervalue their exploitation for tag recommendation purposes in our experiments. Furthermore, we only consider annotated tweets (i.e. those containing a relevant ACL hashtag) thus ignoring the many unannotated posts - the process of identifying these additional tweets

[11] www.flickr.com/services/api/flickr.photos.search.html - last accessed on 18th July 2016.

[12] `acl, acl2012, acl2012acl, aclfest, aclfest2012, aclfestival, aclfestival2012, aclmusicfest, aclmusicfestival, aclmf`

[13] We did not undertake this task as we believe our gathering method to be sufficiently robust in removing noise as well as believing that a manual inspection to not merit the extensive time & cost.

is a very difficult research problem and far beyond the scope of this thesis. Despite this, our crawling approach almost guarantees a set of highly relevant tweets and therefore is suitable for our purposes.

(c) *Wikipedia:* Finally, we crawl the Wikipedia article page related to the ACL festival in this collection as we hypothesise that this will also be a valuable textual resource in order to build tag recommendation models. As with all Wikipedia articles, the page is (i) well curated, due to the wisdom of the crowd, and (ii) extensive, containing much detail regarding, not only the 2012 festival, but also details of previous ACL festivals which may include more relevant information for our purpose (e.g. relevant locations, re-occurring bands *etc*). From this article, we are able to extract 949 distinct terms using the Stanford Parser (De Marneffe et al., 2006).

- *FLICKR-PTR:* In chapter 6 we highlight many problems with the evaluation of existing photo tag recommendation and automatic image annotation works. In particular, the collections used often introduce biases which may be exploited by models in order to "over estimate" their performance. One outcome of this work is a new collection for the evaluation of photo tag recommendation purposes which is more extensively described in chapter 6. This collection is built by querying Flickr for 2,000 popular nouns extracted from WordNet (categorised as `animal`, `artifact`, `body`, `food`, `plant`, `substance`). For these search results, we download the top 2,000 images ordered by relevance (or the maximum number of images if less exist); in total we collect 2M images (under the creative commons library), uploaded by 77k different users, annotated with over 1M tags. On average, each image contains on 15.4 tags which is more than sufficient for evaluation purposes. By collecting many images (i.e. 2,000) from a wide range of topics (which were not selected by ourselves), we are able to create a large scale, real life image collection with PTR evaluation in mind. For testing purposes, we randomly select 1,000 from the described collection where the ground truth (i.e. the user tags) is "clustered" in a crowd sourcing experiment. This collection, as well as the motivation behind many decision in the build process, are described at length in chapter 6.

There exist many other image collections, however, which we could have used in this thesis, most prominently the popular ImageNet (Deng et al., 2009) collection which contains 14M images (as of July 2016) categorised with nouns contained in the WordNet (Miller, 1995) hierarchy. This collection, however, is tailored for *automatic image annotation* purposes where images are annotated with visual classes, thus not reflecting a real-life photo tag recommendation scenario where images can be tagged with various types of tags e.g. verbs, event related hashtags *etc*. Therefore, in order to benchmark against a real life scenario, we train and test our models on user images which have been taken and tagged by *real users* on image sharing website, Flickr; by doing so we create a fairer & more real life evaluation experiment.

PTR baselines In this thesis we compare our tag recommendation approaches against various baselines (BL). These are as follows:

- *Related Tags (REL):* In our first baseline, for a given input tag, we use the recommendations made by Flickr's `tags.getRelated` API method. This method "returns a list of tags *related* to the given tag, based on clustered usage analysis"[14]. We consider this method an *industry strength baseline* due to its implementation and use on the Flickr website. One drawback of this baseline, however, is that Flickr only offers a tag suggestion API for one input tag. Therefore in order to make suggestions for multiple tags, we adopt the following merge process: firstly we weight the recommendation lists for the N input tags. Each rank position of each list is weighted as $1/y$, where y is the rank in the list. Therefore, tags are weighted in an exponentially decreasing fashion, thus promoting the top ranks in each list. The recommendation lists for each tag are then summed and ranked in decreasing order. For new tags where there exists no co-occurrence data, we suggest the most popular tags in the collection. We refer to this system as *REL*.

- *TF-IDF Approach (TF-IDF):* Given a tag co-occurrence matrix (which counts the number of images two tags co-exist in), many existing tag recommendation models can be used in order to make tag suggestions. These models follow the intuition that tags which co-exist highly in previous images will also be tagged together in future images (e.g. `ice` & `snow`, `tennis` & `ball` *etc*). Therefore, in order to make our approach adoptable in future models, we also make recommendations based on a co-occurrence matrix as input. Firstly, we formulate tag co-occurrence:

  > If we assume that in total k unique tags represent the images in a collection of size m, the tag co-occurrence matrix would be a square matrix C_k where the value of the element $C_{t_{ij}}$ represents the number of images that contain both t_i and t_j tags. We define the representation of a tag t_i as a vector $t_i' = (C_{t_{i1}}, C_{t_{i2}}, \ldots, C_{t_{ik}})$ where each dimension corresponds to t_i's co-occurrence value with another tag.

  We make tag recommendations based on these co-occurrence values, using the *TF-IDF* weighting scheme (Jones, 1972; Manning et al., 2008). This approach is as follows: Given a tag co-occurrence matrix C, we firstly normalise each column by scaling by its maximum value, as proposed by Garg and Weber (2008). The co-occurrence vector, t_i', therefore represents the *TF* part of TF-IDF, where *IDF* is the vector of *inverse document frequencies*. In order to build an IDF vector which weights the importance of every tag in our collection, we compute $\log(m/m^{(t_j)})$,

[14] https://www.flickr.com/services/api/flickr.tags.getRelated.html - last accessed on 18th July 2016.

where $m^{(t_j)}$ is the number of images containing tag t_j. Therefore, given one or more input tags, recommendations can be computed by:

$$O_q = (C \times q) \cdot \text{IDF} \tag{2.3}$$

where q is the binary vector of tags used as a query from the image's tag list. For multiple tags, the corresponding contributions are *added* together. Therefore, O_q is a vector of length k, where each element represents the probability of recommending the given tag. For new tags where there exists no co-occurrence data, we suggest the most popular tags in the collection. We refer to this system as *TF-IDF* and consider it as a strong baseline due to its adoption by many other tag recommendation works (Garg and Weber, 2008; Takashita et al., 2010; Zangerle et al., 2013) as well as its performance for many other related tasks (Spina et al., 2012; Trieschnigg et al., 2007; Wang and Manning, 2012).

- *Popular Global Tags (POP):* Finally, we propose a simple, naïve baseline where we suggest the most popular tags in the collection, in descending order, based on the number of images they exist in. This baseline is only compared against in a cold start scenario i.e. where there exist no tags within an image. We refer to this system as *POP*.

In the various works in Part II of this thesis, we explore the exploitation of context in the tag recommendation process. In particular, we aim to address the first of two high level research questions in this thesis:

> *[HL-RQ1]* Can the *context* an image is taken in be exploited for photo tag recommendation purposes in order to complement existing textual evidences? Which contexts are most effective for this task?

In order to fully address this research question, we propose two different works as follows:

- we explore *internal* evidences (i.e. those which can be directly extracted from an image itself e.g. its orientation) for tag recommendation purposes (Chapter 3).
- we explore *external* evidences (i.e. those which can be extracted from media related to an image e.g. social media content posted at the same festival) for tag recommendation purposes (Chapter 4).

2.3.3 Photo Annotation Summary

Over the last two decades many researchers have attempted to "bridge the semantic gap" between an image's appearance and its high level concepts (Smeulders et al.,

2000; Wang and Hua, 2011). To this end, thousands of papers[15] have proposed different automatic image annotation models which draw upon local, mid and global features in order to automatically tag images with relevant concepts. Most prominently, the works employing deep neural networks have gained the largest traction with some achieving near human levels of accuracy (Krizhevsky et al., 2012). Despite these large gains, we believe that some concepts, such as event (e.g. `TITP`) or location (e.g. `Scotland`) related information, are *impossible* to infer solely from visual appearance and that the user must be employed somewhere within the annotation loop.

Photo tag recommendation presents a reasonable compromise where tags are suggested to the user based on those already existing. Unlike the popular area of automatic image annotation, few works have attempted to build tag recommendation systems specifically for images[16]. Photo tag recommendation systems are able to draw on many additional contexts (e.g. time, location, user *etc*), however, which to our knowledge have not been fully exploited by existing works (Garg and Weber, 2008; Sigurbjörnsson and van Zwol, 2008). Therefore, in this thesis we propose to exploit the full context in which an image is captured in order to improve the effectiveness of tag recommendation for photo annotation purposes.

Finally, with respect to evaluation, image annotation and photo tag recommendation have mostly followed the "Cranfield approach" in order to benchmark and compare methods, however, in this process they have over looked many *image specific* evaluation problems which may lead to misleading results. In chapter 6, we present 7 problems with existing image annotation evaluations and 3 with existing photo tag recommendation evaluations.

2.4 Photo Retrieval and Recommendation

Even if an image is sufficiently annotated, retrieval and recommendation are still difficult and open research problems due to tag ambiguity and the small amount of textual content available to index upon. In the following sections, we discuss the various paradigms and recent research works in *photo retrieval* and *recommendation* which attempt to alleviate these issues. Based on this, we motivate the benefit of image context for retrieval and recommendation purposes which are introduced in Part III.

2.4.1 Content based image retrieval (CBIR)

The field of CBIR relates to any work which helps facilitate the retrieval of multimedia content based on *visual appearance*. Since its inception over two decades ago

[15] A search on https://scholar.google.com for "automatic image annotation" returns 5,280 results - last accessed 8th July 2016.

[16] A search on https://scholar.google.com only returns 28 results for "photo tag recommendation" and 81 results for "image tag recommendation" - last accessed 8th July 2016.

(Kato, 1992), the field now encompasses techniques and research from computer vision, human computer interaction (HCI), information retrieval and many more areas. In the following sections we formulate content based image retrieval before discussing prominent models which shaped the field.

Problem Formulation Content-based image retrieval concerns the searching of images from photo databases by exploiting computer vision techniques. More specifically, this branch of image retrieval analyses/queries the visual contents of images (e.g. colour, textures, faces *etc*), rather than metadata (e.g. keywords) which is often missing (Sigurbjörnsson and van Zwol, 2008), in order to return a set of images d in descending relevance.

CBIR Models From the early days of content based image retrieval (Smeulders et al., 2000) to web scale semantic search (Russakovsky et al., 2014), image retrieval has taken a number of different forms and has considered a number of different applications. In the earliest works, authors concentrated on matching images based on their high level visual appearance, such as colour, texture, shape *etc*. The term, Content-based Image Retrieval, was first proposed in 1992 by Kato (1992) who attempted to retrieve images from a database based on the shapes and colours within. The authors attempted to model and match between visual features of an image and features of a human sketch. From this point, an entire field of research emerged with IBM offering one of the earliest commercial products which employed CBIR techniques (Faloutsos et al., 1994; Smith and Chang, 1996), called Query By Image Content (QBIC). This retrieval system indexed images based on various visual features and provided tools which allowed users to filter/rank images based on these criteria.

By the end of the 90's, relying *purely* on visual similarity was considered insufficient due to the so-called semantic gap between low level features and high level concepts within images (Smeulders et al., 2000). Due to this finding, many works instead focused on many more specific applications of content based image retrieval, such as medical image retrieval (Müller et al., 2006), facial recognition (Kong et al., 2005), art retrieval (Chen et al., 2005), number plate identification (Ondrej et al., 2007) *etc*. One popular search paradigm which gained much interest in the late nineties and early noughties (i.e. 2000's) is that of *object recognition* where the goal is to identify exact matches of an item/object within an overall collection. This is a challenging task however, as the queried item may be partially hidden within a given image, rotated, or observed under different lighting conditions *etc*. As previously discussed, the most prominent work in this area was that of Lowe (2004) who proposed the scale-invariant feature transform (SIFT) which attempted to identify the most descriptive local features in an image which were invariant to a number of transformations. Based on this work, many other works have attempted to improve this "local" feature (Bay et al., 2008; Ke and Sukthankar, 2004; van de Sande et al., 2010) as well as employ its discriminative power for a number of purposes beyond that of pure object recognition, such as: du-

plicate video detection (Sivic and Zisserman, 2003), image stitching (Szeliski, 2006), concept detection (Fergus et al., 2005) *etc*.

With the recent rise of social media, and therefore number of photographs online, mainstream search engines, such as Google and Bing, have had renewed interest in content-based image retrieval. One area which has gained much interest in particular is that of query-by-example (QBE) (Flickner et al., 1995), or reverse image search, where the user provides an image as a query in order to capture their search intent. Research interest on this 20 years old paradigm (Flickner et al., 1995) has been renewed with a focus on web scale collections. In particular, many works have attempted to find duplicate (or near duplicate e.g. cropped) images in large collections. For example, Wu et al. (2009) proposed a query-by-example image search model for detecting partial duplicates on a collection of more than one million images. In their work, the author highlighted that a traditional bag-of-visual words approach ignores their geometric relationships which are often crucial for QBE search purposes. Based on this, the authors proposed a method which instead bundled image features into local groups. Ke et al. (2004) proposed a technique which attempted to identify duplicate images which had had common alterations such as contract/saturation adjustment, cropping *etc*. Specifically the authors employed local descriptors, indexed using a locality sensitive hashing technique, in order to match between images. Foo et al. (2007) compared the effectiveness of dynamic partial functions (DPF) and local descriptors for the purpose of duplicate detection; they highlighted that the former was more efficient but the latter more accurate.

Near duplicate detection is often used for filtering purposes in order to avoid displaying duplicate content within the top results of image search pages, however, other works have instead focused on promoting *visual diversity* in the rankings. For example, van Leuken et al. (2009) focused on the visual diversification of image search results. In particular, the authors proposed a method which clusters images based on a number of (weighted) visual features before using a round robin selection process to create ranked lists. In order to encourage work in this area and create a reliable test collection, ImageCLEF recently focused on *photo diversification* in their photo retrieval track (Paramita et al., 2010).

Finally, the popularity of new lifelogging devices has meant that many users now capture *thousands* of spontaneous, hands-free images *everyday*, presenting a new challenge of *personalised large-scale image retrieval*. These logs, which are captured randomly at pre-determined intervals by a wearable camera, compromise of various rich data-types such as: (i) images (ii) video (iii) audio (iv) GPS (v) acceleration sensor data *etc*. Most predominantly, and most relevant to this thesis, many works have attempted to automatically organise the collections of photos. For example, Doherty and Smeaton (2008) attempted to segment these collections into semantic "events" by comparing blocks of adjacent images based on various MPEG-7 descriptors such as: (i) colour layout (ii) colour structure (iii) scalable colour, and (iv) edge histograms. Wang and Smeaton (2012) attempted to categorise lifelogs into *daily activities* (e.g. *"taking*

a phonecall", *"cooking"*) using a density-based approach. The retrieval and organisation of lifelog images presents an exciting new area for multimedia researchers and in this thesis we focus on exploiting the context surrounding an image to automatically summarise a user's day.

Content-based image retrieval is forever evolving and opening new areas of research; recent work has also focused on the retrieval of *3D images* which has applications in many areas, such as medical image retrieval. For example, Gao et al. (2012) proposed an approach which avoided problems associated with determining the distance of 3D objects taken in different views by employing a hypergraph constructed from various 2D images of the objects. Due to this rise in popularity of 3D retrieval, ImageCLEF (Caputo et al., 2014) also created a new track on *liver retrieval* where 3D matrices are provided for training and test purposes.

As Smeulders et al. (2000) hypothesised in 2000, we also believe that the semantic gap cannot be bridged by depending solely on visual contents because often the semantics of an image exist outwith its pixels. In this thesis, we therefore concentrate on exploiting other aspects of an image, in particular its context, for retrieval and recommendation purposes.

2.4.2 Text based Image Retrieval

The most popular paradigm, and the method of choice for major search engines, is that of text, or tag, based image retrieval. In the following section we firstly formulate the problem before discussing prominent works in this area.

Problem Formulation As with traditional document search, in text based image retrieval the overall goal is to rank a set of items (i.e. images) d in descending relevance based upon some textual query q which captures their intent. Since images are annotated with tags with are often noisy and incomplete, however, simply applying traditional ad-hoc search techniques often gives poor results. Therefore, much work has focused specifically on searching for images using *tags*.

Text based Image Retrieval Models Many existing approaches have attempted to estimate the relevance of a tag for a given image in order to facilitate effective retrieval. For example, Gao et al. (2013) proposed to estimate relevance using a hypergraph learning approach for tag based image search. In their method, vertices in the graph denote user images which are represented as bags of visual and textual words which are in turn used to generate hyperedges between vertices. The authors then used an iterative learning approach to calculate hyperedge weights and estimate the importance of given visual words and tags. Li et al. (2010c) attempted to learn the relevance of tags with respect to an image's visual content by analysing the tags of visual similar images. Using a neighbour voting scheme, redundant tags were identified and ultimately tag relevance for a given image was estimated.

Other works have instead focused on handling *tag ambiguity* in tag based image retrieval systems, for example: Cai et al. (2004) attempted to organise image search results into semantic clusters (e.g. `apple`: pictures of the fruit *vs* the technology company). Specifically, the authors proposed a hierarchical clustering approach based on visual & textual features as well as analysis of web links. Similarly, Clough et al. (2005) proposed a hierarchical clustering of concepts for image retrieval purposes by computing co-occurrence relationships between tags. In particular, the author computed document frequency and a statistical relation called sub-sumption in order to determine whether a related tag was a *parent*. Wang et al. (2010b) instead focused on maximising visual diversity in search rankings. Their method, diverse relevance ranking (DRR), attempted to reduce the visual and textual (i.e. tag) similarity in the images in the top ranks by measuring visual similarity based on simple colour & texture features and tag similarity based on an adaptation of the Google distance (Cilibrasi and Vitanyi, 2007) - a normalised metric which measures the semantic similarity between two terms/phrases based on the number of pages returned for each by the Google search engine.

Due to the lack of textual evidence and difficulty of the image retrieval task, an entire body of research has focused on trying to understand the intent of the user. One aspect of this relies on relevance feedback (Zhou and Huang, 2003), where evidences are drawn from implicit or explicit user interactions. Ever since its inception in 1998 (Rui et al., 1998), many works have been proposed which have exploited various types of feedback: for example, (Cheng, 2006) exploited click through data from a commercial image search engine in order to improve precision in the top 20 results. Specifically, the authors combined textual evidence from an image's meta-data as well as visual features in order to optimize the query using the popular Rocchio algorithm (Salton, 1971). Rui and Huang (2000) explored the exploitation of explicit user relevance feedback for image retrieval on a collection of 17,000 images. In their work, the authors presented experimental users with a "degree-of-relevance" slider in order to allow the user to give explicit feedback on the quality of search results, and ultimately to improve retrieval. One cannot depend solely on explicit feedback, however, as users often do not use these features. As a result, research has mostly focused on extracting semantics purely from *implicit* feedback such as: click data (Cheng et al., 2006) *etc*.

Due to the growing interest in tag based image search area, open standardized retrieval challenges have been proposed, such as that by Huiskes and Lew (2008), where readers are encouraged to benchmark tag retrieval methods on a public Flickr collection. This retrieval challenge and the papers previously referenced, however, have mostly benchmarked their retrieval methods on small-medium sized collections from image sharing websites. These images tend to have at least some descriptive textual content (e.g. tags, description, comments, user profile *etc*), whereas on the contrary, companies building search applications for images crawled from websites do not have this luxury. Therefore, in order to overcome this problem, commercial image search engines have instead focused on indexing images based on a number of related

textual sources, such as: (i) meta-data (ii) filenames (iii) surrounding text on a website *etc*. These sources of evidence are often unreliable, however, as filenames are often never changed from when they were uploaded from the camera (e.g. IMG-4238.JPG) and surrounding text may be irrelevant for many reasons (e.g. adverts, related articles *etc*). In this thesis, we therefore consider new contextual features, for retrieval and recommendation purposes, which can be easily drawn from meta-tags (e.g. time taken) or highly actuate visual analysis tools (e.g. face detection) in order to gain *additional* insight into its semantics.

2.4.3 Photo Recommendation (PR)

Finally, in the following subsections we discuss *photo recommendation* approaches and their uses & applications. Firstly, we define the problem before discussing a number of prominent works in this area, supporting our motivation for research in Part III.

Problem Formulation In photo recommendation the overall goal is to rank a set of images d in descending relevance based on *no* existing query; photo recommendation can therefore be viewed as an image retrieval problem where $q = \emptyset$. In PR, evidences are instead drawn from other sources, such as a user's preference to a given topic or similarity to other users.

Photo Recommendation Models Photo recommendation approaches exist to solve a number of different scenarios and are often used to complement photo retrieval systems. One example scenario for photo recommendation is that of suggestion applications on social media and image sharing platforms. For example, Elahi et al. (2013) built a cross-domain user model from a user's Facebook profile, gathering information about their friends, photos, comments and likes as well as background information from an external knowledge base, in order to measure user interests to facilitate photo recommendation. Siersdorfer and Sizov (2009) proposed a collaborative filtering approach for photo recommendation purposes on Flickr. Specifically, the authors built a tripartite network capturing the relationships between users, tags and images in order to suggest photographs. Zheng et al. (2010) instead attempted to suggest Flickr *groups* (i.e. collections of collaboratively sources images on a specific topic) to the user based on a number of collaborative filtering approaches computed from their existing group-participation behaviour. Finally, Flickr offer their own photo recommendation system, called "Interestingness", which, as the name suggests, attempts to recommend *interesting* photographs to the user. Although no detailed specification exists regarding its implementation, the feature is described as modelling *interestingness* based on click-through, comment & favourite data as well as image tags and "many more things which are constantly changing".

Other works have proposed photo recommendation based on an image's aesthetics, which could aid the user in a number of scenarios e.g. selecting the best image(s)

within a personal album for photo editing purposes. For example, Li et al. (2010a) proposed an automatic technique to evaluate the aesthetic of a photo within a larger collection based on a number of features, such as: (i) colour and lighting (ii) composition characteristics (iii) facial characteristics. Similarly, Jiang et al. (2010) also attempted to identify an image's aesthetic value by employing a regression technique trained upon similar features as Li et al. (2010a), such as: (i) colourfulness (ii) contrast (iii) symmetry (iv) the size, number and orientation of faces present.

The previously discussed works either depend solely upon visual features or textual content in order to suggest photos within their given application. In this thesis we instead propose the exploitation of *contextual features* for which there has been a great research focus recently with new dedicated tracks proposed, such as the Context-awareness in Retrieval and Recommendation (CARR) workshop[17] which runs in conjunction with the ECIR conference[18]. In this workshop, many papers have been proposed which exploit contextual features for information retrieval and recommendation problems. For example, Lim et al. (2015) explore a user's social and recent geographical context in order to prediction their location more accurately. From our own analysis of the five years of proceedings in this workshop[19], however, no papers have been considered context in a photo recommendation environment. In our work, we consider various new contextual cues in order to recommended photographs in both a cold start scenario as well as in an environment where there exist sufficient textual evidences. In particular, we propose to address the following research question:

> *[HL-RQ2]* Can the *context* an image is taken in be exploited for image retrieval & recommendation purposes? How can these contexts be used to alleviate the problems associated with retrieval & recommendation on un-annotated images?

Specifically, in this thesis we explore the exploitation of image context for three different tasks:

- firstly, we propose the task of *visual event summarisation* where we attempt to rank the most relevant images for a bursting news story identified on Twitter (Chapter 7).
- recommending photos based on their predicted future popularity on Flickr (Chapter 8).
- suggesting photographs (from lifelog data) which best describe the key moments in a user's day (Chapter 9).

[17] http://carr-workshop.org - last accessed on 18th July 2016.

[18] http://irsg.bcs.org/ecir.php - last accessed on 18th July 2016.

[19] 2011: http://goo.gl/mfvrPC 2012: http://goo.gl/4MS5tR 2013: http://goo.gl/WkpP2S 2014: http://goo.gl/n9v72N 2015: http://goo.gl/CPV1bE - last accessed on 18th July 2016.

2.5 Thesis Problem Statement

Images posted online are poorly annotated by nature due to the extensive effort required (beyond the shooting, editing and uploading phases) in order to describe their image. By uploading to a *social* network, however, users *want*, or expect, their photographs to gain some social interaction after they are uploaded (e.g. "likes", comments *etc*). This causes problems for websites such as Flickr as, in order to encourage these user interactions, they must be able to effectively rank, retrieve and recommend images (in various applications) which requires some understanding of the content. Automatic image annotation approaches have offered some assistance, however, these models present new problems: (i) annotations are often incorrect due to the difficult nature of the task and presence of the semantic gap (ii) automatic approaches are only able to annotate with visual concepts, thus missing key contextual information that could only be tagged by the photographer themselves (e.g. that the image was taken at their son's school's sports day) (iii) also, automatic approaches are often computationally expensive in both the training and testing phases. Therefore, we believe that the user must be included, in some form, within the annotation phase.

As a bridge between manual and automatic annotation, semi-automatic photo tag recommendation models have been proposed which suggest related tags based on existing textual annotations added by the user. Likewise, retrieval and recommendation approaches are generally built under the assumption of descriptive textual evidences being available, which is often not the case. In this thesis, we propose the exploitation of image context in the annotation, retrieval and recommendation of images. By doing so, we hypothesise that photo tagging effectiveness and retrieval performance will significantly improve, especially in situations where no textual evidences exist (i.e. a cold start problem); our hypothesis is driven by the observed value of context in other areas of information retrieval (Jones and Brown, 2004) as well as our own survey on "image tagging trends and tendencies", presented in section 3.3, which shows that various aspects around an image's capture often correlate with its annotations. To this end, we propose 17 internal evidences (i.e. from the image itself) and various external evidences (i.e gathered from textual sources related to an image) for our purposes. It should be noted that many of these features are extracted from the *time* an image is taken, however, as demonstrated by Thomee et al. (2014), many cameras have their timestamps set incorrectly. Therefore, in order to alleviate this problem we propose various other contextual cues which can be more reliably inferred (e.g. photo orientation, device type *etc*).

2.6 Chapter Summary

In this chapter we presented an overview and introduction to information retrieval, photo annotation, automatic image annotation, photo tag recommendation as well as photo retrieval & recommendation. Based on our background work, we motivated our

focus for employing photo tag recommendation models over automatic image annotation approaches due to the lack of coincidence between visual features & annotations (i.e. the semantic gap). We further motivated that automatic image annotation approaches cannot identify *non-object* concepts related to an image (e.g. its location) which account for the majority of tags as used by users on Flickr (as shown in Chapter 3). Therefore in order to address this issue, in the first part of this thesis we explore the exploitation of *context* for photo tag recommendation purposes, specifically addressing our first high-level research question:

> **HL-RQ1.** Can the *context* an image is taken in be exploited for photo tag recommendation purposes in order to complement existing textual evidences? Which contexts are most effective for this task?

In the following chapters we propose various features from the image context, user context & visual appearance as well as novel techniques for their exploitation in the semi-automatic annotation paradigm of photo tag recommendation.

In latter parts of this chapter, we discussed photo retrieval & recommendation paradigms proposed in recent years emphasising that this area is still an open research problem with many unexplored avenues. Specifically, most work has been taken out in the area of *Content Based Image Retrieval (CBIR)*, thus often overlooking simpler contextual cues. To this end, in the second part of this thesis we explore contextual features for photo retrieval and recommendation purposes, demonstrating the value of *context* and addressing our second high-level research question:

> **HL-RQ2.** Can the *context* an image is taken in be exploited for image recommendation and retrieval purposes? How can this context be used to alleviate the problems associated with retrieval on un-annotated images?

Specifically, in part III we propose three new image retrieval & recommendation paradigms which explore context in order to significantly improve upon content-based methods.

Part II

Photo Annotation

In Part II of this thesis we consider the role of *context* in the task of photo tag recommendation. In the following, we focus on exploiting an image's *context* for annotation purposes, as is discussed in Chapters 3 & 4, before discussing the diversification of photo tag recommendation lists in Chapter 5. Finally in Chapter 6 we present problems with the evaluation of automatic image annotation and photo recommendation systems which were uncovered during the research conducted in this part. Overall, in this part we aim to address the following high level research question (i.e. HL-RQ1): "Can the *context* an image is taken in be exploited for photo tag recommendation purposes in order to complement existing textual evidences?"

Chapter 3

Internal Evidences for PTR

The following chapter is published in the following conferences:

- Philip J. McParlane, Yashar Moshfeghi, Joemon M. Jose (2013); On contextual photo tag recommendation, *Proceedings of ACM Special Interest Group on Information Retrieval (SIGIR) 2013, Dublin, Ireland.* ACM New York, NY, USA, pp965-968.
- Philip J. McParlane, Yelena Mejova, Ingmar Weber (2013); Detecting Friday Night Party Photos: Semantics for Tag Recommendation, *Proceedings of European Conference on Information Retrieval (ECIR) 2013, Moscow, Russia.* Springer-Verlag Berlin, Heidelberg,
- Philip J. McParlane, Joemon M. Jose (2013); Exploiting Time in Automatic Image Tagging, *Proceedings of European Conference on Information Retrieval (ECIR) 2013, Moscow, Russia.* Springer-Verlag Berlin, Heidelberg, pp520-531.
- Philip J. McParlane, Stewart Whiting, Joemon M. Jose (2013); Improving Automatic Image Tagging Using Temporal Tag Co-occurrence, *Proceedings of ACM International Conference on MultiMedia Modeling (MMM) 2013, Huangshan, China.* Volume 7733 of the series Lecture Notes in Computer Science, pp251-262.

3.1 Introduction

In the first part of this thesis we consider the task of photo tag recommendation (PTR) where new tags are suggested to the user based on some existing annotations, as discussed in the previous sections. Existing works in this area have mostly considered relationships between tags (Liu et al., 2009, 2010; Sigurbjörnsson and van Zwol, 2008) thus ignoring various contextual signals, despite years of context exploitation in other fields (Baldauf et al., 2007). An image is not taken within a bubble; there exist many different features which may give insight into the semantics of the photograph. Although time and location have been explored in related fields (Zhang et al., 2012b),

no work has considered the full context an image is taken within. For example, many more evidences exist which can be exploited for the purposes of photo tag recommendation, for example: (i) the photographer's gender (ii) the image's orientation (iii) the shooting device type *etc*. In this chapter we consider *internal* image evidences which we define as those:

> ...which are directly extractable from the pixels, context, or user context surrounding an image (e.g. time, location, number of faces *etc*).

Specifically, in this thesis we introduce and compare the effectiveness of 17 *internal* evidences for PTR purposes. We particularly focus on their application in a *cold start* scenario where images contain no annotations or are annotated with new, *unseen* tags. This is especially important as: (i) around 1 in 6 images are unannotated, and (ii) over half of tags are used only once in our FLICKR-COL collection.

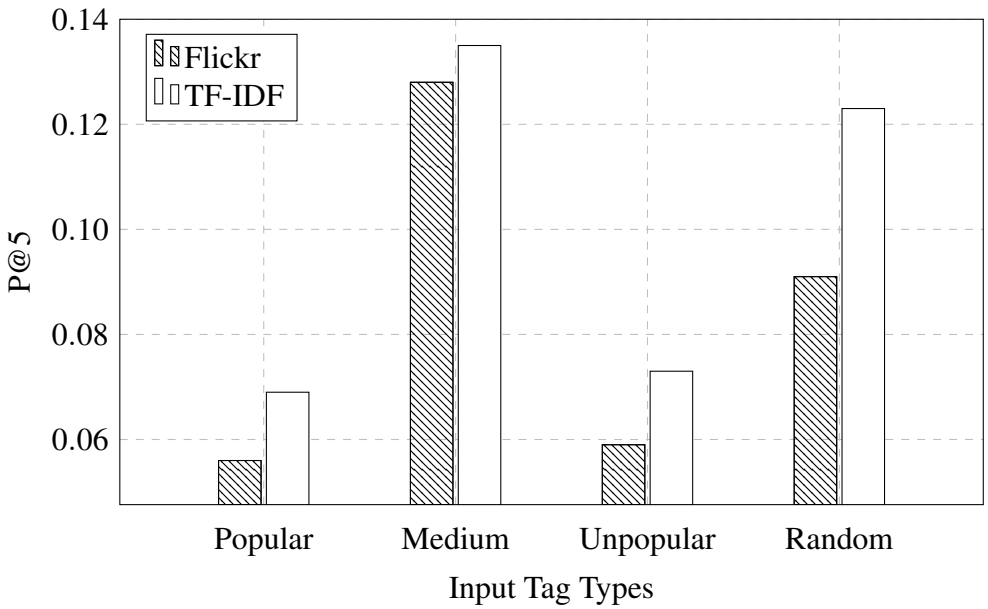

Fig. 3.1 Comparison of annotation performance using one (popular, medium frequency, unpopular or random) tag from images as input to two state-of-the-art tag recommendation models, testing on 1,000 images from the MIR FLICKR 1M collection. *Flickr* refers to tag recommendations made using the Flickr tag recommendation API and *TF-IDF* is the recommendation approach proposed by Garg and Weber (2008).

Even where there exists textual evidence, previous recommendation approaches often offer poor suggestions when the input tag is (i) extremely *popular*, and therefore vague or ill-defined (e.g. `blue`), or (ii) extremely *unpopular*, and therefore lacking in training data. In extreme cases, an input tag may not exist in the training collection and therefore no inferences can be made . To emphasise this problem, we compare tag recommendation performance for two different models[1] (Garg and Weber, 2008) used in

[1]https://www.flickr.com/services/api/flickr.tags.getRelated.html - last accessed on 18th July 2016.

this work (see Section 2.3.2), tested on 1,000 Flickr images from the MIR FLICKR 1M collection (see Section 2.3.2), using different input tags for each approach (based on their popularity). For example, for each test image we compare the recommendations made based on using the most popular tag annotated by the user as input, *vs* the least popular tag, *vs* the median frequency tag and *vs* a random tag in the annotations, benchmarking against the *other* user annotations. Figure 3.1 highlights the varying performance of each tag type where we observe much higher recommendation performance when using the median frequency tag in comparison to using popular or unpopular tags, agreeable with Luhn's hypothesis of term significance (Luhn, 1958) i.e. that the tags with mean frequency in a collection carry the highest value. In our experiments we show that the proposed new evidences can be used to improve recommendation accuracy when input tags are unreliable (i.e. extremely popular or unpopular).

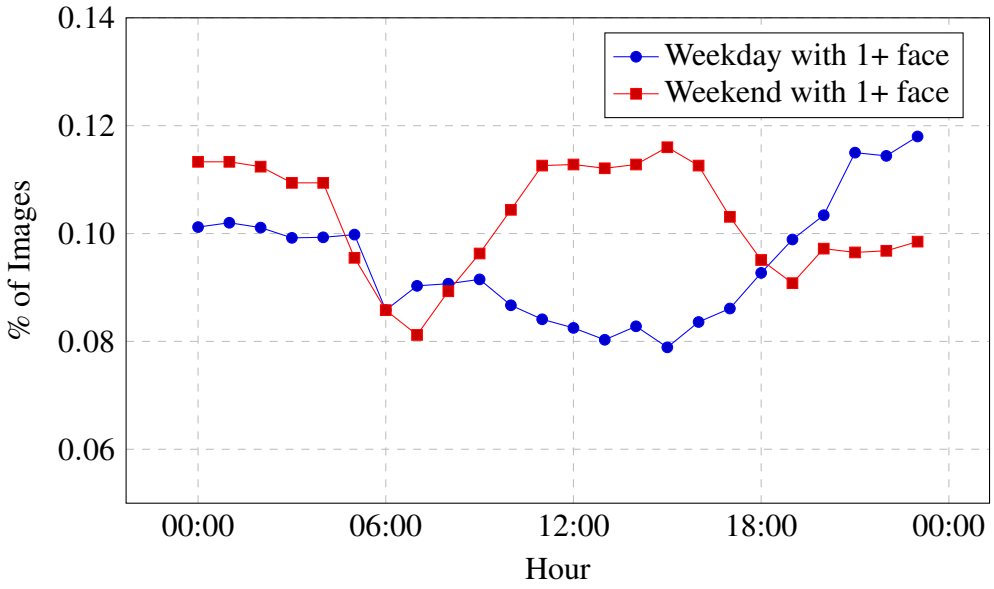

Fig. 3.2 Facial temporal latent relationships

Finally, we consider the *combination* of these evidences for photo tag recommendation purposes which aim to exploit latent relationships between them. To illustrate the latent relationship between many features, Figure 3.2 plots the daily and hourly temporal trends of the average number of faces in a Flickr image for the MIR FLICKR 1M collection. From this Figure we can make a number of observations: (i) on weekdays, the number of images containing faces *decreases* during the day time (i.e. when most people are working) and rises in the evening (i.e. when the majority of people are relaxing at home) (ii) on the contrary, at the weekend, the number of faces *increases* during the day (i.e. when people are socialising) and peaks again in the early hours of the morning (i.e. when many people are at parties). Crucially, the number of faces in an image is correlated with *multiple* temporal features i.e. the time of the day and the day of the week. In our work, it is these latent relationships which we aim to further exploit by using linear combination approaches for this purpose.

In this first work, we address the following research questions:

RQ3.1 Can image context, image content and user context be exploited for photo tag recommendation purposes in order to complement existing textual evidences? Which features are most effective for this purpose?

RQ3.2 Can we capture the latent relationships between the discussed evidences to further improve recommendation accuracy?

RQ3.3 Can these features be used in a cold start recommendation scenario? Can we effectively exploit these features when the input tag(s) of an image are ill-defined or lacking in training data?

The rest of this chapter is as follows: firstly we discuss the various features and evidence combination approaches in Sections 3.2 and 3.2.5. In Section 3.3 we observe how these evidences influence the tagging trends and tendencies of users in our collection. Section 3.4 discusses the evaluation collection, baseline, experimental systems, metrics and evaluation process. Finally, we present the findings of our results in Section 3.5 before concluding in Section 3.6.

3.2 New Evidences for PTR

In the following subsections we formulate our problem in section 3.2.1 before discussing our contextual features in sections 3.2.2, 3.2.3 and 3.2.4. Finally, we detail our feature combination approaches in section 3.2.5.

3.2.1 Formalisation

In this work, we classify a given image into various categories based on a number of different features regarding the image's context, appearance or user context. Formally, given an image, we classify for a number of new features $f = \{f_1 \dots f_g\}$, where each has two or more classes. We define S_f to be the subset of images sharing a common feature class (e.g. S_{night} is the set of images taken at night) and define $TF(S_f)$ & $IDF(S_f)$ to be tag frequency and inverse document frequency vectors computed for images in S_f.

From these subsets, we are able to build contextually, visually and user specific tag co-occurrence matrices C_f, based on a given feature f. For example, C_{night}, is the tag co-occurrence matrix built only on images taken at night (i.e. S_{night}). In our approach, we use these tag co-occurrence matrices in order to make *contextually aware* recommendations. The features introduced in this work (with classes highlighted in *italics*) are as follows:

3.2.2 Image Context

In order to model an image's context, we classify based on the time taken, the specified camera make, the GPS co-ordinates (if these exist) and web context. All of these features can be easily extracted from an image's meta-data tags (e.g. its exchangeable image file format) or using the Flickr API. We select the first set of features from the time an image is taken, due to the observed value of temporal information in traditional information retrieval (Alonso et al., 2011). We note, however, that camera timestamps are often set incorrectly with around 10% of photographs on Flickr more than 24 hours "incorrect" (Thomee et al., 2014). "Correcting" these timestamps is beyond the scope of this thesis and we therefore explore other image contexts, such as device metadata (e.g. device type, orientation *etc*) and online context (e.g. image view count) to alleviate this problem as these details are likely to be more reliable; additionally we proposed these features based on our analysis of user "tagging trends & tendencies" discussed in section 3.3.

1. **Time**: images are classified as either taken in the *morning* (06:00 to 11:59), *afternoon* (12:00 to 17:59), *evening* (18:00 to 23:59) or *night* (00:00 to 05:59).
2. **Day**: images are classified as either taken at the *weekend* (Fri-Sun) or on a *weekday* (Mon-Thu).
3. **Season**: images are classified as either taken in *winter*, *spring*, *summer* or *autumn*.
4. **Country**: images are classified as taken in a particular *country*. In total, 31% of images in our collection contain GPS data where images are taken in 219 distinct countries with the most popular countries being *USA*, *UK*, *Spain* and *Italy*.
5. **Device**: images are classified as taken on either a *mobile phone* or *camera*. This is achieved by manually classifying the 100 most popular camera makes (from an image's meta data) by an assessor into each category.
6. **Flash**: images are classified as flash *on*, *off* or *unknown*, based on whether the flash fired, inferred from the Exchangeable image file format (EXIF) meta-data.
7. **# Views**: we classify an image as having a *high*, *average* or *low* number of views on Flickr. In order to group images into these categories, we select (view count) thresholds such that the sets contain similar numbers of images.
8. **# Comments**: we classify an image as having a *high*, *average* or *low* number of user comments on Flickr. As before, we select thresholds so each set contains a similar number of images.
9. **Orientation**: images are classified based on their orientation which we compute based on the pixel size of an image, or specifically the relationship between their height and width. An image is either: *landscape* (i.e. width > height), *portrait* (i.e. height > width) or *square* (i.e. width = height).

3.2.3 Image Content

Although we hypothesise that an image's context is important for tag recommendation purposes, we do not underestimate the power of *visual features*. In this work, we also exploit an image's visual appearance in our tag recommendation model by using well known, high precision classification techniques; the value of computer vision has been demonstrated across various applications (Szeliski, 2010) and we hypothesise that these techniques will also provide additional signal for photo tag recommendation purposes. Specifically, we classify using state-of-the-art methods as described below:

1. **Scene #1**: images are classified to be one of the following scenes: *city*, *party*, *home*, *food* or *sports*. We classify images by training a multi-class support vector machine (SVM) on the popular image feature GIST (Oliva and Torralba, 2006), which has been used in the past to classify an image's scene with state-of-the-art performance. To build the relevant training collections, we use those 25k images which were manually classified, as one of the given scenes, by Mechanical Turk users for the ImageCLEF 2009 task. A full description of how these images were classified in the report by Nowak and Dunker (2010). From this, we train a multi-class SVM using 5-fold cross validation to classify all the images in our collection. Best performance is achieved using the Radial basis function (RBF) kernel with parameters $C = 2$ and $\gamma = 2^{-3}$, where 52.6% accuracy for classifying for the 5 scenes is achieved.
2. **Scene #2**: images are also classified to be one of the following scenes: *indoor*, *outdoor*, *macro* or *portrait*. To classify images, we use the same process as before, training a multi-class SVM on the GIST feature extracted from those ImageCLEF 2009 images manually classified as one of the given scenes. Best performance is achieved using a Radial basis function (RBF) kernel with parameters $C = 3$ and $\gamma = 2^{-7}$. where 47.4% accuracy for classifying for the 4 scenes is achieved.
3. **# Faces**: using the popular HAAR Cascade methods of face detection (Viola and Jones, 2001), we count the number of faces in each image, classifying as: *0*, *1*, *2* or *3+*.
4. **Colour**: we also attempt to extract an image's dominant colour. Specifically, we classify images as being prominently *white*, *black*, *red*, *green* or *blue* based on the most popular colour in an image. This is implemented by averaging the RGB values of each pixel in an image, and selecting the colour with minimal Euclidean distance from each of the pre-defined colour's RGB values (e.g. black = (0,0,0)).

3.2.4 User Context

Finally, we attempt to draw further evidences from the photographer themselves as we hypothesise details of the user may give insight to the semantics of their photographs. There exists work in computational linguistics which has shown vast differences in the language used by different population sets and demographics (Rayson et al., 1997); therefore, we hypothesise that by exploiting information regarding the user (such as

their gender) we will offer tag suggestions which are more relevant. In order to achieve this, we specifically classify images based on the user for which an image is taken by, based on details from their Flickr profile. We classify each image for the following categories:

1. **Gender**: as Flickr does not disclose a user's gender, we infer this by classifying based on their first name, if this exists. To do so, we collect publicly available 1990 US census data[2] detailing the most popular male and female names - we employ this US name database due to the mostly American demographic of Flickr users. A user is classified as *male* or *female* if their first name exists in each of these lists, or otherwise, *unknown*. In the case where a name is unisex (e.g. Stacey), we use the gender which is most popular for the name in question.
2. **Account**: we classify users based on whether they have a *pro* or *free* Flickr account. A paid subscription to a pro Flickr account offers more storage space and no advertisements. We use this feature to weakly infer whether a photographer is a professional or a hobbyist.
3. **# Images**: we classify a user as having uploaded a *high*, *average* or *low* number of photographs to Flickr. In order to group users into these categories, we select (upload count) thresholds such that the sets contain a similar numbers of users.
4. **# Contacts**: we classify a user as having a *high*, *average* or *low* number of Flickr contacts. We select thresholds so each set contains a similar number of users.

3.2.5 Evidence Combination

In our work we compare the *individual* effectiveness of these features for the photo tag recommendation task in order to determine their value. However, in order to exploit latent relationships as well as combine their value for photo tag recommendation purposes we also propose a combination approach as described in the following.

Problem Formalisation In our approach, we output a single vector, O_q where each element uniquely references every tag in our collection, for every feature f; in our work 17 different O_q vectors are output for every image. In evidence combination the goal is to most effectively consider, or "combine", the suggestions made based on every feature so that we maximise tag recommendation performance. Alternatively, given a $f \times k$ matrix, the goal is to output a $1 \times k$ recommendation vector which considers all features. The following sections detail our feature combination approach for this purpose.

Combination Approach In this work we combine using *weighted* (WLC) and *non-weighted* (NLC) linear combination approaches. For an $f \times k$ matrix, we compute the

[2]http://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html - last accessed on 1st July 2016.

$1 \times k$ vector by calculating the sum of each matrix column. For NLC, we simply take the average score for each tag, over all vectors. For WLC, we first weight each row by its individual recommendation "performance". Specifically, each weight is computed as the proportional tag recommendation improvement (using *precision at five* as a measure) for each feature over the baseline (TF-IDF) which considers global tag co-occurrence values, when testing on a validation set of 500 images. These weights are normalised to sum to one.

3.3 Tagging Trends and Tendencies

To further motivate the exploitation of the proposed feature for tag recommendation purposes, we delve further into their trends and tendencies as described in the following sections.

Image Context The *context* an image is taken in gives strong evidence to the likely contents of an image without considering the pixels themselves. For example, we can observe from Figure 3.3 that an image taken at 1am is seven times more likely to be tagged with `party` than at 10am. This rises to over 16 times on a Saturday. Similarly, many tags also have a geographical significance (e.g. `eiffeltower`) as identified by existing works (Rattenbury et al., 2007; Zhang et al., 2012b). For example, Rattenbury et al. (2007) automatically extracted *place* and *event* tags from geo-tagged Flickr images using naïve burst detection methods from signal processing literature (Vlachos et al., 2004), where an event tag was any term which exhibited significant *temporal patterns* and a place tag was any term which exhibited significant *spatial patterns*. Similarly, Zhang et al. (2012b) attempted to cluster tags based on geolocation and temporal trends allowing for the construction of tag cluster visualisations. These works, however, only consider a small part of an image's context (i.e. geographical and temporal factors). In this work, we explore far beyond these evidences by proposing a multitude of new contextual image features for our purpose, extracted from aspects such as: (i) the device type, (ii) the orientation *etc*.

For example, even the seemingly trivial feature of image *orientation* can offer insight into an image's semantics. In Table 3.7[3], we compute the most *significant* tags, sorted by descending order, for each feature; this is computed as the fraction of images tagged with tag t in images classified as x, minus the fraction of images tagged with t in all images. Other more elaborate types of measure could have been used such as TF-IDF, log-likelihood or Chi-squared (Kilgarriff, 2001) in order to reduce noise, however, we believe using raw frequency to be sufficient given the large size of our MIR-FLICKR 1M collection containing 9.85M annotations.

By calculating these values we identify the tags which occur significantly more in

[3]This figure is located at the end of this chapter due to space constraints.

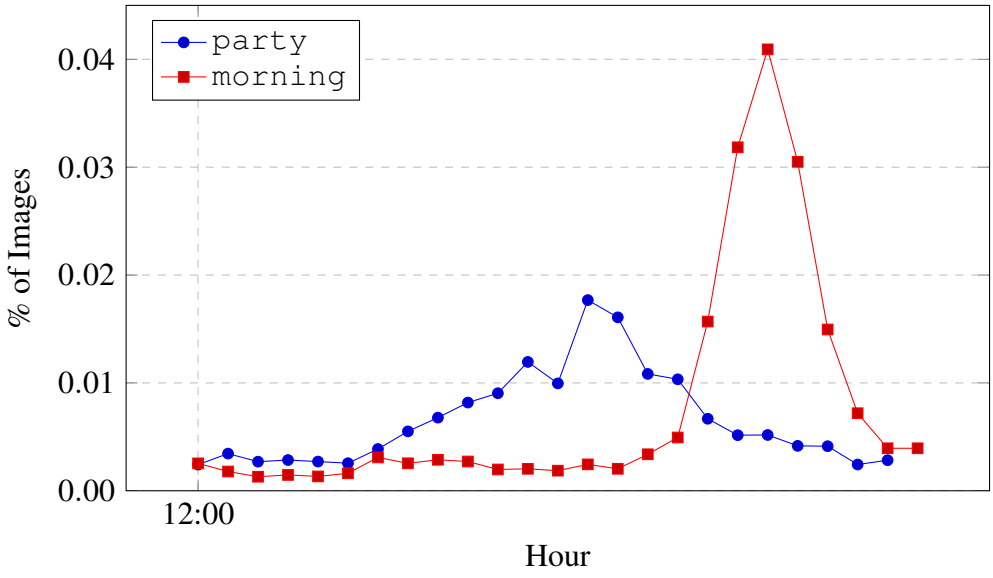

Fig. 3.3 Tagging temporal relationships

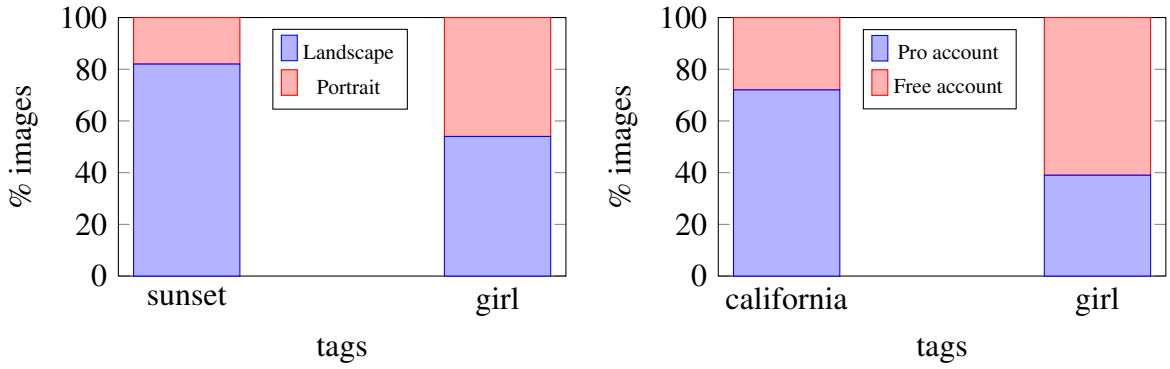

Fig. 3.4 *(Left)* % images taken in orientations. *(Right)* % images taken for account types.

a subset of images sharing a common feature, than in the global set. It can therefore be observed from Table 3.7 (and the left graph of Figure 3.4) that there exists a strong relationship between an image's orientation and the tags it is annotated with. For example, concepts which are more suitable for a wider shooting view (e.g. `sunset`) are used significantly more in photographs shot in landscape orientation. It may be argued that these contexts may not be the *sole reason* that we exhibit different tagging patterns and that there may be other latent features at work; this argument cannot be fully rejected despite our intuition from extensive manual inspection of the tags produced and analyses proposed in this section. In any case, whether the contexts proposed are the sole reason, or not, is irrelevant for our purpose - what *is relevant* is that there is a clear tagging distribution difference from feature to feature and therefore valuable signal for tag recommendation purposes.

Similarly, the type of device an image is taken on also influences the probable scene of an image, as can be observed in Table 3.7. For example, cameras are generally used

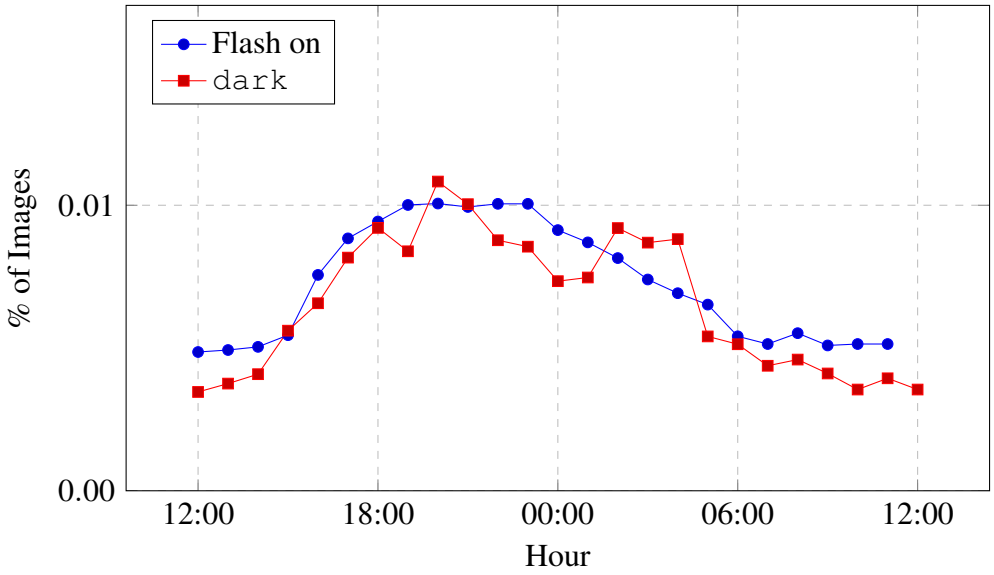

Fig. 3.5 Camera flash temporal trends

more for nature shots (i.e. `nature`, `flower`) than mobile phones which are more adopted in social scenarios. Finally, whether the flash fired influences the probable scene of an image, as detailed in Figure 3.5, where the fraction of images where the flash fired vs the fraction containing the tag `dark` is shown to correlate. We hypothesise that these types of trends and tendencies are of value to tag recommendation models.

Image Content Aside from an image's context, its visual appearance can given insight into the tags an image is most likely to be annotated with. Based on the most significant tags for each visual scene in Table 3.7, we observe a *logical* correlation. For example, images classified as "city" are more likely to be annotated with `architecture` than images classified as other visual scenes. This coherent relationship is present for almost all of the visual scene types. Additionally, the number of faces present in an image offers insight into its scene, objects and context. From Table 3.7, we observe that images without a face are more likely to be of outdoor, nature scenes (i.e. `sky`, `sunset`, `sea`) than those with many faces (i.e. 3+), which are more likely to be of a social occasion (i.e. `party`, `girls`).

User Context Finally, we observe that evidences can also be drawn from the photographer themselves where simple demographics, such as gender, can result in distinct tagging tendencies, as is detailed in Table 3.1. Firstly, we observe that the most significant tags for each gender are very different, with men more interested in tagging an image's general location (i.e. `usa`, `nyc`, `california` *etc*) than women who are

[3] 'Flash on' has been scaled by 10 for presentation purposes.

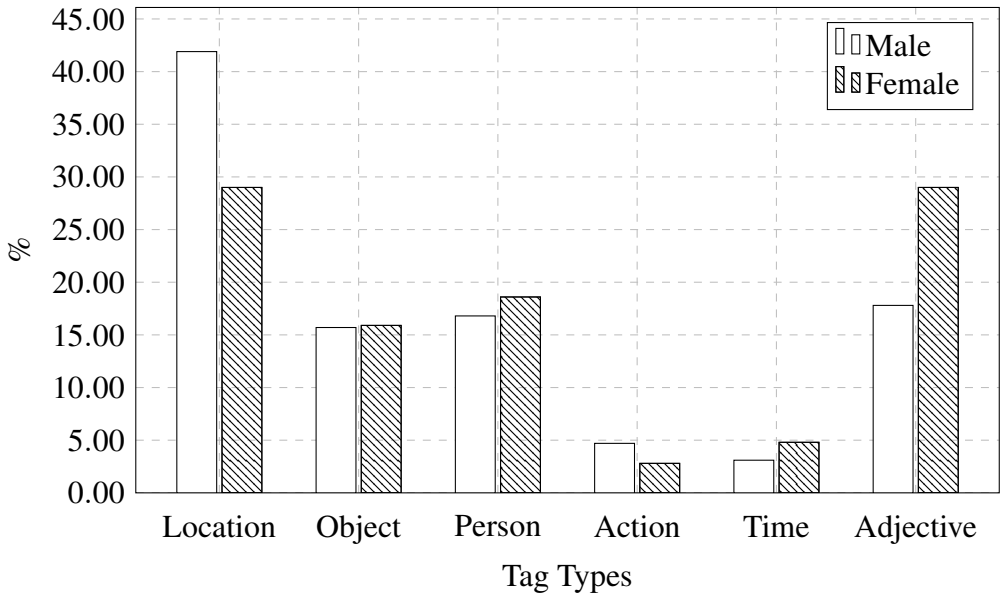

Fig. 3.6 Gender tagging trends

more likely to take part in Flickr challenges[4] (i.e. `365days`, `secondlife`) and take photos of themselves (i.e. `selfportrait`, `me`). Secondly, we categorise the 500 most significant tags for each gender, using WordNet (Miller, 1995), for the same categories used by Sigurbjörnsson and van Zwol (2008); the results of which are expressed in Figure 3.6. This categorisation reconfirms that men are more interested in tagging an image's location than women. Further, we observe that women are more interested in tagging an image with adjectives than men, with the most prominent being `pink` and `cute`. Additionally we note that work in computational linguistics (Rayson et al., 1997) has also observed that the language used by users correlates with gender with other works in information retrieval showing that gender can also be predicted from the author's text Burger et al. (2011). Given this, in our work we look to *flip* this correlation by instead predicting the user's *language* (i.e. tags) based on their *gender*.

Further, users whom own a Flickr pro account use different tag sets than those with free accounts, as seen in Table 3.7. Figure 3.4 highlights the difference in proportion of images taken by users with professional and free accounts for the tags `california` and `girl`. We observe that users with professional accounts are more likely to take photos of `california` than users with free accounts, perhaps reflecting the high salary of users based in California[5] (a professional account costs either $49.99 or $499.99 per year). Conversely, users with free accounts are more likely to take photos

[4]Users on Flickr are able to start challenges, or competitions, where there is some objective to be carried out by the user. For example, in the `365days` group https://www.flickr.com/groups/365days/, the objective is to post 1 image per day - last accessed on 18th July 2016.

[5]https://en.wikipedia.org/wiki/List_of_U.S._states_by_income - last accessed on 18th July 2016.

#	Male	Female
1.	california	selfportait
2.	usa	365days
3.	sanfrancisco	me
4.	unitedstates	pink
5.	america	italy
6.	hdr	cat
7.	night	sp
8.	newyork	secondlife
9.	nyc	food
10.	australia	love

Table 3.1 The most "significant" tags for each gender, where *significance* for an annotation is defined as the percentage of images tagged within a given subset (e.g. S_{male}), minus the percentage tagged in the full collection.

of people or themselves (i.e. `girl`), highlighting that these users use Flickr for more social purposes than those with pro accounts.

Summary As detailed in this section, there exist many "tagging trends and tendencies" for images which are taken in different contexts e.g. male *vs* female photographer, night *vs* daytime image *etc*. In particular, we have shown concrete examples where some aspect of the image's context has altered the annotation likelihood for certain tags e.g. images taken in landscape mode are much more likely to contain `sunset` than those taken portrait. It is these kinds of latent tagging tendencies which we look to exploit for photo tag recommendation purposes; detailing all of the underlying annotation tendencies would fill an entire thesis on its own, therefore, this section simply introduces the idea which is exploited in the rest of this chapter and thesis.

3.4 Experiments

In the following subsections we discuss details of our evaluation which compares our various approaches against 3 baselines. In particular, we detail our evaluation procedure in section 3.4.1 before discussing our baselines and experimental approaches in sections 3.4.2 and 3.4.3.

3.4.1 Evaluation Procedure

In this work, we evaluate against the tags assigned by the user of an image testing on the MIR FLICKR 1M dataset (as detailed in Section 2.3.2).

In our evaluation we approach the tag recommendation scenario as that of a ranking problem. Specifically, we compare the top ranked tags, when using various input types

(e.g. popular vs non-popular) and lengths (e.g. cold start or using multiple input tags). In order to address **RQ3.1** (see section 3.1), the first evaluation scenario selects N uniformly random tags from an image's tag set as a query to the given recommendation model. In order to address **RQ3.3**, we compare recommendations when suggesting using (i) the most popular tag in an image (ii) the least popular tag in an image (iii) the median frequency tag in an image (iv) one random tag in an image (v) no tags (i.e. a cold start).

For both approaches, the top five tags are retrieved and used as recommendations, with evaluation metrics computed against the *other* tags in an image's tag set. In particular, we compute precision (P@1, P@5), success (S@5) and Mean Reciprocal Rank (MRR) as discussed in Chapter 2.3.2. All of these metrics are commonly used and have been adopted by previous work in photo tag recommendation (Garg and Weber, 2008). Finally, we compute statistical significance paired t-tests comparing the experimental approaches against our *TF-IDF* baseline.

3.4.2 Baseline Systems

In our experiments, we evaluate performance against three baselines (BL). The first two consider textual evidence (i.e. $N > 0$) in the recommendation process, whereas the final baseline is only used in a cold start scenario (i.e. $N = 0$), as described below:

- **Related Tags (REL):** In this BL, for a given input tag, we use the recommendations made by Flickr's `tags.getRelated` API method as described in Section 2.3.2. We refer to this system as *REL* and consider it to be *industry standard* due to its use on the Flickr image sharing website and therefore by its millions of users.

- **TF-IDF Approach (TF-IDF):** This baseline, referred to as *TF-IDF*, recommends tags using the methodology described in Section 2.3.2, where tag co-occurrence measures are taken from the global tag co-occurrence matrix, C. This method was proposed in the popular photo tag recommendation work[6] by Garg and Weber (2008) which has subsequently been adopted by other authors (Takashita et al., 2010; Zangerle et al., 2013). Furthermore, the TF-IDF weighting scheme has been shown to outperform many state-of-the-art methods in other related areas of information retrieval (Spina et al., 2012; Trieschnigg et al., 2007; Wang and Manning, 2012) and as a result we consider this method as a *strong baseline*.

- **Popular Global Tags (POP):** Our final benchmark is a naïve baseline where we suggest the most popular tags in our collection. This baseline is only compared against in the cold start scenario i.e. where there exist no tags (to recommend upon) within an image. We refer to this system as *POP*.

[6]This work has been cited by 159 other works as of the 10th of July 2016.

3.4.3 Experimental Systems

In our experiments, we propose four methods for tag recommendation as explained in the following subsections. The first system considers N tags as well as a *single* feature:

- **Individual Feature Recommendation (f):** Recommendations are made using the TF-IDF methodology described in Section 2.3.2 replacing the tag co-occurrence matrix C, with the tag co-occurrence matrix C_f and the IDF vector with $IDF(S_f)$ i.e. those computed on images sharing the common feature f. We refer to this system using the *feature name* (f) as listed in Section 3.2 (e.g. *Time* is the approach which draws co-occurrence measures from the matrix C_{Time}).

As previously discussed, we attempt to combine the value of all the proposed features using two combination approaches, as discussed in Section 3.2.5, as denoted below:

- **Non-weighted Linear Combination (NLC):** Given recommendation vectors from *all* features, we use the non-weighted linear combination approach described in Section 3.2.5. We refer to this system as *NLC*.
- **Weighted Linear Combination (WLC):** Given recommendation vectors from *all* features, we use the weighted linear combination approach described in Section 3.2.5. We refer to this system as *WLC*.

Linear combination is a popular approach for combining predictions and is used within many recent recommender systems (He et al., 2010; Stern et al., 2009; Yang et al., 2007); aside from their effective performance, it is their ease of implementation & therefore comparability which makes them a desirable choice. Other more elaborate combination methods, such as those employing machine learning (Hou and Pelillo, 2013), have also been proposed in recent years, however, a comparative study of feature combination approaches is beyond the scope of this thesis and we leave the adoption of these models as future work.

Finally, we present a method which makes tag recommendations in a *cold start*, when $N = 0$, by making suggestions based on *all* features:

- **Cold Start Recommendation (CSR):** For a test image we compute $TF(S_f) \cdot IDF(S_f)$ for each classified feature $f_1 \dots f_{17}$, resulting in 17 (i.e. $|f|$) recommendation vectors. Given that the output of all these features form a $17 \times k$ matrix, we use the WLC combination approach, as discussed in the previous section, in order to combine the vectors and suggest tags i.e. by taking a weighted average of the predictions based on each feature. We refer to this system as *CSR*.

3.5 Results

In the following section, we analyse the effects image context, content and user context have on the tag recommendation process before discussing combination and cold

start methods. Specifically, we compute the proposed metrics against the top 5 and 10 ranked tags, as output from our recommendation approaches. Firstly, in section 3.5.1 we consider the performance of features individually (**RQ3.1**) before using a combination in section 3.5.2 (**RQ3.2**), when selecting N uniformly random tags from an image's ground truth as input to the tag recommendation model. We then consider the exploitation of the various features in a cold start scenario (**RQ3.3**) in section 3.5.3. As previously discussed, our research questions in this chapter are as follow:

RQ3.1 Can image context, image content and user context be exploited for photo tag recommendation purposes in order to complement existing textual evidences? Which features are most effective for this purpose?

RQ3.2 Can we capture the latent relationships between the discussed evidences to further improve recommendation accuracy?

RQ3.3 Can these features be used in a cold start recommendation scenario? Can we effectively exploit these features when the input tag(s) of an image are ill-defined or lacking in training data?

3.5.1 Individual Features

Considering *image context*, every feature (except country) improved recommendation accuracy by on average 19% (for P@5) over our TF-IDF baseline, as observed in Table 3.2. In particular, temporal features were the most discriminate with the *time* of day and *season* giving largest improvement (for P@5), highlighting the strong temporal trends present in tags. The features extracted from the camera's EXIF (i.e. device type and whether the flash fired) also resulted in a small recommendation improvement suggesting these aspects also influence the likely scene of an image. Finally, the number of image views and comments an image receives were also successfully exploited for tag recommendation purposes highlighting a correlation between an image's popularity and its tags; conversely certain tags attract much of the attention for a photograph. This is further supported by Table 3.7, where we observe that certain tags (e.g. `girl`, `portrait`, `woman`) are viewed significantly more often than others. For example, images tagged with `girl`, are viewed on average 3x more often than those tagged with `flower` (1,453 views vs 453 views). There also exists a similar relationship between tags and the number of comments they provoke. For example, images tagged with `nature` contain more than double the number of comments on average than those tagged with `2008` (19.2 comments vs 9.2 comments). Therefore, predicting whether an image will become popular in the future could be useful in order to promote tags which are generally associated with popular images; this avenue is explored in chapter 8.

On average, features computed from an image's *visual appearance* resulted in the highest recommendation accuracy, with each feature improving suggestions by 36%

	REL	**TF-IDF**	**Time**	**Day**	**Season**	**Country**	**Device**	**Flash**	**Views**	**# Com**	**Orient**
P@1	0.161	0.180	0.286**	0.253*	0.284**	0.192**	0.202*	0.184**	0.195**	0.195**	0.277 ***
P@5	0.159	0.182	0.272**	0.228**	0.266**	0.176**	0.196 ***	0.186**	0.186**	0.190**	0.251**
S@5	0.411	0.460	0.612**	0.583*	0.609**	0.434*	0.546**	0.478**	0.510**	0.507*	0.604*
MRR	0.280	0.291	0.404**	0.368**	0.399*	0.275**	0.323**	0.289**	0.308**	0.304**	0.395**

Table 3.2 Performance of *image context* features for PTR selecting $N = 3$ tags as *input*. The statistical significance results against the baseline (TF-IDF) are denoted as * being $p < 0.05$, ** being $p < 0.001$. Due to space constraints: # Com = # Comments, Orient = Orientation.

(for P@5) on average over our TF-IDF baseline, as observed in Table 3.3. Specifically, the dominant colour in an image was the most discriminative feature outperforming both state-of-the-art scene classification methods. We hypothesise that this is due to the number of users tagging images with high level colours in an image; in fact, 20% of the 50 most popular tags on Flickr are colours. Aside from colour and scene, the feature which counted the number of faces in an image also significantly improved recommendation accuracy further highlighting the connection between the number of people in an image and its annotations.

	REL	**TF-IDF**	**Scene 1**	**Scene 2**	**# Faces**	**Color**
P@1	0.161	0.180	0.304 ***	0.224	0.252 **	0.289 ***
P@5	0.159	0.182	0.270 ***	0.214 ***	0.225 ***	0.280 ***
S@5	0.411	0.460	0.615 ***	0.558 *	0.574 *	0.628 ***
MRR	0.280	0.291	0.414 ***	0.343 ***	0.366 ***	0.412 ***

Table 3.3 Performance of *image content* features for PTR ($N = 3$). The statistical significance results against the baseline (TF-IDF) are denoted as * being $p < 0.05$, ** being $p < 0.01$ and *** being $p < 0.001$.

Considering the *user context*, every feature improved recommendation accuracy by on average 31% (for P@5) over our TF-IDF baseline, as observed in Table 3.4. The feature which counted the number of contacts a user had produced the most discriminative signal for tag recommendation purposes. From further investigation, we find that users with many contacts have similar tagging tendencies to *other* users with many contacts (and similarly for users with few contacts). For example, users with many contacts annotate their images with 20% more tags and with far more popular tags than those with few contacts. Therefore, there exists a relationship between the number of contacts a user has and their tagging tendencies which are exploited in our model.

We can conclude that there exist a number of strong relationships between tags and image context, image content and user context, of which are captured and exploited

	REL	TF-IDF	Gender	Account	Images	Contacts
P@1	0.161	0.180	0.193 ***	0.246 **	0.285 ***	0.290 ***
P@5	0.159	0.182	0.185 ***	0.230 ***	0.257 ***	0.279 ***
S@5	0.411	0.460	0.468 ***	0.570 *	0.592 *	0.627 ***
MRR	0.280	0.291	0.288 ***	0.363 ***	0.395 ***	0.414 ***

Table 3.4 Performance of *user context* features for PTR purposes ($N = 3$). The statistical significance results against the baseline (TF-IDF) are denoted as * being $p < 0.05$, ** being $p < 0.01$ and *** being $p < 0.001$.

in our recommendation approach. Overall, the time an image is taken, its high level scene/colour and the user's popularity/activity prove to be the most effective features for this purpose (**RQ3.1**).

3.5.2 Evidence Combination

In the following section we address **RQ3.2**, analysing the effects of evidence combination. As observed in (a) of Table 3.5, by considering *all* features using linear combination approaches the highest recommendation accuracy is achieved (where P@5 is 14% higher than the best performing individual feature). In particular, using a simple weighted combination we achieve *double* the recommendation accuracy (for P@5) in comparison to our REL baseline and improvement of up to 75% (for P@5) over our TF-IDF baseline. This highlights that image context, content and user features are complementary in the recommendation process. Thus for **RQ3.2**, we can conclude that evidences contain latent relationships which can be combined in order to further benefit tag suggestions.

	REL	TF-IDF	NLC	WLC
P@1	0.161	0.180	0.331 ***	0.340 **
P@5	0.159	0.182	0.314 ***	0.320 ***
S@5	0.411	0.460	0.658 ***	0.667 ***
MRR	0.280	0.291	0.452 ***	0.463 ***

(a) Feature Combination (N=3)

	POP	CSR
P@1	0.050	0.068 ***
P@5	0.045	0.066 ***
S@5	0.110	0.130 ***
MRR	0.075	0.085 ***

(b) Cold Start (N=0)

Table 3.5 Combination and cold start approaches. The statistical significance results against the baseline (i.e. (a) TF-IDF (b) POP) are denoted as * being $p < 0.05$, ** being $p < 0.01$ and *** being $p < 0.001$.

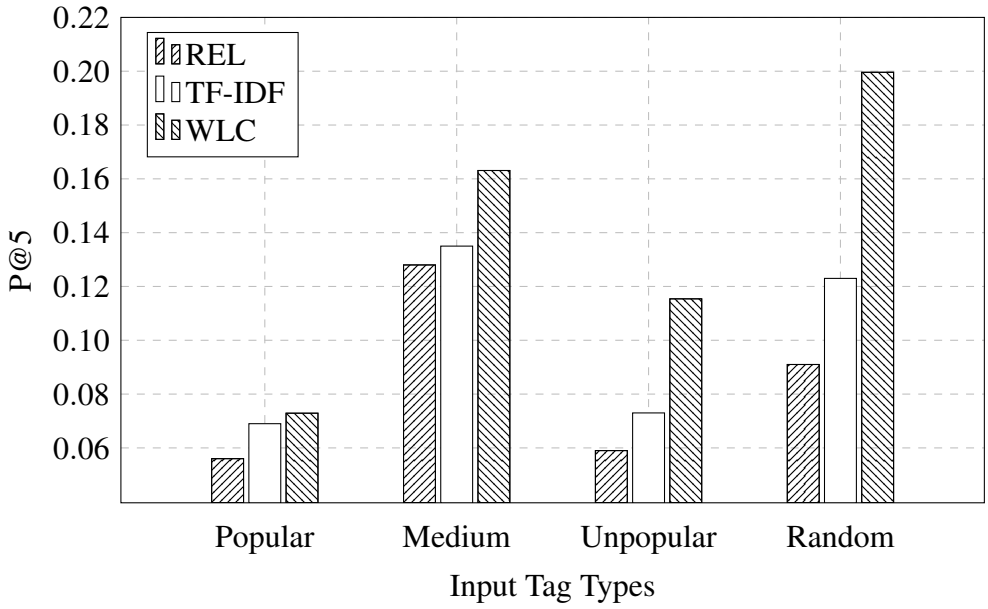

Fig. 3.7 Recommendation performance (N=1) for different input tag types

3.5.3 Cold Start Recommendation

In the following section we address **RQ3.3**, analysing the effects of recommendation performance in a cold start situation. First we consider the situation where a single tag is used for recommendation. Figure 3.7 shows the recommendation performance for the various input types discussed in Section 3.4.1. We observe that we are able to significantly out perform both baselines by using an evidence combination approach for each input type. The greatest proportional increase is observed for *unpopular* tags where recommendation accuracy is more than double that of the REL baseline. Crucially, the evidences proposed can be exploited in the situation where "bad" tags (i.e. those lacking in training data) are used as input. The effect is not as significant for popular input tags, however; we hypothesise this is because popular tags occur in the majority of images and therefore do not correlate strongly with any feature in particular.

Finally, from Figure 3.8 we can observe that by exploiting a combination of the proposed features we are able to significantly outperform the POP baseline in a cold start scenario. In fact, the effectiveness of the proposed features achieves recommendation accuracy that is comparable in the case of recommendation using 1 popular tag (see Figure 3.7). Therefore we can conclude for **RQ3.3** that the proposed features can be employed without the presence of any textual evidences alleviating the problems associated with annotation in a cold start scenario.

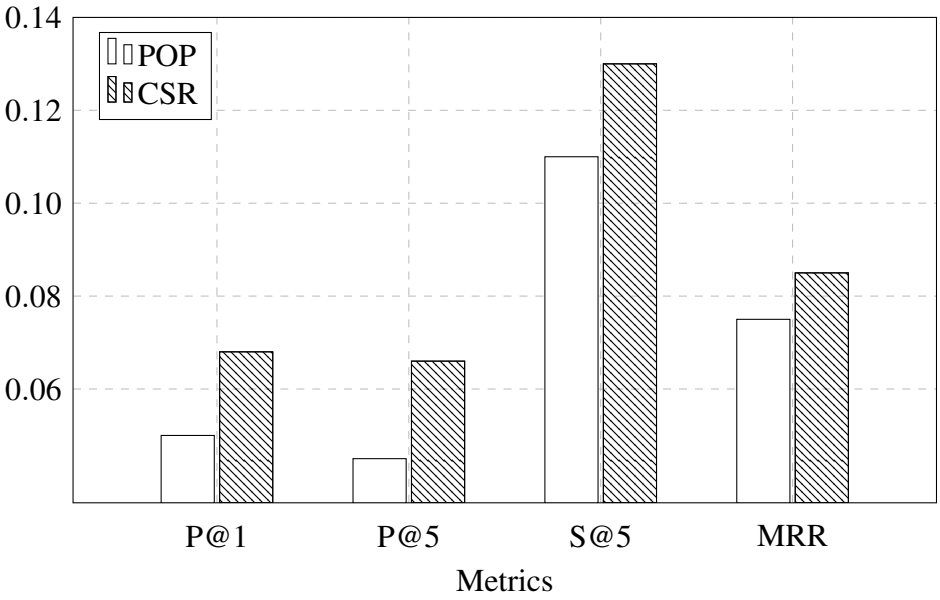

Fig. 3.8 Cold start recommendation performance (N=0)

3.5.4 Manual Inspection

In the following section we manually analyse the recommendations made for an example image from our test collection, highlighting the differences in the suggestions made by each system. Table 3.6[7] summarises the details of this image: its user tags, the automatic classifications we have made, the inputs used for recommendation, the recommendations themselves, and the computed P@5 scores for each method.

The input tag `sharks` is somewhat ill-defined as the term is often used in other contexts, which do not relate to the animal, such as sports team names e.g. the ice hockey team "San Jose *Sharks*". Therefore, this query is ambiguous, resulting in many incorrect hockey related tag suggestions (e.g. `nhl`, `hockey`). Given this, there is an inherit motivation to reduce this ambiguity in order to increase the relevance of recommendations returned to the user; in this work, we achieve this by exploiting an image's context, visual appearance and user context.

The systems: *Time, Country, # Comments, Scene #1* and *Faces* are able to resolve this ambiguity by removing hockey related tags from the top five suggestions (these rows are highlighted in red). The time of day successfully achieves this, as NHL ice hockey matches are usually played in the evening[8], whereas this image was taken in the afternoon. As the image is taken in Australia (i.e. where ice hockey is not a popular sport), *Country* is able to remove the ice hockey tags from the recommendation list as we would expect the co-occurrence of `hockey`/`nhl` with `sharks` to be low or non-existent. It is not as immediately obvious why considering high number of comments removes the hockey tag suggestions, but on closer inspection we see that `predator`

[7]This figure is located at the end of this chapter due to space constraints.
[8]https://www.nhl.com/schedule - last accessed on 18th July 2016.

(the most promoted tag in *# Comments*) invokes more user comments on average (13.2 comments per image) than `nhl` (3.2 comments per image) and `hockey` (3 comments per image). Therefore, as the image has a high number of user comments, `predator` is promoted over `hockey` and `nhl`.

All of these discussed features fall into the image *context* category; however, the visual appearance of an image also plays an important role in the tag disambiguation process. For example, *# Faces* resolves this ambiguity as the image in question contains no (human) faces; on the contrary, we would expect an image taken at an ice hockey match to contain many faces. Further investigation confirms our hypothesis where 34% of images tagged with `nhl` contain at least one face, whereas only 11% of images tagged with `shark` contain at least one face.

Finally, the combination of all features (i.e. *NLC* and *WLC*) produces very diverse lists of popular tags, covering various distinct aspects, such as: (i) location (i.e. `australia`) (ii) ice hockey (i.e. `nhl` and `hockey`) (iii) shark related tags (i.e. `wildlife`) and (iv) camera meta-data (i.e. `canon`). In this particular example, this does not result in the highest recommendation performance, although we would expect over all images in our test set that suggesting a *diverse* list of *popular* tags to be a successful recommendation strategy, and is the reason behind the high P@5 measure.

3.6 Chapter Summary

In this chapter we introduced 17 novel evidences for the purposes of photo tag recommendation. These features were computed from the image context, image content and user context, covering a wide range of aspects from the photo shooting process, such as the time taken, orientation, number of faces present *etc*. By exploiting these features individually, as well as in a combination, we were able to significantly outperform two existing photo tag recommendation baselines when testing on the MIR-FLICKR 1M collection.

We conclude that temporal, orientation, high level scene/colour and a user's online presence to be the most effective features for exploitation in a photo tag recommendation system. Ultimately, however, by combining these features and exploiting the latent relations between them, highest accuracy is achieved. This work also analysed the effects of exploiting these features in a cold start and when the input tag is vague (i.e. very popular) or sparse (i.e. very unpopular). We conclude that the exploitation of these features are most effective when the input tags are unpopular or of medium frequency. Also, we conclude that a combination of the discussed features can be used to recommend tags without the presence of any textual evidence. Finally, we showed through detailed analyse of the recommendations that by considering these features, we were able to alleviate the tag ambiguity issue associated with tag recommendation models.

The results of this work highlight that even seemingly irrelevant contextual evi-

dences (e.g. whether the flash fired) can be used as evidence to improve a system's performance. In addition, many of these features are very lightweight to compute (e.g. time, orientation *etc*) and can be extracted from a multitude of content types aside from photography. For example, all of the discussed features could also be extracted from video, with many suitable for any kind of content (e.g. time *etc*). Therefore, future works should not overlook simple contextual features in favour of computationally expensive content-based approaches as often they can achieve similar (or better) performance, or can be combined to improve upon existing features. Finally, the methodology proposed in this chapter, where tag co-occurrence matrices are computed on a *subset* of semantically similar documents (e.g. images taken by the same gender photographer), could easily be generalised to other application and scenarios.

Recommendations for an Example Image

User Tags

`5d australia breach canon canon5d danger eos eos5d jaws greatwhite greatwhites greatwhiteshark hunter shark sharks water`

Classifications

Time:	Afternoon	*Day:*	Weekday
Season:	Summer	*Country:*	Australia
Camera:	Camera	*Flash:*	unknown
# Views:	High	*# Comments:*	High
Orientation:	Landscape	*Scene 1:*	Food
Scene 2:	Outdoor	*# Faces:*	0
Colour:	Green	*# Gender:*	Male
Account:	Free	*# Images:*	Low
# Contacts:	Low		

Tag Recommendation Inputs

`hunter, sharks, eos`

System	Top 5 Recommendations	P@5
REL	canon,cat,fish,nature,bird	0.2
TF-IDF	canon,nhl,hockey,predator,upclose	0.2
† Time	canon,shark,400d,aquarium,wildlife	0.6
Day	canon,nhl,hockey,predator,upclose	0.2
Season	canon,shark,camden,league,jersey	0.4
† Country	australia,canon,5d,shark,wildlife	1
Device	canon,nhl,predator,upclose,hockey	0.2
Flash	canon,nhl,hockey,predator,upclose	0.2
# Views	canon,nhl,pretty,hockey,taiwanese	0.2
† # Comment	canon,predator,upclose,wildlife,curious	0.4
Orientation	canon,nhl,hockey,sanjose,burns	0.2
† Scene 1	canon,upclose,wildlife,predator,animals	0.4
Scene 2	canon,nhl,upclose,predator,wildlife	0.4
† # Faces	canon,predator,upclose,wildlife,animals	0.4
Colour	canon,wildlife,nhl,nature,australia	0.6
Gender	canon,nhl,hockey,400d,australia	0.4
Account	burns,nhl,canon,hockey,sanjose	0.2
# Images	canon,shark,sanford,canon5d,icehockey	0.6
# Contacts	nhl,canon,450d,canoneos450d,playoffs	0.2
NLC	canon,nhl,wildlife,hockey,australia	0.6
WLC	canon,nhl,wildlife,hockey,australia	0.6

Table 3.6 Comparison for an example test image. The rows which begin with † resolve the tag ambiguity problem.

f	Class	Most Significant Tags
Time	morning	nature, sunrise, morning, flower, bird, snow
	afternoon	nature, flower, macro, winter, london
	evening	night, sunset, lights, light, film, longexposure
	night	film, night, vintage, party, newyork, america
Day	weekend	canon, nikon, 2008, people, girl, festival
	weekday	film, 365days, selfportrait, 365, me
Season	winter	winter, snow, christmas, ice, 2007
	spring	spring, flower, 2009, flowers, macro
	summer	summer, 2008, festival, flower, beach
	autumn	autumn, fall, leaves, halloween, november
Country	USA	usa, california, unitedstates, sanfrancisco, ny
	UK	london, uk, england, geotagged, scotland
	Spain	spain, espana, barcelona, catalunya, madrid
	Italy	italy, italia, toscana, tuscany, mare
Dev	Mobile	cameraphone, iphone, mobile, moblog, nokia
	Camera	nikon, canon, nature, flower, macro
Fla	On	macro, cat, cute, party, flash
	Off	canon, flower, eos, usa, unitedstates
#V	High	girl, portrait, woman, hdr, explore
	Low	flower, macro, cat, flowers, nature
#Cm	High	abigfave, aplusphoto, anawesomeshot, nature
	Low	2008, california, art, graffiti, sanfrancisco
Or	Landscape	macro, sunset, car, flower, water
	Portrait	portrait, girl, polaroid, woman, northcarolina
Scene 1	city	sky, architecture, hdr, sunset, night, city
	party	portrait, fisheye,smile, girls, rock, lomo
	home	portrait, flower, people, girl, selfportrait
	sports	sky, beach, sea, water, clouds, sunset, blue
	food	food, macro, flower, cat, portrait, red
Scene 2	Indoor	portrait, selfportrait, 365days, me, girl, self
	Outdoor	sky, sunset, water, clouds, hdr, landscape
	Portrait	portrait, 365days, me, selfportrait, girl
	Macro	macro, food, flower, portrait, rose, bokeh
# Faces	0 faces	sky, sunset, clouds, water, nature, landscape
	1 face	portrait, girl, woman, people, selfportrait
	2 faces	portrait, girl, me, people, selfportrait
	3+ faces	portrait, me, party, selfportrait, girls
Colour	Red	red, food, cat, orange, yellow, pink
	Green	green, nature, macro, flower, garden, grass
	Blue	blue, sky, clouds, water, sea, landscape
	Black	night, light, bw, lights, canon, sunset
	White	polaroid, snow, vintage, winter, beach
Gen	Male	california, usa, sanfrancisco, unitedstates
	Female	selfportrait, 365days, me, pink, italy
Acc	Pro	california, usa, 2008, sanfrancisco, geotagged
	Free	portrait, girl, secondlife, aplusphoto, wildlife
#I	High	2008, usa, california, sanfrancisco
	Low	nikon, bw, canon, hdr, portrait
#Cn	High	film, usa, california, bw, nikon
	Low	postcard, geotagged, algarve, ephemera

Table 3.7 Most significant tags for each feature, where *significance* for an annotation is defined as the percentage of images tagged within a given subset (e.g. S_{male}), minus the percentage tagged in the full collection. Due to space restrictions: Dev = Device, Fla = Flash, #Cm = # Comments, Or = Orientation, #I = # Images, #Cn = # Contacts.

Chapter 4

External Evidences for Photo Tag Recommendation

The following chapter is published in the following conferences:

- Philip J. McParlane, Joemon M. Jose (2014); Exploiting Twitter and Wikipedia for the Annotation of Event Images, *Proceedings of ACM Special Interest Group on Information Retrieval (SIGIR) 2014, Gold Coast, Australia.* ACM New York, NY, USA, pp1175-1178.
- Philip J. McParlane, Joemon M. Jose (2014); A Novel System for the Semi Automatic Annotation of Event Images, *Proceedings of ACM Special Interest Group on Information Retrieval (SIGIR) 2014, Gold Coast, Australia.* ACM New York, NY, USA, pp1269-1270.

4.1 Introduction

In the previous chapter, we focused on exploiting *internal* evidences extracted directly from the image itself. However, out-with the isolation of an image and its immediate context there exists a larger *online context* which can also offer clues to the probable tags relevant for an image. For example: (i) is the given image taken at a large scale social event? (ii) if so, is there any related textual content online? (iii) and can we exploit this textual content for tag recommendation purposes? In the following chapter we consider these *external* evidences, defined as those *"which are collected from some other (textual) resource related to an image"* for photo tag recommendation purposes.

Firstly we exploit *text-based social media* microblogging streams (i.e. Twitter) in order to train our tag recommendation approaches. The main advantage of using this content primarily concerns the high *quantity* (i.e. coverage) and fast *speed*, in comparison to photographs uploaded to image sharing platforms. For example, over 600M

messages[1] and 215M images[2] are posted on average to Twitter every day; in comparison, (only) 80M & 1M photographs are uploaded to the most popular image sharing websites[3], Instagram[4] and Flickr respectively[5]. Figure 4.1 further highlights this problem for content posted about to the Austin City Limits 2012 three day music festival (see Chapter 2.3.2), which we define as any image or tweet posted/uploaded *during* the festival and containing a predefined keyword[6]; from this graph we can observe the large difference in volume for Flickr vs Twitter data.

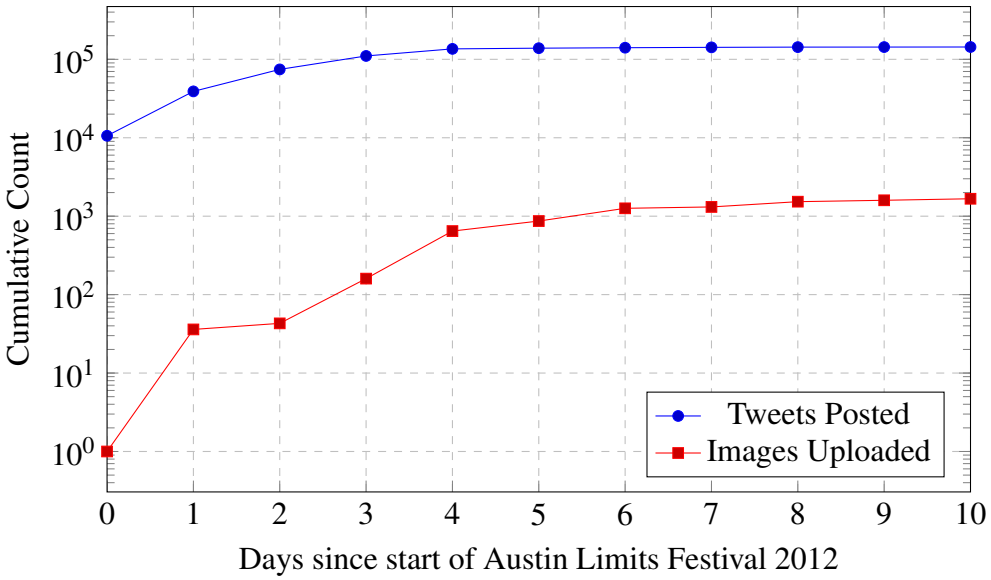

Fig. 4.1 Volume of ACL Tweets vs Flickr Images (note the logarithmic scale)

Aside from quantity, we also find that images posted to image sharing websites are often uploaded long after they are taken, meaning that the data is "out-of-date". For example, on Instagram over 48M and 27M images are tagged with `#latepost` or `#latergram`, respectively, implying that the image was uploaded some time after it was taken. After some investigation, we also observe this "time lag" problem on Flickr for our Austin City Limits collection where images are uploaded, on average, around 50 days later than they are taken. Figure 4.3 further emphasises this problem for the MIR-FLICKR collection where less than 1 in 3 images are uploaded within 24 hours

[1] http://www.internetlivestats.com/twitter-statistics/ - last accessed on 1st February, 2016

[2] http://goo.gl/CHmJVz - Based on 36% of tweets containing images, correct as of August, 2012.

[3] https://en.wikipedia.org/wiki/List_of_photo-sharing_websites - correct as of February, 2015

[4] http://www.internetlivestats.com/ - correct as of February, 2016

[5] http://goo.gl/OGPgyT - correct as of February, 2014

[6] `acl2012, acl2012acl, aclfest, aclfest2012, aclfestival, aclfestival2012, aclmusicfest, aclmusicfestival, aclmf`

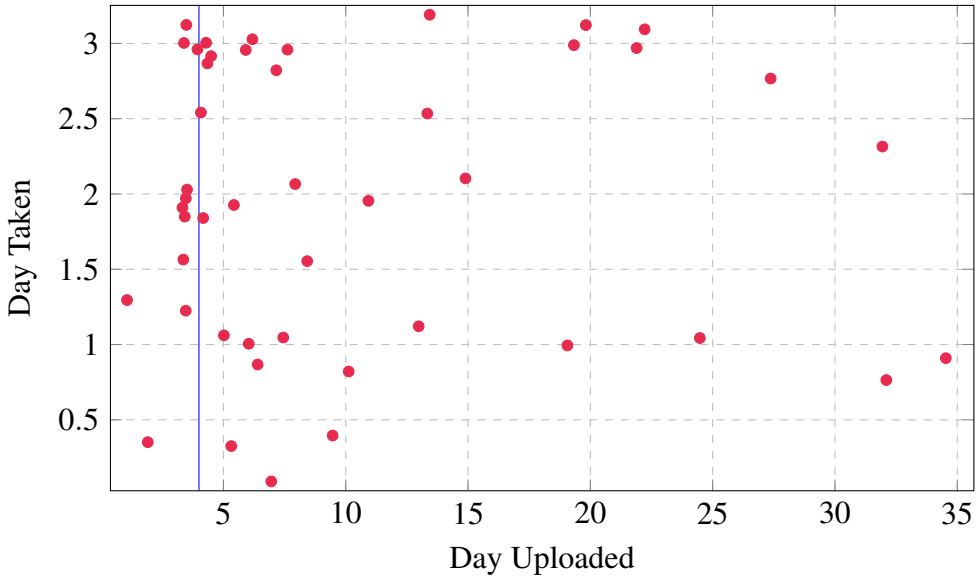

Fig. 4.2 Comparison of an image's taken vs upload time. Vertical line indicates the *end* of the three day festival.

of capture. We hypothesise that this tendency may happen for a number of reasons: (i) users may not be able to upload due to poor network coverage (ii) or even if they have network coverage, they might not want to upload to avoid paying expensive data charges (iii) the user may want to edit their photograph (e.g. adjust colour levels) on a laptop before uploading (iv) finally we note that timestamps are often set incorrectly (or not at all) on cameras, however, correcting these timestamps is beyond the scope of this thesis and has been addressed recently by Thomee et al. (2014) - in this work we accept this limitation and consider timestamps to reflect their true capture time.

Therefore considering the time lag problem, if we train tag recommendation models on this out-of-date content, the predictions will also be out-of-date due to the delay between users *taking* and *uploading* photographs. To overcome this problem, in this work we instead propose to build tag recommendation models based on fast moving Twitter data.

One problem with using Twitter data for recommendation purposes, however, concerns the *low quality* (e.g. informal language) and amount of *noise* present within streams, for example: (i) conversational chatter (ii) advertisements (iii) bot posts *etc*. We hypothesise that this aspect of microblog post may have a detrimental effect for tag recommendation purposes. Wikipedia, in comparison, offers a structured data source containing less irrelevant content, whilst maintaining fast update speeds (Osborne et al., 2012). Due to the curated nature of Wikipedia, in this work we also propose to use its content as a more reliable source of information in order to "counter" the noisy nature of Twitter for photo tag recommendation purposes. In this work, we propose to address the following research questions:

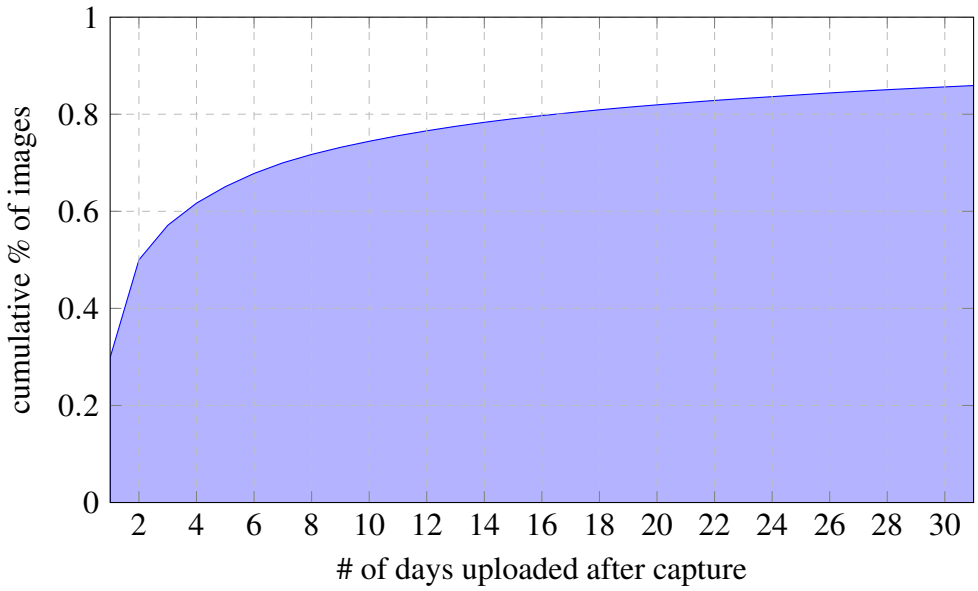

Fig. 4.3 Time lag problem: how many days after an image is taken is it uploaded to Flickr

RQ4.1 Can noisy social media streams, such as Twitter, be exploited in order to annotate images online? How can we alleviate the amount of noise within these streams?

RQ4.2 Can Wikipedia content also be exploited in order to offer reliable photo tag recommendations?

The rest of this chapter is as follows: Section 4.2 details our recommendation methodology whilst we discuss our evaluation procedure in Section 4.3. Section 4.4 details the findings of our results before summarising in Section 4.5.

4.2 Methodology

In the following section we detail how we exploit Twitter and Wikipedia data for the purposes of photo tag recommendation. Firstly, we discuss the specific task of annotating images taken at social events in section 4.2.1 before detailing how Tweets and Wikipedia articles are retrieved and exploited for our task in sections 4.2.2 and 4.2.3.

4.2.1 Annotating Event Images

Photographs taken at social and world events present an interesting challenge for annotation models as there exists much evidence from many disparate sources (i.e. Tweets, Flickr images and Wikipedia articles). Given the amount, varying quality and types of data present, there are many challenges regarding its exploitation for PTR purposes. The following sections detail the challenges and exploitation of each data source.

4.2.2 Twitter Data

Using Twitter data presents a number of challenges for tag recommendation; the largest problem being that of *noise* where tweets are *short*, contain *misspelt* words, *colloquial* expressions and often *irrelevant* information. In order to overcome the problem of irrelevant content, we consider only those tweets containing predefined tags (as detailed in the introduction of this chapter) which refer to the event in question. We address the identification of hashtags for an event manually as this is not the purpose of this work; however, in a real world scenario, we would rely on an event detection model, such as the state-of-the-art approach described by McMinn and Jose (2015), in order to cluster tweets and identify relevant event hashtags.

Using this approach we are able to address noise from a tweet *topic* relevance perspective, but not from a tweet *content* perspective. In order to suggest tags which are relevant for *images* we firstly conduct part-of-speech (POS) tagging, using the popular Stanford Parser (Manning, 2011). This POS tagger, which claims to have near perfect accuracy (Manning, 2011), identifies the *term types* (e.g. noun, adjective *etc* - specified in Table 4.1) for a body of text. Applying this model to those terms in each of our collection (see figure 4.4) we firstly observe that Flickr images are mostly annotated with nouns and entities. In comparison, tweets have the highest fraction of verbs (with the most common being `be`, `have` and `get`), highlighting the active nature of the users and the informal content of their posts. Further, there are more terms in the Twitter collection which cannot be identified (i.e. those classified as other) also likely due to the informal nature of the language on social media (e.g. misspellings, emojis *etc*). We highlight, however, a limitation of our work in that the part-of-speech tagger is trained on formal text from news articles and suggest that future works consider POS taggers which are trained on informal language collections (Owoputi et al., 2013).

As nouns and entities account for the vast majority of tags on Flickr (77%), photo tag recommendation systems should bias towards suggesting these term types in order to match the annotations used by users. Therefore, in our approach we only use nouns & entities from Twitter and Wikipedia for recommendation purposes, thus motivating our work on part-of-speech tagging.

Category	**Stanford POS Categories**
Noun	NN, NNS
Entity	NNP, NNPS, Person, Location, Organization
Verb	VBD, VBN, VB, VBG, VBZ
Adjective	JJ, JJR

Table 4.1 Part of speech term type classifications

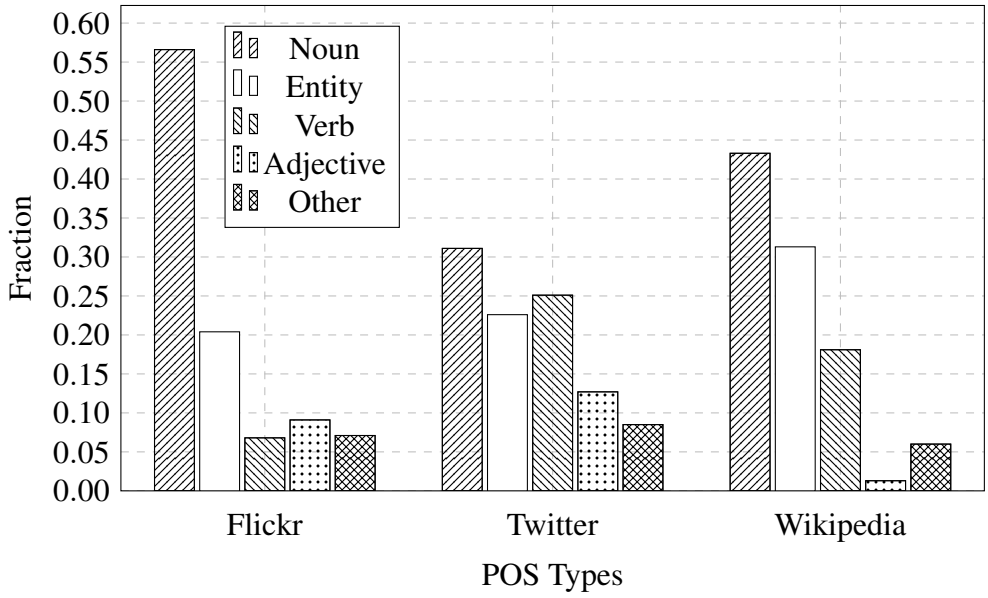

Fig. 4.4 Fraction of term types per collection.

4.2.3 Wikipedia Data

Our approach assumes we are able to identify the relevant Wikipedia article for an event in question in order to exploit its textual content for recommendation purposes. We achieve this process automatically by using the Wikipedia API's URL resolving function[7]. Using this API call, queries (e.g. `ny`) are resolved to the relevant article URL (e.g. https://en.wikipedia.org/wiki/New_York). Therefore, we employ the same process for our problem by querying this API function for the various hashtags used by the ACL music festival - as introduced in the first section of this chapter. In our scenario, the `ACLMF` hashtag matches against the `http://en.wikipedia.org/wiki/ACLMF` URL redirect for the Austin City Limits music festival Wikipedia article. As before, we classify each term within this Wikipedia article using the described Stanford POS tagger. From Figure 4.4 we observe that Wikipedia offers a more factual representation of an event, detailing more nouns & entities (75%) and less adjectives & verbs than Twitter. Due to these different characteristics, we hypothesise that a combination of all sources will achieve highest topic coverage and ultimately highest recommendation performance.

4.3 Experiments

In the following section we detail the various evaluation and baseline systems before discussing our evaluation procedure.

[7]https://www.mediawiki.org/wiki/API:Query#Resolving_redirects - last accessed on 10th July 2016.

4.3.1 Systems

In this work, we compare recommendations computed from Twitter and Wikipedia (as well as a combination), against a naïve and an industry strength tag recommendation baseline. Firstly, we introduce our systems which offer suggestions from a *single* source:

1. **Flickr(POP):** firstly, we compare against a naïve baseline which suggests the most popular tags on Flickr. We propose this baseline to replicate the cold start scenario where we have no training data related to a new event (or an image initially contains no tags) to make suggestions upon. This baseline is equivalent to the *POP* baseline introduced in Chapter 2.3.2.

2. **Flickr(REL):** secondly, we use an industry strength baseline by using those tag recommendations made on the Flickr website. Specifically, we consider the top tags as suggested from the `getRelated` API method[8], for the input tag, `acl`. This baseline is equivalent to the *REL* baseline introduced in Chapter 2.3.2.

3. **Twitter(T/TP):** in our first Twitter approach, we suggest the most frequent terms within the related stream of Tweets, referred to as *T*. In our second approach, we employ POS tagging by suggesting only the most frequent nouns and entities from this stream in an attempt to reduce noise and improve recommendation accuracy, referred to as *TP*.

4. **Twitter(TF-IDF(N)):** to determine whether fast moving microblogging streams can also be employed for the construction of tag co-occurrence matrices and the "time lag" problem associated with Flickr images avoided, we propose a TF-IDF tag recommendation strategy similar to the approach detailed in chapter 2.3.2. This strategy builds a term co-occurrence matrix C based on those terms present in the related Twitter stream for the given event. Given N random tags from an image, we are able to make suggestions using the TF-IDF model where each tweet is considered as a document and $C_{t_{ij}}$ counts the co-occurrence of two terms t_i and t_j in the tweet stream. We refer to this system as *TF-IDF(N)*.

5. **Wikipedia(W/WP):** in our first Wikipedia approach, we suggest the most frequent terms within the related Wikipedia article, referred to as *W*. In our second approach, we employ POS tagging by suggesting only the most frequent nouns and entities (WP) in an attempt to reduce noise and improve recommendation accuracy, referred to as *WP*.

Secondly, we combine recommendations from Twitter and Wikipedia using the following methods:

[8] www.flickr.com/services/api/flickr.tags.getRelated.html - last accessed on 18th July 2016.

1. **Intersection (∩):** We combine recommendations from multiple sources into one list by selecting only the *intersecting* tags, weighted by its position in the original lists. This weighting scheme is computed as $1/p$, where p is the tag's position in a list, thus giving precedence to those in higher ranks. The weights from each list for each tag are summed. The top tags, ordered by decreasing weight, are used as suggestions.

2. **Union (∪):** We use the same combination strategy as before, however, we consider the union of the given lists, thus increasing the recommendation coverage.

4.3.2 Evaluation Procedure

In this work, we train and test our approaches on the ACL2012 collection, as described in Chapter 2.3.2. Given tag recommendations made by one of the systems, detailed in the previous section, we evaluate against those tags annotated by the user. Using this evaluation procedure we compute standard metrics used in existing photo tag recommendation work (Garg and Weber, 2008): Precision at Five (P@5) and Mean Reciprocal Rank (MRR). The details of these metrics are discussed in Chapter 2.3.2.

4.4 Results and Discussion

Firstly, from Table 4.2, we observe that by suggesting frequent nouns and entities from a related stream of tweets (TP) we are able to significantly outperform our naïve baseline, supporting our hypothesis (**RQ4.1**) that social media sources can be used to alleviate the time lag problem associated with existing photo tag recommendation models. We observe the importance of using part-of-speech tagging methods as a technique to address noise in Twitter for PTR by comparing the large difference in accuracies between the systems which suggest only nouns/entities (TP) vs all terms (T). We improve upon these techniques in a more elaborate TF-IDF model (TF-IDF(N)) which makes suggestions based on Twitter term co-occurrence data for N input tags, achieving up to 31% recommendation accuracy for P@5. Therefore, term co-occurrence matrices can be built on "up-to-date" social media streams for new events where there lacks sufficient Flickr training data. Offering suggestions from Twitter does not achieve the same recommendation accuracy as our state-of-the-art baseline (REL), however, suggesting that our approach could be implemented as a weighted trade-off between the two sources as the amount of Flickr training data for an event increases.

Secondly, from Table 4.2, we observe that Wikipedia can also be exploited for tag recommendation purposes, however its application is not as effective as when recommending on Twitter data, perhaps due the narrower coverage of Wikipedia articles. The most effective recommendation strategy combines suggestions based on both Twitter and Wikipedia data (TF-IDF(1) ∩ WP) highlighting the complementary nature of these

	Baselines		**Individual**						**Combination**		
	POP	*REL*	*T*	*W*	*TP*	*WP*	TF-IDF(1)	TF-IDF(2)	TP ∪ WP	TP ∩ WP	TF-IDF(1) ∩ WP
P@5	0.05	0.53	0.00	0.11*	0.23*	0.10*	0.27*	0.31*	0.10*	0.33*	<u>0.47*</u>
MRR	0.08	0.69	0.00	0.51*	0.61*	0.51*	0.52*	0.45*	0.51*	0.57*	<u>0.69*</u>

Table 4.2 Recommendation comparison. Statistical significance against F denoted as * $p < 0.05$. Underline denotes the highest performing experimental approach.

evidences and supporting our initial research question (**RQ4.2**). Specifically, using an intersecting combination approach for both sources, we achieve accuracy which is almost comparable with our state-of-the-art baseline (REL) highlighting that Wikipedia can be used as a "filtering" mechanism for up-to-date, yet noisy, suggestions from Twitter.

4.5 Chapter Summary

In this work we developed an automatic approach for annotating event images by exploiting relevant social media and Wikipedia data. Specifically, we proposed photo tag recommendations based on significant nouns and entities present in tweets and Wikipedia data related to the Austin City Limits 2012 music festival. In this work, we highlighted the merit of computing recommendations based on these streams as an alternative to recommending based on Flickr data (which is often sparse and out-of-date due to users uploading images long after they are taken). In order to address noise present in social media streams, we applied natural language processing techniques and combined recommendations made with those computed from structured Wikipedia data. This work proposes a new area for image annotation research, and for this purpose we have released our test collection online[9].

One drawback of this work, however, concerns the small test collection used, due to the restrictive nature of the Twitter search API i.e. you can only retrieve 3,200 tweets posted within the previous week for a given query[10]. To overcome this, we suggest that future works (i) evaluate on larger (e.g. the Superbowl) and varied (e.g. protests, natural disasters *etc*) events (ii) consider other social media streams (e.g. Instagram, Google+, Facebook) (iii) combine tweets from the *search* and *streaming* APIs[11] (iv) extract information from related blogs (such as Wordpress, Tumblr *etc*), as well as from their user comments.

[9]http://mir.dcs.gla.ac.uk/resources/ - last accessed on 18th July 2016.
[10]https://dev.twitter.com/rest/public/search - last accessed on 18th July 2016.
[11]https://dev.twitter.com/streaming/overview - last accessed on 18th July 2016.

Chapter 5

Photo Tag Recommendation Diversification

5.1 Introduction

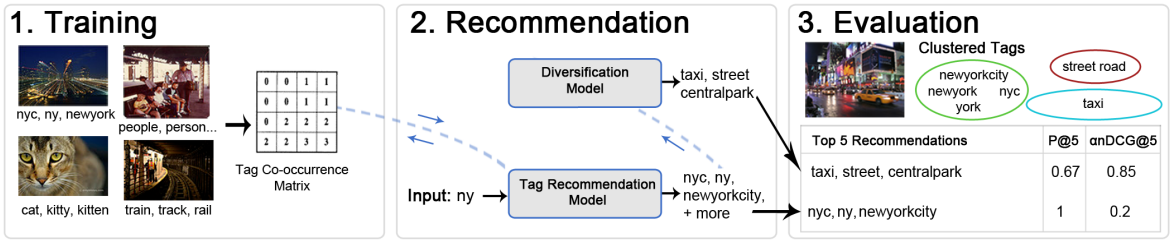

Fig. 5.1 The tag recommendation process: Training, recommendation and evaluation on collections containing synonyms

In the previous chapters we focused on photo tag recommendation approaches which attempted to maximise the *relevance* of tags in the top ranks. Although we were able to achieve promising results, the suggestions offered often contain many *synonyms* (e.g. `newyork` & `ny`) and *tag inflections* (e.g. `cat` & `cats`). Figure 5.1 gives an overview of *why* synonyms are promoted and the problems this poses for evaluation purposes.

Firstly, existing recommendation models employ tag co occurrence matrices on *user annotated images* (Garg and Weber, 2008; Sigurbjörnsson and van Zwol, 2008) which are often tagged with synonyms. Table 5.1 shows the top 5 co-occurring tags for two popular keywords; as can be seen, these tags co-occur highly with many *synonyms* i.e. users tag images with many redundant keywords. As a result, tag recommendation models often suggest these duplicates to the user, which offer *no added descriptive value* to the image. Figure 5.1, shows the recommendations (i.e. [`nyc, ny, newyorkcity`]) made by our TF-IDF recommendation model for the input tag `newyork`, trained on a collection of 0.5M Flickr images. Although all of the tags

Tag	Top 5 Co-occurring Tags
nyc	newyork, ny, newyorkcity, manhattan, city
cat	animal, kitty, kitten, feline, cute

Table 5.1 Top co-occurrences computed on 0.5M Flickr images. Synonyms are underlined.

returned are *technically relevant*, they are also synonyms of the input tag. Therefore, the user is *not* offered any useful tags as the top ranks contain only tags describing an aspect of the image which has already been annotated. For example, for the image in the evaluation stage of Figure 5.1, many sub-topics[1] of the image are missing in the recommendation list (e.g. [timessquare, street, taxi]). Therefore, there is a need for the diversification of suggestions in the tag recommendation process to remove these *synonyms* and offer a *more diverse* list of possible concepts to the user; in this work, we apply a diversification technique proposed in traditional web search, namely Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998), to re-rank the suggestions made by a recent photo tag recommendation model.

Secondly, existing models have been *tested* on user images, where no measures are made to ensure that synonyms are removed from image ground truths (Garg and Weber, 2008; Sigurbjörnsson and van Zwol, 2008). Therefore, approaches which suggest many synonyms will achieve higher precision and recall scores than those models which promote diverse tag lists (based on the intuition that these models will match more of the synonyms in the ground truth than a diversified recommendation list). For example: as can be observed in Figure 5.1, the ground truth tags in the evaluation image (right) contain many tags which describe the same aspect (e.g. ny/ nyc/ newyorkcity *etc*). Therefore, a recommendation model which suggests [nyc, newyorkcity, ny] would achieve high precision/recall scores, despite the tags offering *no additional semantic information*. However, a model which promotes tag *novelty* and *diversity* (e.g. [taxi, street, centralpark]) would achieve a lower evaluation score, due to the lack of diversity in the image's ground truth. We believe that a tag recommendation model should not suggest tags which relate to an already, well described aspect of an image (e.g. newyork in Figure 5.1), but instead offer a *wider range* of *novel* tags. Therefore, in this work we attempt to build a *new test collection*, where synonymous tag recommendation lists achieve lower evaluation metrics than diverse lists, without reducing or compromising the annotations made by the photographer.

In the following chapter, we attempt to address the following research questions (RQ):

RQ6.1 How can we effectively diversify tag recommendation lists in order to reduce the

[1]We define an image sub-topic to represent a visual *aspect* (e.g. sky, clouds) or non-visual *concept* (e.g. hot, warm)

number of synonyms in the top ranks?

RQ6.2 How do we overcome the problems with evaluating tag recommendation approaches on synonymous user annotations (i.e. where non-diverse recommendation lists achieve higher evaluation metrics)?

The rest of this Chapter is as follows. Firstly, we summarise works in text based diversification in Section 5.2. Section 5.3 introduces our crowd-sourced experiment, before we detail our diversification approach in Section 5.4. Finally, we detail our experimental setup and preliminary results in Section 5.5 before summarising in Section 5.6.

5.2 Background Work

Diversification techniques were originally introduced for re-ranking the results of text based information retrieval (IR) systems. Given a query, IR systems rank documents according to their estimated relevance, however, searchers' queries are often ambiguous or have multiple facets (Spärck-Jones et al., 2007). For example, *Java* is an ambiguous query since it has different interpretations e.g. the programming language, the island, and the coffee. Further, *java programming language* is a multi-faceted query since it has several aspects, e.g. development kit download, language specifications, tutorials, courses, and books. Ambiguous or multi-faceted queries are an issue for search engines; originally, this ambiguity was not addressed by retrieval algorithms.

Research now focusses on promoting *diverse* results lists which have a wider coverage of the query sub-topics, thus offering a solution to search ambiguity; from a multimedia perspective, work now also promotes visually diverse search results lists with the image retrieval task at ImageCLEF now also evaluating against diversification metrics (Paramita et al., 2010). In text-based diversification there generally exist *explicit* and *implicit* approaches (Drosou and Pitoura, 2010). Explicit approaches model the query aspects in their diversification approach and then maximise the coverage of the selected documents with respect to these aspects. In such approaches, the aspects of a document are gathered from external evidences, rather than from the content of the document itself. In contrast to explicit approaches that rely on external evidences to identify the query aspects, in implicit approaches, the similarity between the contents of the documents is used for diversification purposes.

The Maximal Marginal Relevance (MMR) model (Carbonell and Goldstein, 1998), an *implicit* model, was initially introduced in 1998 which diversified document ranks by promoting document novelty. Documents were promoted if they were dissimilar to the documents already in the new diversified list. This poster gave birth to an entire area of diversification research promoting many prevalent works, such as those by Agrawal et al. (2009) and Santos et al. (2010). Agrawal et al. (2009) proposed, IA-Select, a greedy diversification algorithm which attempts to maximise the probability that an average user will find some relevant information among the top search

ranks. Santos et al. (2010) introduced xQuAD, an explicit diversification model which promotes those novel documents which achieve maximal coverage over a number of query sub-topics. xQuAD is widely seen as the best performing state-of-the-art for diversification and has been extended and applied to a number of different domains outside that of general web search re-ranking, such as intent aware retrieval (Santos et al., 2011) and personalised content recommendation (Vallet and Castells, 2012). Vallet and Castells (2012) looked to combine personalisation and diversification by introducing an explicit random variable in the xQuAD model. Their approach was evaluated on a web search task, achieving significant improvements over the IA-Select and xQuAD models. Vargas et al. (2012) offered a relevance-orientated reformulation of xQuAD which was evaluated on both a web search and movie recommendation task. Finally, Santos et al. (2011) attempted to diversify web search results, using xQuAD, by learning the relevance of retrieval models for each intent of a query. Despite the amount of interest focussed on diversity in research years, its application to photo tag recommendation has been ignored. In this work we propose the diversification of tag recommendation lists to promote *novel* tags which cover as many of the different *aspects* of an image as possible.

5.3 Building a Test Collection

Firstly, we conduct a crowdsourced experiment, as detailed in section 5.3.1, which asks users to 'group the synonyms or the tags which refer to the same aspect' of 1,000 Flickr images, which are annotated with at least seven tags to ensure sufficient ground truth is available to test upon (randomly selected from the FLICKR-COL collection as described in Chapter 2.3.2). The methods we use to ensure high submission quality are detailed in section 5.3.2 with our results detailed in section 5.3.3. By undertaking this experiment, we are able to build a test collection where the ground truth contains sub-topics for each image, as required for diversification evaluation i.e. matching against tag groups opposed to individual annotations.

Crowdsourcing (i.e. outsourcing a task to a network of online workers) experiments have grown in popularity in recent years (Eickhoff and Vries, 2013; Hirth et al., 2010; Zuccon et al., 2013) and have been adopted to carry out tasks which are often difficult for computers but easy for humans e.g. image classification (Deng et al., 2009; Nowak and Rüger, 2010). Recently, Nowak and Rüger (2010) showed that by using a majority voting scheme for an image annotation task, the quality of worker judgements were in-line with those made by experts. The ImageNet collection was also built using a crowdsourced experiment where internet images were mapped to WordNet "nodes" (Deng et al., 2009). In this work, we instead use crowdsourcing to group tags of Flickr images into related "sub-topics" in order to create a test set for diversification evaluation and to analyse tag redundancy on photo sharing websites. The details of this experiment are discussed in the following sections.

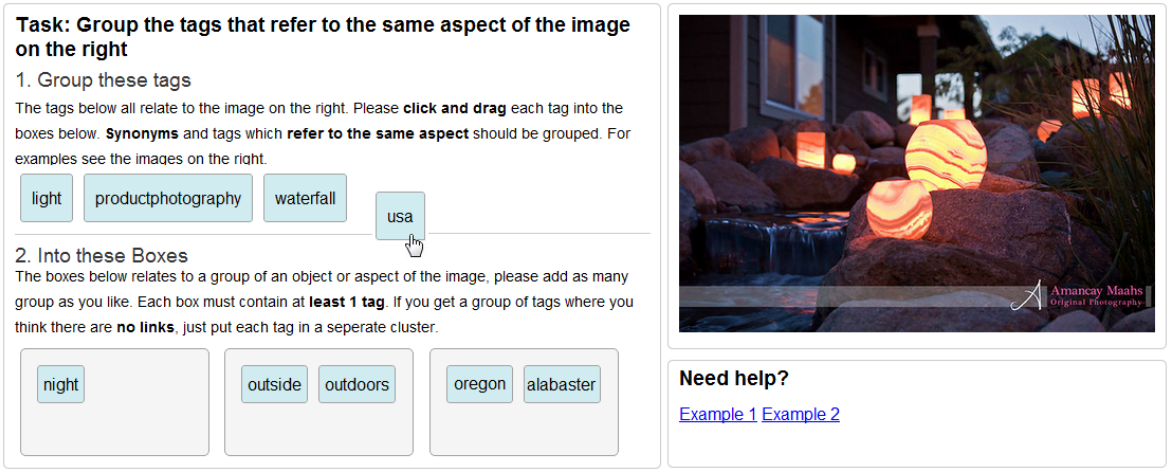

Fig. 5.2 Crowdsourced tag clustering interface

5.3.1 Experiment Procedure

We conduct this experiment, for the discussed 1,000 Flickr images, on the popular Amazon Mechanical Turk[2] platform. On this platform, human intelligence tasks (HITs) are taken out by various workers called 'Turkers'. In our experiment, only those Turkers with the *Master Qualification*[3] are able to accept our HIT. On acceptance of our HIT, users are presented with the following task description:

1. **What is required of you:** You will be presented with an image with the tags describing its contents. You must group the *synonyms* or the tags which refer to the *same aspect* of the image.

2. **Details:** You will be presented with 20 images. You may skip up to 3 images. You have a maximum of 45 minutes to finish the experiment. To group tags, simply click and drag them into the displayed boxes, then click submit. All of the tags must belong to one group, and every group must contain at least one tag. This experiment is supported for Firefox and Chrome (Res 1024+).

3. **Finally:** You must judge at least 17 images and be a *fluent English* speaker.

Figure 5.2 shows a screenshot of the interface that was used to construct the diversified test collection. Users were able to click and drag each tag of the image into an existing cluster or add it to a new cluster. Each tag had to exist in a single cluster, and a cluster could contain one or more tags. A *video* tutorial and two *example images* accompanied the task description, allowing the worker to fully understand what is expected of them before accepting the HIT. Turkers are paid if they agreed to and

[2]https://www.mturk.com/ - last accessed on 18th July 2016.

[3]'Workers who have demonstrated excellence in a type of HIT, for instance categorization, are awarded the Master Qualification'

carried out the conditions of the experiment. On acceptance of these terms, the worker is presented with a registration questionnaire asking for the following details: TurkID, age, sex, occupation, education level and proficiency in English. On completion of the registration form, the Turker is presented with the first image of the user experiment where they can begin to group the tags into related image aspects.

5.3.2 Ensuring Quality

One of the major problems with crowdsourcing, however, is that workers often spam or try to complete tasks with as little effort as possible in order to maximize their profits. This can lead to poor quality submissions. Many existing works have resolved this problem by introducing a number of *'honeypots'* (Vuurens et al., 2011) i.e. tasks where the correct 'answers' are already known. We therefore introduce a number of honeypot images, which aimed to identify spamming users. Specifically, for every 20 images, we present the user with *three* images where the tags had been pre-grouped by an expert. Care is taken in creating these clusters to ensure that there is no ambiguity in the groupings. Creating these honeypots allows us to indicate users whom completed the HIT without *reasonable effort*. Any user which grouped the tags of these honeypot images differently than the expert is blocked and their work is discarded. Also, the work of any Turker whom describes their English level as less than "fluent" is also removed.

To further minimise spam, each image has its tags clustered by three *different* workers. We compute the final clusters using the following aggregation scheme (Nowak and Rüger, 2010): we first compute a co-occurrence matrix which counts the number of times two tags co-exist in the same cluster, for each image, as judged by the three users. Two tags are clustered together if they co-existed in at least two of the three user clusters. Tags which are never grouped by any of the three users, or are clustered just once, are assigned to their own cluster. The clusters are iteratively built, where clusters are merged if they contain a common co-occurring tag.

5.3.3 Crowdsourcing Results

In total 197 different Turkers accepted the hit, with 20 users failing to pass the honeypot test. Therefore, work is accepted from 177 Turkers. Each HIT (20 images) was completed in 23 minutes and 40 seconds, on average. Turkers were paid between $1 and $3 for their work, which equates to $5.98/hour on average. From the entry questionnaire, around 70% of users said English was their *first language* and around 30% described their English proficiency as *fluent*. Further, 49% of Turkers were female and 51% male, with an average age of 34. Finally, 80% of workers described their education level as "college" or higher (undergraduate or postgraduate).

Each image in our test collection contains around 9.87 clusters (with each containing around 2.18 tags), on average. Considering that images in our test set were

	
(autumn, fall) (bus) (cannon) (maryland, md) (pumpkin) (target)	(art, sculpture) (aurora) (balance, balanced) (rocks, stack, stones)
	
(1960s, 1950s) (fern, leaves) (casserole, cups, pitcher, plates) (china, collectibles, collections) (midcenturymodern, modernism)	(bw) (floor) (italy, rome, vatican) (museum) (tourist)

Table 5.2 Cluster aggregation output: 4 random images

annotated with 21.5 tags on average, this indicates that more than half of the tags in our test collection are *redundant* (assuming each tag in a cluster describes a single image aspect). Table 5.2 shows the aggregated clusters of four random images from our collection. As is observed, the clusters in these examples are highly relevant (e.g. `autumn`, `fall`) and demonstrate expert knowledge (e.g. that `md` is an abbreviation of `maryland`).

5.4 Methodology

As discussed in Chapter 2.3.2, the goal in *traditional* tag recommendation evaluation is to recommend a set of tags, p_j, given a subset of tags, q_j, from d_j ($q_j \subset d_j$), so that it

maximizes $p_j \cap (d_j - q_j)$, where d_j denotes the set of tags annotated for the j-th image in our collection. In this work we evaluate for *diversity*, using intent aware metrics. Therefore, in order to evaluate for diversity we group the tags in d_j into c clusters (computed in our crowd-sourced experiment), each representing an image *'sub-topic'*. The goal in *diversification* tag recommendation evaluation is to recommend a set of tags which maximises the *cluster* coverage i.e. where recommended tags exist in the most *clusters*. In our work we suggest tags using the TF-IDF recommendation model (see Chapter 2.3.2) before re-ranking them using the MMR approach discussed in Section 5.4.2. Firstly we discuss our BOW tag representation in section 5.4.1.

5.4.1 Tag Representation

In traditional IR and textual based diversification approaches, a document can be represented using a bag of words (BOW); in our scenario, however, we are re-ranking tags which are uni-gram terms, restricting the use of such a model. Therefore, we must introduce a vector based approach to represent tags, for their use in the diversification model. In this work we represent a tag based on its co-occurrence with the other tags in the collection using the TF-IDF vector representation described in Chapter 2.3.2. This vector representation allows for comparability of tags using standard IR similarity measures.

5.4.2 Tag Diversification

Given an *existing tag suggestion list R* (i.e. from a tag recommendation method), we look to re-rank these suggestions into a new *diversified list*, *S*. The overall goal in diversification is to re-rank the list in order to maximise the coverage of image subtopics in the top ranks. In our work we adapt the Maximal Marginal Relevance (MRR) diversification model for our purpose which we define formally as follows:

Formal Definition Let $sim(t_i^{'}, t_q^{'})$ denote the cosine similarity between the vector representations of tag t_i and input tag t_q; this can be regarded as a measure of relevance for $t_i^{'}$ to $t_q^{'}$. We consider the situation where $|R|$ tags have been ranked, and the ranking function considers which tag has to be ranked next:

$$t_n^{'} = \text{argmax}\,[\lambda\, sim(t_n^{'}, t_q^{'}) - (1-\lambda) \max_{t_j^{'} \in R} sim(t_n^{'}, t_j^{'})] \tag{5.1}$$

where λ is a parameter that controls the impact of diversification on the selection of tag $t_n^{'}$: i.e. $\lambda = 1$ uses no diversification.

Informal Definition Less formally, MMR is used to maximise both relevance and novelty in the top ranked tags by re-ranking the *existing tag list*, as suggested by our TF-IDF approach, into a new *diversified list*. It works by iterating over the tags in the

existing tag list and computing the maximum cosine similarity score with those tags in the new *diversified list* (which is initially empty and therefore this measure is initially equal to 0). This score is then subtracted from the relevance score in the *existing tag list* and the tag added to the *diversified list* with the updated score. This process is repeated until all the tags in the *existing tag list* have been processed. It can be observed that if a similar tag already exists in the *diversified list* when a tag is being processed, its score is "demoted" as it presents little novelty to the recommendation list (and has a high maximum cosine similarity).

5.5 Experiments and Results

In our experiments we build our tag co-occurrence matrices using the images in the FLICKR-COL collection, removing the 1,000 images used in our test set. Our evaluation procedure is as follows:

For every image in our test collection, we select a random tag from the image's tag list to query the described recommendation model. For our baseline (i.e. *TF-IDF*) approach we use these suggestions for evaluation purposes. In our diversification approach, we re-rank the top n recommended tags ($n = [10, 20, 30, 50]$) using the MMR method, described in Section 5.4.2. For both approaches, we evaluate the top tag suggestions against the clustered sub-topics defined in the crowdsourced experiment, computing common diversification metrics: (i) *α-Normalised Discounted Cumulative Gain (α-nDCG@N)* (Clarke et al., 2009) ($\alpha = 0.5$), and (ii) *Intent-aware Expected Reciprocal Rank (ERR-IA@N)* metric (Clarke et al., 2009). These metrics are detailed in Chapter 2.3.2. For each metric, we evaluate for cut-offs $N = \{5, 10\}$.

	α-nDCG		***nERR-IA***	
	@5	**@10**	**@5**	**@10**
TF-IDF	0.165	0.159	0.169	0.164
+MMR	0.176* (*+6.7%*)	0.169* (*+6.3%*)	0.179* (*+5.9%*)	0.173* (*+5.5%*)

Table 5.3 Results ($n = 20$); statistical significance against TF-IDF denoted as * $p < 0.05$

As can be observed from Table 5.3, by employing the techniques discussed in the previous sections we are able to offer a more diverse list of tag recommendations which cover more of the relevant "sub-topics" within images, highlighted by the statistically significant increase to the α-nDCG and nERR-IA metrics at both cutoffs. From these results, we achieve highest performance when re-ranking the top 20 tags (i.e. $n = 20$) using a fairly low λ value, where we observed a local maxima when $\lambda = 0.4$. Alternatively, we observe that using an a*ggressive diversification strategy*, where relevance and diversification scores are treated almost equally, for a *small selection* of the top ranked tags (i.e. the top 20 out of a vocabulary of 853k different tags) to be the most effective strategy. This may imply that only a small number of tags are ever relevant

for any given image, but often these tags have semantic overlap (i.e. synonyms) which we should attempt to reduce in the top ranks in order to propose a more *useful* list of tags to the user.

5.6 Chapter Summary

This chapter proposed the diversification of photo tag recommendation lists. Firstly, we highlighted problems, caused to both tag recommendation and evaluation, with users tagging images with synonymous tags. To summarise, training on synonymous tags results in synonymous recommendations, which are in turn are evaluated against synonymous ground truths, thus yielding misleading performance measures. In order to overcome this problem, we conducted a crowdsourced experiment to quantify the level of tagging redundancy on Flickr as well as to create a fair test collection for photo tag recommendation diversification evaluation. We then adapted the Maximal Marginal Relevance diversification model for tag recommendation purposes in order to increase the coverage and reduce the synonymity of suggestions made by a recent photo tag recommendation approach, significantly improving recommendation performance.

The results of this work highlight an important problem faced by *all* tag recommendation approaches (and not just PTR) where one must ensure that documents do not contains duplicates, or synonymous, ground truth in order to build effective models and evaluate fairly. It should not be understated the importance of this observation as all previous tag recommendation works may report misleading results. We therefore propose that all future tag recommendation works should compute diversification centric metrics to measure the true "usefulness" of a suggestion list, as a recommended tag which has already been annotated by the user in some other form (e.g. `ny` for an image tagged with `nyc`) offers *no additional value* from a annotation and therefore retrieval/recommendation perspective.

Chapter 6

AIA and PTR Evaluation

The following chapter is published in the following conference:

- Philip J. McParlane, Yashar Moshfeghi, Joemon M. Jose (2014); Collections for Automatic Image Annotation and Photo Tag Recommendation, *Proceedings of ACM International Conference on MultiMedia Modeling (MMM) 2014, Dublin, Ireland.* ACM New York, NY, USA, pp133-145.

6.1 Introduction

Despite the intense research focus and number of works on the annotation of images published in the last two decades, a comparison of approaches is difficult due to the lack of a unified evaluation framework and collection. A review of the 20 most popular automatic image annotation papers[1] showed that at least 15 different collections were tested upon[2]. These collections vary in characteristics and hence introduce biases of their own into the evaluation, highlighting the need for a single test collection which is *representative* of images uploaded to image sharing websites. Additionally, the most prominent works in photo tag recommendation all use their own collections (Garg and Weber, 2008; Liu et al., 2009; Sigurbjörnsson and van Zwol, 2008).

Aside from the large number of collections used to benchmark annotation models, we have identified *seven* flaws which may result in misleading performance measures and therefore the incomparability of state-of-the-art models. The problems are as follows: (i) *class ambiguity*, in the form of synonyms e.g. testing for `ocean` vs `sea` (ii) *testing on unnormalised collections*, where SOTA models are able to boost annotation performance by promoting popular tags (iii) *low image quality* (iv) *lack of image*

[1] Selected by searching http://citeseerx.ist.psu.edu/ for "`automatic image annotation`", ordered by descending citation count (Dec'12)

[2] The collections were: Corel5k, Corel30k, ESP Game, IAPR, Google Images, LabelMe, Washington Collection, Caltech, TrecVid 2007, Pascal 2007, MiAlbum & 4 other small collections.

meta-data (v) *lack of image diversity* (vi) *using location tags as ground truth*, and (vii) *copyright restrictions.*

For photo tag recommendation, we have identified the *three* problems with the collections used by Sigurbjörnsson and van Zwol (2008) & Garg and Weber (2008): (i) *using crowdsourced ground truths*: only the photographer of an image understands the true content and context of an image (ii) *synonyms in the ground truth*: models which promote synonyms in their suggestions are promoted over those models which suggest diverse recommendations, and (iii) *lack of distribution*: currently tag recommendation works test on their own private collection.

In particular in this chapter we aim to address the following research questions:

RQ6.1 What issues are there with existing AIA & PTR evaluation collections (if any) and do they present biases which may lead to misleading evaluation metrics?

RQ6.2 How are we able to more fairly evaluate automatic image annotation & photo tag recommendation approaches?

The rest of this chapter is as follows: firstly, we detail related works in the evaluation of PTR and AIA models in Section 6.2. Section 6.3 fully details the problems associated in automatic image annotation evaluation before introducing Flickr-AIA. In Section 6.4 we further detail the problems associated in photo tag recommendation evaluation before introducing Flickr-PTR. Finally, we summarise in Section 6.5.

6.2 Background Work

Effectively evaluating multimedia annotation models has been a contentious topic for researchers in recent years. In particular, a number of issues associated with the evaluation of AIA models have been identified in various publications: Westerveld and de Vries (2003) highlighted a range of problems with the Corel collection, such as the fact that images are grouped into *coherent themes*, resulting in misleadingly high performance measures. Athanasakos et al. (2010) compared two existing models showing that the high performance reported was more to do with the evaluation scheme and test set instead of the approach itself. Muller et al. (2002) highlighted issues with using the Corel image collection, in that many models test on a different subset of this collection resulting in different performance measures. In our work, we discuss new biases, resolving them in new evaluation collections.

The popularity of works which specifically consider multimedia datasets has resulted in new tracks at major conferences, such as the "dataset track" at the ACM Multimedia System (MMSys) conference where accepted papers are rewarded with free hosting services in order to encourage experimentation on common collections. The track which has ran since 2013 has resulted in 40 new multimedia collections covering various tasks from football player tracking (Pettersen et al., 2014) to hand gesture detection (Hsiao et al., 2014). Most relevant to our work is that by Mousselly-Sergieh

et al. (2014). In this work, the authors present a collection of 14 million geo-tagged images where a heuristic based tag pre-processing stage is taken out removing stop-words and "non-descriptive tags" (e.g. photo *etc*). Aside from MMSys, there exist other initiatives, such as the community driven MediaEval Benchmark[3] (Larson et al., 2015) which also encourages research in open multimedia evaluation in the form of an annual workshop, which runs alongside the ACM Multimedia (MM) conference.

Despite the amount of work taken out on multimedia dataset evaluation and the number of new collections proposed, many modern AIA works (Makadia et al., 2010) still evaluate on small out-dated collections (i.e. low resolution, <100k images) which do not reflect real world scenarios where models must be able to handle *millions* of images. Some collections, however, are beginning to gain popularity; in particular, Deng et al. (2009) introduced the ImageNet collection, consisting of 3.2M images (which is constantly being extended), which is structured into synonym sets of the WordNet lexical database (Miller, 1995). Huiskes et al. (2010) introduced two Flickr collections containing 25K (Huiskes and Lew, 2008) and 1M images (Huiskes et al., 2010). However, these collections are not created with defined train/test subsets for annotation or tag recommendation evaluation. Further, these collections fail to address a number of the issues presented in our work such as tag ambiguity and normalisation. Despite the increase in the availability of computational power, in the form of MapReduce clusters and multi-core machines, the computationally intensive task of image annotation on this volume of images is out of the reach of many, and therefore a more manageable collection is desirable for most. In our work, we propose new collections for image annotation and photo tag recommendation purposes which aim to overcome a number of discussed problems.

6.3 Automatic Image Annotation Evaluation

The purpose of an image annotation evaluation collection is to benchmark a given annotation method, for a number of image classes or scenes, based purely on its visual discriminatory power. Therefore, these classes should be distinct (and not ambiguous) and easily identifiable by a human based purely on their appearance. The images in this collection should reflect real, user images and should cover a diverse range of images for each class; alternatively, the images should be taken in different locations, by different users, in a number of different lighting conditions, on a range of devices. By doing so, annotation models would be benchmarked for as close to a real world scenario as possible. In the following, we introduce three popular annotation collections in section 6.3.1 before detailing the problems they pose for fair evaluation in sections 6.3.2 and 6.3.3. Finally we detail our new collection which aims to tackle the issues presented in section 6.3.4.

[3]http://www.multimediaeval.org/ - last accessed on 18th July 2016.

Collection	Images	Tags	Ambiguity	Time/Loc	Free	Size	Train	Test	I/T
Corel	5k	374	9.6%	×	×	160px	4.5k	0.5k	88
ESP	22k	269	9.7%	×	✓	156px	20k	2k	377
IAPR	20k	291	12.7%	✓	✓	417px	18k	2k	386
Flickr-AIA	312k	420	0%	✓	✓	719px	292k	20k	2,304

Table 6.1 Comparison of the collections (i) Ambiguity: % of tags where there exist at least one synonym (ii) Size: average dimension in pixels (iii) Time/Loc: whether time taken and location details are included (iv) I/T: average # images per tag

6.3.1 Existing Collections

We consider the following collections: *Corel* (Duygulu et al., 2002), *ESP Game* (von Ahn and Dabbish, 2004) and *IAPR* (Grubinger et al., 2006). These collections are selected as they have been used to benchmark *many* AIA models of recent years (Athanasakos et al., 2010; Makadia et al., 2010). We use the same *methods*, *training* and *test* subsets as used by Makadia et al. (2010). These collections, along with the collection introduced in this work (Flickr-AIA), are summarized in Table 6.1.

One popular collection which has been omitted and is related to this work is the MIR-FLICKR collection (Huiskes et al., 2010). We have not considered this collection as it is not setup with image annotation evaluation in mind; they contain user tags rather than high level, visual, classes. However, these collections have been used in the ImageCLEF 2009 annotation task, where the referred 25k collection was annotated using a crowdsourced experiment. Despite this, the collection was only made available for the participants in this task and is no longer publicly available. Therefore, researchers are unable to compare new annotation approaches on this testbed. Additionally, a collection of 25k images, is too small by modern standards. In this work we introduce a larger collection for AIA evalation which is freely available.

6.3.2 Annotation Model

To demonstrate the issues with the given collections, we conduct a number of experiments using the annotation model described by Makadia et al. (2010). The method models the problem of image annotation as that of image retrieval using a k-nearest neighbour (KNN) approach ($K = 10$, as used by Makadia et al. (2010)). Seven features are extracted from images, three colour histograms in various channels (RGB, HSV and LAB), two texture descriptors (HAAR and Gabor filters) and two quantized versions of the texture features. Each feature vector is normalised, with visual similarity between images computed using the average of the seven distances (for each feature pair).

Each distance is scaled by its *maximum distance*, for the given feature, within the training set. The L_1 distance is used for all features, apart from the LAB descriptor, where the K-L divergence measure is used. N tags (using $N = 5$, as used by Makadia

et al. (2010)) are transferred from the nearest neighbour (ordered by frequency in the training set). If the number of tags in the nearest neighbour is $< N$, tags are transferred from the surrounding neighbourhood. The top tags, ranked by the *product* of tag occurrence in the neighbourhood and co-occurrence with the nearest neighbour, are selected. This model is used to highlight problems with testing on unnormalised collections.

6.3.3 Problems

As briefly discussed, we have identified seven problems with using existing popular image annotation evaluation collections. In the following sections we detail these problems in greater depth.

1. Tag Ambiguity One of the major problems with these collections concerns the classes they use as ground truth. All three collections contain synonyms (e.g. `america`/`usa`) or visually identical classes (e.g. `sea`/`ocean`). For the purposes of generic image annotation, a model should not have to differentiate between synonyms, as often (from analysing the visual contents), this is impossible. For example, consider, as a human, differentiating between an image of the `sea` or the `ocean`. To illustrate this problem, we use WordNet (Miller, 1995) to classify keyword pairs as synonyms i.e. those keywords which contain a common synonym set. After a list of potential synonyms is generated, pairs which are seen to be incorrect by an assessor (e.g. `ball`/`globe`) are removed.

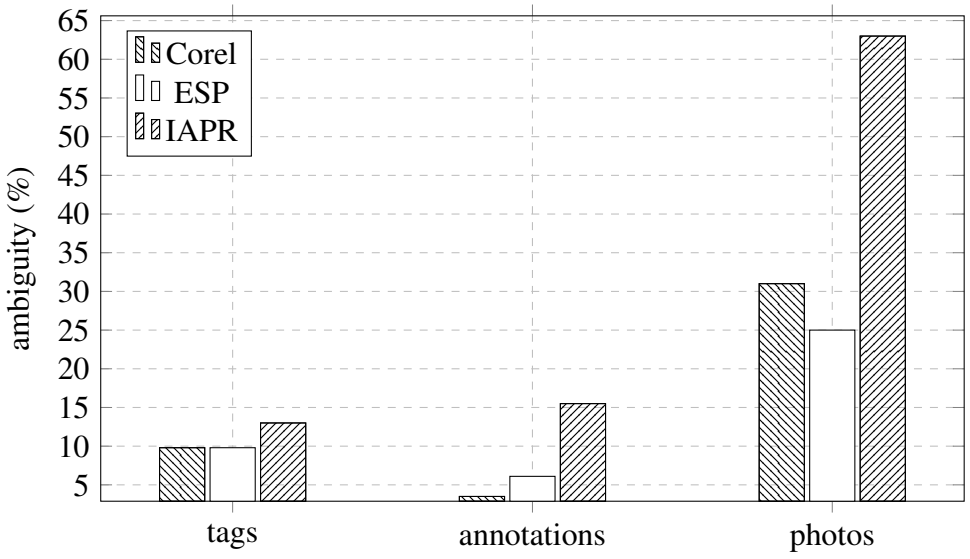

Fig. 6.1 Ambiguous tags: those tags which have at least one synonym. Ambiguous annotations: those tags assigned to images which have at least one synonym. Ambiguous photos: photos containing at least one ambiguous tag.

Using this approach we identify 36, 26 and 37 *ambiguous tags* (i.e those tags which

have at least one synonym) for the Corel, ESP and IAPR collections, respectively. Figure 6.1 highlights the percentage of *ambiguous tags* present in each collection. Around *one* in *ten* tags in each collection is deemed ambiguous. This equates to 15% of all photo annotations in the IAPR collection meaning a model may under perform by up to 15%, as for each *ambiguous* annotation in the ground truth, the model may predict the synonym. Therefore evaluating on these collections may result in misleading performance measures. For example, if an image's ground truth is [`home`, `sea`] and it is annotated with the tags [`house`, `ocean`] it will achieve precision and recall scores of 0. This is clearly a bias experimental framework as luck plays a *major* role in the scoring of evaluation measures. Table 6.2 summarises the most occurring synonyms pairs.

Collection	Top Synonym Pairs
Corel	`field/lawn, polar/arctic, ice/frost, ocean/sea`
ESP	`home/house, rock/stone, baby/child, child/kid`
IAPR	`woman/adult, building/skyscraper, rock/stone`

Table 6.2 Top synonyms for each collection

2. Unnormalised Collections One of the main issues with the evaluation of existing annotation models lies in the unbalanced nature of collections. By nature, the classes used in image collections follow a long tail distribution i.e. there exist a few *popular* tags and many *unpopular* tags. For the evaluation of annotation models, this leads to a bias experimental setup for two reasons:

- *Selection Bias*: Popular tags exist in more training and test images. Therefore, annotation models are more likely to test their annotation model on these keywords, purely because a popular tag is more likely to exist in a random test image than an unpopular tag.
- *Prediction Bias*: Due to the wealth of training data available for popular keywords, annotations models are more likely to annotate images with these tags, as they are more likely to be correct.

The unbalanced nature of collections therefore allows for potential *"cheating"* where models promote popular tags over less popular tags. To fairly measure a model's annotation accuracy based *purely on visual content*, models should not be able to exploit attributes of collections, such as tag popularity.

To demonstrate the hypothesis that popular keywords can be exploited to increase annotation accuracy, we split each collection into three vocabulary subsets representing the *popular*, *medium frequency* and *unpopular* tag sets. We denote the full vocabulary as *entire*. We select each subset so that each contains *approximately* the same number

of keywords (i.e. one third), from the overall vocabulary. Using the annotation model described in Section 6.3.2, we annotate the images in each collection four times, annotating only with tags in each tag subset (and the entire set). Precision and recall measures are then computed against the tags in the ground truth, which *exist* in the given subset.

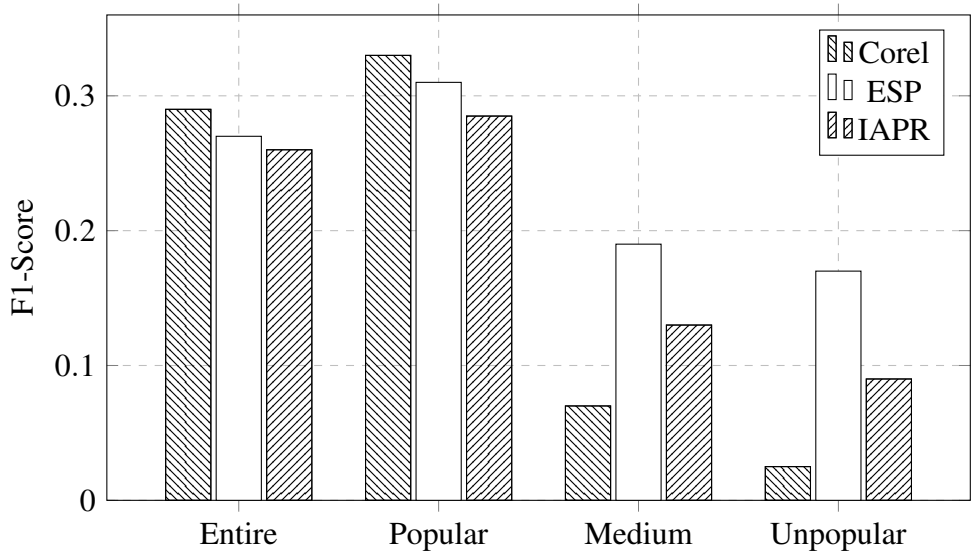

Fig. 6.2 Normalised annotation for each collection. Due to space constraints: Pop = Popular, Med = Medium, Unpop = Unpopular. F1-Score = $2(P*R)/(P/R)$, where P = precision R= recall.

Figure 6.2 shows the results of this experiment. We observe that *popular* keywords are easier to annotate than *less popular* tags. Additionally, when we annotate the images purely with popular tags, we achieve higher accuracy than the collection as a whole. Therefore, models may exploit this collection characteristic by promoting popular tags, leading to higher than expected measures for precision and recall. This annotation trend is observed across all collections.

It may be argued that by normalising, we are creating an *unrealistic* test set. However, if AIA models are benchmarked *purely* on visual features, we are measuring a model's *true* discriminative visual annotation power, without the bias of promoting popular tags. We believe that the "popularity" weighting of a tag could be applied *after* annotation on visual appearance has taken place. In our test collection, we propose two ground truths, an unnormalised (real life) and normalised version (containing only medium frequency tags). We hypothesise that by improving annotation accuracy on the normalised ground truth, we will improve a model's visual discriminatory power, thus increasing accuracy on a real life collection. We encourage researchers to report evaluation metrics on both ground truths to ensure a model is not exploiting the long tail distribution and is annotating well on visual appearance.

3. Quality of Images The small size and poor quality of images in many collections often make it difficult to extract semantics from the visual contents of images, due to the lack of resolution and visual artefacts present. Despite this, the images contained in modern evaluation collections are often very small (see Table 6.1). The quality and size of images used in evaluation collections must increase to reflect those images taken on high resolution smart-phones and digital cameras.

4. Lack of Meta-data AIA is being more recently viewed from an information retrieval perspective, rather than that of content analysis, where time and location (Zhang et al., 2012b) are being exploited in the image annotation process. Despite this, all the collections used fail to include time, location and user meta-data. Therefore to allow deeper contextual analysis of images in the annotation process, every detail of an image's meta-data should be made available.

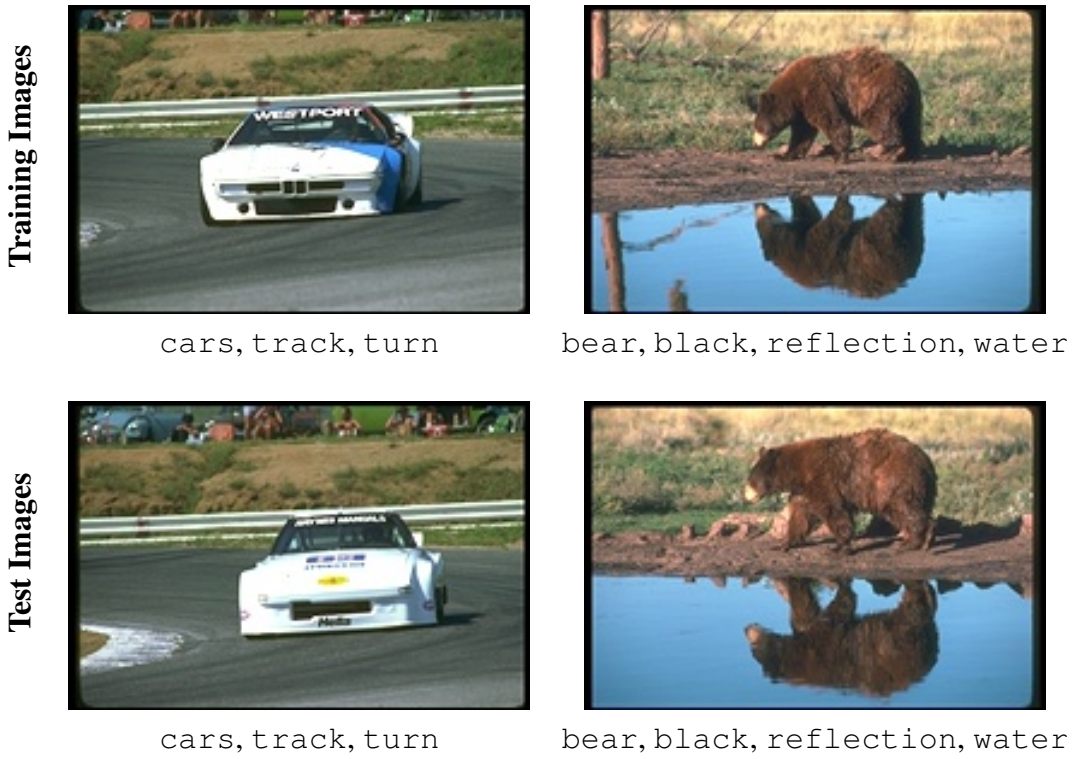

Fig. 6.3 Duplicates in test and train sets for Corel.

5. Lack of Diversity Images in the described collections are often taken by the same user, in the same place, of the same scene/object, using the same camera (Westerveld and de Vries, 2003). This leads to natural clustering in image collections, making annotation easier due to high inter-cluster visual similarity. This also causes problems such as duplicate images in the test and train set, making annotation easier, as observed in Figure 6.3.

6. Identifying Location As highlighted by Huiskes and Lew (2008), identifying a location from an image is often impossible. Despite this, two of the three image collections contain ground truth classes which are locations (e.g. `scotland`). These should not be used for image annotation evaluation purposes.

7. Copyright The most popular baseline collection, Corel, is not freely available and is bound by copyright. To allow for the easy comparison of annotation models, a collection should be at least *free* and *distributable*.

6.3.4 Flickr-AIA

In this following section we detail the process used to build the Flickr-AIA collection, which aims to resolve these problems. In total, we present two test collection ground truths for 20k images, one with a normalised ground truth (i.e. where the image classes contain roughly the same number of test images), and one without (i.e. a real life scenario). We refer to how we address each problem by referencing the problem number in parenthesis e.g. (1).

Building the collection In this collection, we apply a number of transformations to the FLICKR-COL collection (described in Chapter 2.3.2). This collection is built by first querying Flickr for 2k popular nouns extracted from WordNet (Miller, 1995) (categorised as `animal, artifact, body, food, plant, substance`). The top 2k images, which contain the creative commons license, (7) location, user and time meta-data (4) and at least one tag, for each search are then considered for use in our collection. Using this approach, we collect images covering a wide range of topics (5). We download the "largest" available compressed size version (not the original) for each image (3), ensuring high resolution and small file size.

Initially we collect 2M images before a number of pre-processing stages are taken out to resolve the discussed issues. As ground truth we use the tags assigned by the Flickr users; this has a number of advantages and disadvantages. By using user annotations, we are able to collect a *large* number of images, in comparison to the manually collated ground truths used in the Corel and IAPR collections. However, user tagging is often *noisy*, where tags do not refer to the visual contents of an image. In order to remove these tags deemed irrelevant for image annotation we use the following approach:

Collection Cleaning As most tags are irrelevant for the task of image annotation evaluation (e.g. non-visual classes), we first undertook a cleaning stage to remove these unsuitable tags. Specifically, using three assessors[4] we manually removed those tags which fell into the following categories, device information (e.g. `d60`), Flickr

[4]The assessors consisted of myself and two colleagues at Yahoo! Barcelona, namely: Ingmar Weber & Yelena Mejova.

awards (e.g. `excellent photograph`) and Flickr groups (e.g. `5photosaday`), from the top 1,000 most frequently occurring tags, as we believe these tags are either impossible or unsuitable to predict based on visual features. After removal of these redundant keywords we consider only the top 500 tags, ranked by descending number of users, for use in the collection. This removes tags which are used by only a few users (i.e. noise) and keeps popular classes which are more likely to be well known objects/concepts (i.e. potential image classes). We use WordNet to classify the remaining tags. Only *nouns* which are *not* categorised as the `noun.time` or `noun.location` sub-categories are used in the collection (6). By selecting nouns, we consider only visual objects, ignoring concepts difficult to identify e.g. verbs such as `talk`. Time and location tags are omitted as they are also difficult or impossible to annotate based purely on visual content (Huiskes et al., 2010) e.g. `Romania` or `2010`. Filtering stages (5) and (6) remove 80 tags deemed irrelevant for annotation purposes, leaving the final vocabulary size at 420.

Promoting Diversity As identified by Westerveld and de Vries (2003), previous collections, such as Corel, often cluster images into coherent themes, where image similarity is high. This makes it easier for AIA models as, for every test image, there are likely to be many images in the training set which are almost visually identical. We therefore limit the number of images taken by a user to 20 to promote visual diversity (5).

Removing Synonyms We remove synonyms in the remaining tag set using the same method as described in Section 6.3.3, by grouping tags which co-exist in a common WordNet synonym set. Specifically, this task was undertaken manually by myself, where each pair of tags which existed in the same WordNet synset were manually evaluated in order to differentiate between them only being *semantically related* (e.g. `world` vs `globe`) or actual *visual synonyms* (e.g. `home` vs `house`). Based on this, those tag pairs which were deemed "visual synonyms" were grouped together, and otherwise ignored in the case of "semantically related". This task could have been crowdsourced, however, due to the small size of the collection (i.e. 420 tags) we believe this is unnecessary. In total, 49 synonym pairs are identified and merged (1). The details of the final collection are shown in Table 6.1 (see page 94).

Test Sets From this collection, we remove 20k random images for testing purposes, leaving the rest for training. As previously discussed we offer two ground truths to test against for these images (i) *full ground truth* i.e. traditional evaluation where images contain all annotations (ii) *normalised ground truth* i.e. where only those middle frequency classes are selected (2). Specifically, we select only those tags which occur in the middle third of tags ordered by frequency i.e. tags #140 to #280. By offering this normalised ground truth, we are able to test annotation models based purely on their visual discriminative power, removing the bias from offering popular tags. It should

Collection	# Training	# Test	Tags	Freely Available	Ground Truth
Sigurbjornsson	52M	331	3.7M	×	Crowdsourced
Garg	50M	9k	-	×	User Tags
Flickr-PTR	2M	1k	1M	✓	Clustered User Tags

Table 6.3 Collection comparison (i) I/T = average # images per tag (ii) T/I = average # tags per image

be noted, however, that we are not recommending that researchers test *only* on this normalised ground truth due to its unrealistic composition. Instead, we propose that researchers attempt to optimize for both test sets, and by doing so build more visually discriminative models which do not *overfit* the training set.

6.4 Photo Tag Recommendation Evaluation

In photo tag recommendation, the typical evaluation approach is to take a small number of tags from an image and attempt to predict the other tags. As predictions are made based on textual features, the range of ground truth classes can take a larger number of classes than those used in AIA. Differing to that of AIA evaluation, ground truth tags can also refer to both an image's visual content (e.g. `man`), its context (e.g. `london`) & non-visual concepts (e.g. `religion`). In the following, we first highlight problems with test collections used by two existing tag recommendation methods in sections 6.4.1 & 6.4.2. Finally we detail our new collection in section 6.4.3, Flickr-PTR, which is built for the purposes of tag recommendation evaluation in mind.

6.4.1 Existing Collections

In this work we consider the evaluation collections for tag recommendation used by Sigurbjörnsson and van Zwol (2008) & Garg and Weber (2008). Unfortunately these collections are not freely available making any analysis or comparison with our collection difficult; however, we detail what is described in the respective papers, along with details of our new collection, Flickr-PTR, in Table 6.3.

6.4.2 Problems

In this work we identify three problems with using existing popular photo tag recommendation collections for evaluation purposes, as detailed in the following section.

1. Crowdsourced ground-truths The test collection used by Sigurbjörnsson and van Zwol (2008) compares predictions against a crowdsourced ground truth for 331 images. We agree with Garg and Weber (2008), that the ground truth of an image can only be identified by the user whom the photograph is taken by. For example,

consider a user's holiday/vacation photographs: only they (and those present at the time) will know various details regarding the images, such as: (i) the people present (ii) the locations visited (iii) the event attended *etc*. Therefore, an approach which tags images using a crowdsourced experiment will result in substandard annotations. Garg and Weber (2008) follow this notion by adopting user tags as image ground truth, however, we identify an issue with this approach which may give mis-leading results, as described in the following subsection.

2. Synonymous Ground Truths One of the issues with using user tags is that, by nature, users tend to tag images with multiple synonyms, as described in Chapter 5, in order make their image searchable for the various versions of the same entity. For example, instead of tagging an image solely `newyork`, many images also include a number of synonymous tags e.g. `ny`, `nyc` and `newyorkcity`. In our (unfiltered) collection (containing 2M Flickr images), 52%, 43% and 35% of images tagged with `newyork` are also tagged with `nyc`, `ny` and `newyorkcity`, respectively. As described in the previous chapter, this poses evaluation problems where models which simply promote synonyms achieve higher precision/recall scores than models which promote tag *novelty* and *diversity* in their rankings. In this work, we address this problem by clustering the tags in user images into related aspects, allowing for intent-aware metrics to be computed (e.g. αnDCG) instead of the traditional precision/recall metrics which ignore diversity.

3. Free Distribution One of the largest problems with these collections is that they are not available for distribution, making comparison with new recommendation models difficult. In our work, we download a manageable number of Flickr images which use the creative commons license, allowing for easy distribution.

6.4.3 Flickr-PTR

In this following section we detail the process used to build the Flickr-PTR collection, which aims to resolve these problems. As before, we refer to how we address each problem by referencing the problem number in parenthesis e.g. (1).

As with Flickr-AIA, we begin with the unfiltered collection containing 2M creative commons Flickr images (3), which was created by selecting the top 2k image results (containing location, time and sufficient annotation information) from the Flickr search API for 2k popular WordNet nouns. As before, we consider the user annotations as ground truth (1). The role of a training set in tag recommendation differs from image annotation, in that images can be categorised with a wide range of tags, whereas images in an AIA training set are only categorised for a small number of *visual* classes. Therefore, for Flickr-PTR, we chose *not* to remove the noisy tags from the collection allowing for a real-life evaluation scenario. Our main contribution, however, lies in our test collection, where tags are clustered into coherent aspects. In order to overcome the

discussed problems with synonyms, we cluster tags which describe the same aspect of 1,000 random images using a crowdsourced experiment. This experiment is discussed in detail in section 5.3, however, to summarise: we ask 197 "turkers" to group together (using a drag-and-drop web interface) Flickr tags which "refer to the same aspect" of 1,000 images. The results of this experiment show that more than over half of tags within images are redundant (i.e. synonyms of other annotations within the image).

By conducting this experiment, we are able to build a test collection where the ground truth describes *aspects* for each image (2), rather than tags, as required for diversification evaluation. The details of this experiment and the resultant testset are described in depth in Chapter 5.3.

6.5 Chapter Summary

This chapter highlighted a number of problems which exist in using *three* popular image annotation and *two* popular photo tag recommendation evaluation collections, thus addressing RQ6.1. Most importantly, synonyms exist in annotation ground truths for all collections, which may result in misleading performance measures. This problem was most prominently demonstrated for the IAPR collection, where we showed that annotation models may "under-perform" by up to 15% where a synonym is predicted instead of the exact annotation (e.g. predicting `sea` instead of `ocean`). Other issues included: (i) poor image quality (ii) lack of visual diversity (iii) copyright issues *etc*. The findings of this chapter bring into question the validity of those works which benchmark on these collections, of which there are many[5], and as a result we encourage researchers to consider these problems when evaluating future methods.

For photo tag recommendation evaluation, we also show that by training and testing on synonymous annotation sets, poor (i.e. non-diverse) tag recommendations achieve higher precision and recall measures than those which promote diversity in the top ranks. Again, due to the problems discussed, most prominently with crowdsourced tagged ground truths, we question the validity of the results in the discussed papers and also encourage researchers to consider these problems in future evaluations.

Due to the problems introduced in this chapter, we proposed two new evaluation collections, namely Flickr-AIA and Flickr-PTR, which aim to overcome these issues and are created with fair evaluation in mind, thus addressing RQ6.2. These collections have been made publicly available[6], allowing for future comparative studies to be carried out.

In today's evaluation framework, image annotation models and tag recommendation systems have mostly attempted to predict single, uni-gram terms. We hypothesise, however, that in order to build more sophisticated models, one should instead attempt

[5]The IAPR (Grubinger et al., 2006) and ESP (Von Ahn, 2009) papers are cited 228 & 426 times respectively

[6]Available at http://mir.dcs.gla.ac.uk/resources/

to annotate images with *phrases* or *descriptive sentences* (e.g. man running across grass), which more closely matches real world search queries as well as provides a deeper level of descriptive value. Aside from the exponential increase in complexity from an annotation perspective, evaluation frameworks will need to consider the challenges posed by these new ground truths. In particular, one must be able to equally represent synonymous *sentences*; an extremely difficult task due to the number of ways it is possible to describe the same scene in the English language (e.g. teenager jogging on a field). Therefore, we hypothesise that annotation test sets may become redundant as researchers look to to crowdsourcing for all testing purposes, where humans give a graded judgement for a model's predictions (e.g. "running man" is correct, but not as descriptive as "man running across grass").

Part III

Photo Recommendation

In Part III of this thesis we consider the role of *context* in the task of *photo recommendation*[a]. Specifically, we propose three new tasks in which we exploit the context a photograph is taken in, or exists in, during the recommendation process. In Chapter 7, we propose the task of visual event summarisation in which we attempt to retrieve images relevant for an event automatically detected on social media. In Chapter 8, we propose image popularity prediction where we attempt to identify whether an image is likely to be popular in the future, which could be exploited for content recommendation purposes. Finally, in Chapter 9 we explore various methods for the ranking of images collected on life logging devices in order to summarise a user's day.

[a]We use the terms recommendation and retrieval interchangeably as recommendation can be seen as a top-N retrieval task where no query exists i.e. a cold start.

Chapter 7

Visually Summarising Social Media Events

The following chapter is published in the following conference:

- Philip J. McParlane, Joemon M. Jose (2014); "Picture the scene..." Visually Summarising Social Media Events, *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM) 2014, Shanghai, China.* ACM, New York, NY, USA, pp1459-1468.

7.1 Introduction

Given the rise of Twitter and similar microblogging platforms in recent years, there has been a research focus on using their data to automatically detect news events[1] while they happen (McMinn et al., 2013; Sakaki et al., 2010; Weng and Lee, 2011). With the wealth, coverage, speed, lack of censorship and unbiased nature of microblog posts, they present many advantages over traditional journalistic media. Once an event is detected (i.e. tweets clustered), one must summarise it succinctly in order to convey the event topic to the user; many works have attempted to create short (i.e. 1 or 2 sentences) textual summarises to this effect (Aggarwal and Subbian, 2012; Sakaki et al., 2010; Sankaranarayanan et al., 2009; Sharifi et al., 2013), however, due to the the large scale of the data, automatically summarising these events is a non-trival task which often produces poor results e.g. informal language, irrelevant summaries *etc*.

In this work we exploit *images* to automatically summarise social media events. Images present a number of advantages over text for summarisation purposes; they are: (i) able to quickly convey an idea or atmosphere (e.g. consider the famous "Tiananmen Square tank man" image[2]) (ii) naturally "multilingual", in that even an illiterate person

[1] "An event is a significant thing that happens at some specific time and place" (McMinn et al., 2013)

[2] https://en.wikipedia.org/wiki/Tank_Man - last accessed on 18th July 2016.

is able gain much insight from an image (iii) can express what words cannot convey (e.g. consider watching a football games on TV vs listening to it on the radio). On the contrary, text is often: (i) slower and more laborious to digest in that we have to move our eyes and process each letter individually (ii) written in a different language for which the user cannot understand (iii) poorly written (e.g. online slang, character constrained microblog posts *etc*). In this work we develop a technique which can automatically identify the most relevant images on social media in order to best describe a news story which has began to "trend".

Automatically identifying relevant and representative images for events detected on microblogging websites presents a number of difficult challenges for researchers, however. Most importantly, how do we overcome noise present in these streams in order to *select* and *rank* the most relevant images for summarisation purposes? Images posted on these websites pose many problems as they are: (i) often irrelevant (e.g. internet memes, screenshots *etc*) (ii) often duplicates, or near-duplicates, of other posted images (iii) often lack diversity and capture the same "moment" (iv) or of low quality. Therefore, new methods of *selection* and *ranking* are required in order to overcome these problems. In this work, we propose a number of techniques which aim to maximise *relevance* as well as topic *diversity* in the rankings of images automatically collected for event summarisation purposes.

Not all events are suitable for event summarisation, however. For example, consider a meeting between world leaders which happens behind closed doors i.e. where there is no access for journalistic photographers. Therefore, in this work we also address this problem by considering which event types are most suitable for *visual* summarisation purposes.

In the following, we attempt to address the following research questions (RQ):

RQ7.1 Can we use images to automatically summarise events detected on microblogging streams?

RQ7.2 How can we effectively *select* and *rank* images relevant to a social media event? How do we overcome the challenges of noisy and irrelevant images?

RQ7.3 Does adding images alongside text improve the summarisation effectiveness? Do images help the user to identify an event's topic and key entities (i.e. people, location *etc*)?

RQ7.4 Which event "type" (e.g. sports, politics) is best suited for visual summarisation?

The rest of this chapter is as follows: firstly we detail works in event detection and summarisation in Section 7.2. Section 7.3 formulates the problem of visual event summarisation before further discussing image selection, ranking and presentation problems in Sections 7.4, 7.5 & 7.6, respectively. In Section 7.7 we describe our crowdsourced evaluation before discussing the results of this in Section 7.8. Finally, we summarise this chapter in Section 7.9.

7.2 Background Work

In the following section we detail the works in event detection and summarisation, discussing how our work differs from these publications.

Event Detection Event detection has been a research focus since 1998 when the Topic Detection and Tracking (TDT) project (Allan et al., 1998) began which aimed to automatically monitor and detect events in broadcast media. In recent years there has been a renewed focus on event detection with the rise of social media; much work has focused on overcoming the new challenges presented by detecting events on large-scale and noisy microblog data.

Sankaranarayanan et al. (2009) proposed the one of the first systems which aimed to detect breaking news and events from tweets. Using significantly filtered tweet streams, the authors applied clustering techniques weighted using a time-decayed cosine function. Similarly, Aggarwal and Subbian (2012) used clustering and growth rate thresholding in order to detect events on Twitter. Finally, Sakaki et al. (2010) attempted to detect tweets referencing natural disasters in order to issue early warning alerts. By using a simple keyword filtering technique, the authors were able to classify tweets as *event related* or *not* using a Support Vector Machine (SVM).

Automatically detecting events, is a related, but distinct problem from that of event *summarisation*. In event *detection*, the overriding goal is to cluster related tweets into coherent events as they happen. In event *summarisation*, the objective is to *succinctly* and accurately describe an event in a way which covers as many its different aspects as possible (Das and Martins, 2007). Effectively and concisely summarising a cluster of related tweets is a non-trival task, however, and as a result has been a research focus in recent years.

Event Summarisation One of the main applications of event summarisation, is that of automatically identifying key moments in *scheduled* broadcast television programs e.g. football matches, political events *etc*. Chakrabarti and Punera (2011) attempted to summarise structured and re-occurring sports events by deriving their underlying state representation using Hidden Markov Models (HMM). By first filtering noisy posts using various criteria, the authors found their underlying latent space before selecting summary tweets using a TF-logIDF representation. Similarly, Nichols et al. (2012) also attempted to summarise sporting events on Twitter by considering temporal volume spikes to determine key moments within an event. The authors first applied filtering techniques to Tweets, such as removing spam, off-topic and non-English posts, before extracting key sentences based on a number of grammar/language heuristics. In a similar work, Zubiaga et al. (2012) attempted summarisation of scheduled events on Twitter by first detecting sub-events through analysis of volume peaks. Key tweets were selected as event summaries using a term frequency and Kullback-Leibler divergence weighting scheme. Aside from sport, Shamma et al. (2010) exploited microblog

Human Summary	**Automatic Summary**
Chicago turns 177 years old	Happy 177th birthday to the best city in the world Chicago
Jury in Florida loud music murder trial stuck on murder charge	The college I want to go to would be in Florida
California drought sparks call to ban Fracking and protect water	Bieber dies in a car accident on highway in Hollywood

Table 7.1 Problems with SOTA text based summarisation (Sharifi et al., 2013) vs human summaries.

posts for the segmentation and summarisation of the 2008 USA presidential debate. The authors produced an interface which displayed automatically segmented video, trending topics and Tweet geolocations.

Summarising *scheduled* events, however, significantly reduces the complexity of the problem in that most of these events have (i) well defined start and end times (ii) well defined "moments" (e.g. football goals, political speeches *etc*), and (iii) well defined hashtags. Therefore, the stream can be more easily parsed and understood with specific prior knowledge (e.g. knowing when half time is *etc*). In our work we attempt the more challenging task of summarising *unscheduled* events which are (i) often unexpected (ii) cover a wide range of topics (e.g. from earthquakes to movie releases) (iii) and have no defined start and end times (meaning tweets are often posted long after the event finishes). Therefore, event summarisation methods must be generic enough to cover this range of topics and be able to deal with noisier, less specific data streams.

Recent research has also focused on the summarisation of these *unscheduled* events: Sharifi et al. (2013) developed a summarisation model which computed effective summaries by creating two partial summary graphs on each side of popular topic phrases. Further investigation, however, showed that an adaptation of the simpler TF-IDF algorithm produced summaries which were just as good, if not better. Marcus et al. (2011) introduced TwitInfo, a system for visualizing and summarizing events detected on Twitter. Specifically, they identified peaks within Twitter streams allowing users to explore events by geolocation, sentiment and popular URLs. Long et al. (2011) also proposed a similar summarisation approach which attempted to cluster posts on the Sina microblogging website by selecting topic words before building a graph based topic co-occurrence model for event tracking purposes; those posts which had the highest coverage of the cluster topics were used to produce summaries.

As previously discussed, using text to summarise events can result in poor performance due to the level of noise present on social media streams and complexity of the task, as can be observed in Table 7.1. This table compares summarises using a state-of-the-art model (Sharifi et al., 2013) against a human summary, when attempting to describe various "tweet clusters" (i.e. events) included in a recent large-scale

social media event detection collection (McMinn et al., 2013). As can be observed, the summaries range from overly *opinionated* (e.g. "Happy 177th birthday to the best city in the world Chicago"), to *irrelevant* (e.g. "The college I want to go to would be in Florida") to *incorrect* (e.g. "Bieber dies in a car accident on highway in Hollywood"). In our work we overcome these problems, as well as those already discussed, by introducing images, automatically selected, ranked and presented from noisy microblog streams, in the event summarisation process. Specifically, we propose a number of image selection, ranking and presentation methods in order to describe events automatically detected by a recent event detection model (McMinn et al., 2013).

Visual Event Summarisation Recent related work has attempted to create visual timelines of social events and celebrities using content posted on image and video sharing websites. Del Fabro and Boszormenyi (2012) attempted to summarise four major social events using images and video by querying Flickr and YouTube for relevant content posted between two given timestamps. They performed clustering in order to identify key "moments" and used view and like counts in order to select the most relevant content. Sahuguet and Huet (2013) attempted to build a visual timeline, using videos, in order to summarise major events for the lives of celebrities (e.g. Mark Zuckerberg). In their work, they used Google Trends[3] to extract important time segments and keywords which were used to retrieve relevant videos from YouTube for summarisation purposes.

Related to that of event summarisation is *video summarisation* where the overall objective is to effectively summarise, or shorten, (long) videos into shorter clips whilst maintaining as much of the semantics & story as possible. In 2007 the TRECVID Video Summarization workshop (Ove, 2007) was held in conjunction with the ACM Multimedia conference where the goal was to benchmark methods in content based video summarisation, specifically focusing on shot boundary detection, video search and automatic concept detection. In its first year, 31 different research teams attempted to produce short video summarises of British Broadcasting Corporation (BBC) series. The papers proposed for this workshop ranged from simple approaches which "fast forwarded" video (Christel et al., 2008), to sophisticated methods for detecting and shortening repetitive scenes, using (i) face detection (ii) camera motion (iii) and colour layout features (Naci et al., 2008). Other methods attempted to build storyboard type summarisations (Bredin et al., 2008) by identifying key shots using linear discriminant analysis. Video summarisation is a related, yet distinct task to that of event summarisation for a number of reasons: (i) firstly, both the source and output content type is different resulting in new challenges for researchers (i.e. video *vs* text, images + video) (ii) further, the content used in the video summarisation task is generally of higher quality (i.e. professional drama series) to that of content present of social media websites. In our work we instead consider the task of summarising *automatically detected events* using images from *noisy* social media streams.

[3]http://www.google.com/trends/ - last accessed on 18th July 2016.

7.3 Problem Statement

In event *detection*, the problem is to take a set S of streaming microblog posts and cluster (both semantically and temporally) into a set of events E which each contain a subset of posts S_e related to the event in question. In traditional event *summarisation*, the problem is to take these subsets S_e and produce a sentence which best describes the topic of the posts within. In this work, we propose *visual event summarisation* which, given a subset of posts S_e, attempts to select relevant images I_e related to the event before ranking them in a way which maximises relevance and diversity in the top ranks. From this rank, the top image(s) are used alongside text in a visual event summary. In the following sections we discuss our methodology for addressing the many problems involved in achieving automatic visual event summarisation.

7.4 Image Selection

Given a set of tweets related to an event S_e, the first challenge is to collect and select a subset of relevant and representative images. As discussed in the following sections, there are a number of problems identified with using images posted by users on Twitter:

1. **Lack of images:** despite the extensiveness of textual content, images make up only a small fraction of tweets posted on Twitter and as such, there often exist only few relevant images for smaller, localised events. To be able to effectively summarise events, however, we rely on a wealth of content; therefore, in this work, we also focus *beyond tweets* in order to gather images for our purpose, as described in section 7.4.1.

2. **Near-Duplicate images:** users post and retweet many duplicate or near-duplicate images on Twitter; in order to avoid summarising an event using identical images, an initial phase of near-duplicate detection must be taken out as described in section 7.4.2.

3. **Irrelevant Images:** users often post images which are either completely irrelevant, or are relevant but unsuitable for event summarisation purposes (e.g. internet memes, screenshots *etc*) as described in section 7.4.3.

7.4.1 Lack of images

Despite the wealth of *textual* content posted on Twitter, images make up only 4% of Tweets; therefore, for smaller events, collecting images solely from microblog posts may be insufficient in order to create a meaningful visual summary. In order to overcome this data sparsity problem, we extend our collection by extracting images from *websites* (i.e. URLs) contained within tweets referring to the given event. This has a number of advantages in that: (i) URLs posted in Tweets are often news websites

or blogs which generally contain high quality content and images (ii) URLs exist in 44.6% of tweets in our collection; therefore, there is a wealth of diverse content which can be used for our purposes.

Automatically extracting information from semi-structured data sources has been explored in the past by a number of works (Chang et al., 2006; Etzioni et al., 2008); however, selecting relevant images (with respect to the article) from websites presents a new challenging task. For example, images contained on websites are often irrelevant with respect to the content of the article (e.g. adverts, social buttons, thumbnails, logos *etc*). In order to select the most relevant images from URLs, we use the following heuristics in the selection process:

1. **Adverts:** images which are equal to the dimensions of standard web banner advertisements[4] are ignored.

2. **Irrelevant Graphics:** images with filenames containing the phrases "logo", "facebook", "twitter", "google" are also ignored. By doing so, most social media buttons and logos are filtered out.

3. **Thumbnails:** images which are less than 200px wide or 200px high are also ignored as they are too small for summarisation purposes.

4. **Image placement:** we also consider an image's placement on a webpage in the selection process with the hypothesis that images relevant to the article content will appear early in the HTML file. We therefore select only the first 5 images referenced within an HTML document. Further, we ignore all images referenced within the `<head>` of an HTML document apart from the `og:image` element; this tag was introduced in the Open Graph protocol[5] created by Facebook aimed at building a rich social graph of websites. The Open Graph describes the `og:image` tag as "an image URL which should represent your object within the graph" and is therefore suitable for selecting images which are most representative for a given website.

For each event, we therefore download the images which pass this criteria, from the websites referenced within the tweets. Although following these heuristics will result in a percentage of false positives, by filtering out images which are small, placed near the bottom of webpages, in advertisement format or have a filename which implies irrelevance, we are able to quickly and easily filter out the most irrelevant content without extensive visual analysis, which is desirable in large microblogging collections.

[4]http://commons.wikimedia.org/wiki/File:Standard_web_banner_ad_sizes.svg - last accessed on 18th July 2016.

[5]http://ogp.me/ - last accessed on 18th July 2016.

7.4.2 Near Duplicate Image Detection (NDID)

Due to the *sharing culture* present on microblogging websites, there exist a large amount of duplicate content online. The same image is often hosted on many different servers and retweeted in many different social circles. Further, unlike duplicate text content which can be easily matched, images may be different at a *file level* (e.g. filesize, filename *etc*), but almost identical at a visual/semantic level (e.g. taken by a different device, watermarked, compressed *etc*). Therefore, in order to avoid summarising events using duplicate images, we must be able to automatically identify and cluster them.

In our work we detect duplicate images and near-duplicate images using a popular hashing function technique (Tang et al., 2012). Hashing functions are used to generate *fixed-length output strings* which act as a shortened reference to its initial data (e.g. text documents, audio files, images *etc*). These functions were initially created for cryptographic purposes (Rivest, 1992), however, in recent years, their application has been experimented for near-duplicate image detection. For example, Chum et al. (2008) proposed two new image similarity measures using locality sensitive hashing (LSH). Specifically the authors proposed a method which used a weighted intersection of SIFT keypoints in image pairs in order to detect duplicates. Foo et al. (2007) instead compared the effectiveness of *dynamic partial functions (DPF)* and hash based counting techniques in order to detect visual duplicates within the image rankings of a commercial search engines.

In this work we used a related hashing method called the Perceptual Hash (pHash) which has been shown to give high detection accuracy for resized, cropped and exposure compensated images (Tang et al., 2012). The pHash function produces an output string, for both colour and black & white images, by first normalising an image through a number of interpolation and filtering phases before converting to the YCbCr space and extracting invariant "moments", which are concatenated together in order to represent an image. We choose to detect visual duplicates using a hashing function due to its high performance while maintaining low computational expense in extraction and matching phases. By adopting this method, we ensure its scalability in large microblog collections.

For each image in our collection we first compute its pHash[6] string before employing single pass clustering on all images in our collection, using the hamming distance for comparison purposes. Specifically, images are added to an existing cluster if their hamming distance is small enough ($T < 8$ as suggested by existing work (Tang et al., 2012)), otherwise the image is added to a new cluster. Using this method, duplicate images with the following alterations are grouped together (Tang et al., 2012): (i) brightness adjusted (ii) contrast adjusted (iii) gamma corrected (iv) 3x3 Gaussian lowpass filtered (v) JPEG compressed (vi) watermark embedded (vii) resized (viii) rotated slightly. Tang et al. (2012) demonstrate that the pHashing approach is able to sig-

[6]Using the tools available at http://www.phash.org/ - last accessed on 18th July 2016.

nificantly outperform a popular Singular Value Decomposition (SVD) baseline (Kozat et al., 2004), achieving a 0.1% false positive rate over 4,950 comparisons for images with these adjustments.

7.4.3 Irrelevant Images

Aside from the problem of filtering duplicate images, the major problem we must overcome in the selection process is that of removing *irrelevant* images. Before we discuss irrelevant images on social media, however, we define what we believe an "ideal" image to be for summarisation purposes: (i) firstly, it should be of high quality i.e. taken on a high resolution device (ii) topically relevant i.e. capturing a key moment in the event (iii) and follows "recommended photography practises" (Kelby, 2012) (e.g. lighting conditions *etc*). The majority of images on social media, however, are taken by amatuer photographers, often on smartphones, and therefore do not fall into this category. Instead our problem becomes that of *reducing* the amount of noisy images present in our visual summaries. Exploiting the wisdom of the crowd (i.e. images which are posted/retweeted many times in our case) has been used in other works as one such quality control measure (Kittur and Kraut, 2008), however, there exist a wealth of images on social media which are *popular yet irrelevant* for our purposes (e.g. memes) forcing us to consider other methodologies. In figure 7.1 we observe 3 image types which cover a broad range of unsuitable images posted on Twitter:

1. **Memes:** there exist many "memes" (or funny images with captions) which are often popular and topically relevant (e.g. (a) of Figure 7.1 which refers to the 2014 conflict between Ukraine and Russia) but are not suitable for event summarisation purposes.

2. **Screenshots:** there exist many screenshot images (e.g. (b) of Figure 7.1), which may contain tweets relevant to the event but are not useful for event summarisation.

3. **Reaction images:** there are also a wealth of reaction images posted (e.g. (c) of Figure 7.1) which are used to evoke the user's emotion but are neither relevant or suitable.

It is therefore in our interests to automatically filter out these images due to their unsuitability for summarisation purposes. For (b) and (c), they share a common feature in that they are almost entirely computer generated, or "synthetic". Therefore, in order to identify these images, we implement a synthetic image detection model introduced by Wang and Kan (2006). In their work, the authors extract colour and edge histograms from 600 images[7], which have been manually identified as *synthetic* or *natural* (i.e. real photographs), in order to train an SVM classification model. The authors demonstrate the effectiveness of their approach by conducting an extensive experiment where over 95% classification accuracy is achieved. Therefore, in our work

[7]http://wing.comp.nus.edu.sg/npic/ - last accessed on 1st February 2015.

(a) (b) (c)

Fig. 7.1 Unsuitable images for event summarisation

we employ this technique to discriminate between real life photographs and computer generated images, crawled from microblog posts. More specifically, we firstly extract the following features from images in our collection:

1. **Colour histogram (CH):** we compute colour histograms by "binning" the pixel frequency of each colour range in both RGB and HSV colour spaces, as adopted by Wang and Kan (2006). We hypothesise that colour histograms are effective for differentiating between real and computer generated images as photographs tend to contain random distributions of colour where synthetic images (i.e. covering many memes, screenshots and reaction images) are more likely to employ flat colour & linear gradients in their composition. Further, synthetic images are more likely to contain brighter, more saturated hues which are generally not present in nature.

2. **Edge Histogram (EH):** we also compute edge histograms using the popular MPEG-7 standard, proposed by Manjunath (2002), which captures the local edge distribution of images, categorising into vertical, horizontal, 45 & 135 degree and non-directional "bins". We also hypothesise the effectiveness of edge histograms for our purpose as real photographs are more likely to contain random edge distributions and are less likely to contain continuous, perfectly straight lines (i.e. edges) which are often present in logos, graphics and other computer generated media.

Representing each image as a normalised concatenation of the CH (in RGB and HSV spaces) and EH feature vectors, we train a two-class SVM using 5-fold cross validation on the synthetic image classification collection proposed by Wang and Kan (2006). This collection contains 350 *synthetic* diagram images (e.g. computer graphics), 250 *synthetic* map images and 15,600 *natural* photographs. Due to the unbalanced natured of this collection (i.e. 600 synthetic images vs 15,600 natural photographs), we randomly select 600 natural photographs in order to create a 50/50 split between the two classes. From our experiments, the best performance is achieved using the Radial basis function (RBF) kernel with parameters $C = 2$ and $\gamma = 2^{-3}$, where 94.5% accuracy for classifying images as synthetic or natural is achieved. Using this technique, we are

able to remove screenshots, reaction images and other unrelated computer generated content automatically from our overall image test collection.

Automatically detecting *meme images* (e.g. (a) of Figure 7.1) is more difficult, however, as memes often contain different textual captions which hide much of the background image causing near-duplicate detection methods to fail. Instead we employ local features which are more able to robustly match between heavily altered images by instead attempting to match small invariant "*patches*", opposed to the full image. Specifically, we match all images in our collection against a crawled database of 587 meme background pictures based on the popular Scale-invariant feature transform (SIFT) (Lowe, 2004). Originally introduced for matching between different views, angles and lighting conditions of an object or scene, we instead employ SIFT features in order to identify memes. We consider an image to be a meme, if it matches a certain percentage of its keypoints with a meme background image. We select a 25% matching threshold due to previous work (Foo and Sinha, 2007) which experimented with the percentage of keypoints matched for cropped images (we consider the crop alteration to be the most suitable for our purpose as meme images with their captions cropped leave only the background image). Using this approach we are able to automatically identify and filter out meme images from tweets.

7.4.4 Selection Systems

In the follow section we discuss two *selection* strategies for gathering images from Tweets and related websites. We define the following selection systems which are referenced in later sections of this chapter:

1. **Filtered Twitter images (TWR):** In this system, we select only the *most popular* (i.e. frequency) images referenced within *Tweets* related to events. These images are selected using the techniques described in Section 7.4.2 & 7.4.3. We hypothesise that images which exist frequently within event tweet clusters will often be relevant for the given event as users often retweet the most interesting (i.e. pseudo relevant) content (i.e. exploiting the wisdom of the crowd). Further, by filtering memes, visual duplicates and computer generated images, we hypothesise that only the most relevant images will be selected.

2. **Filtered Website images (WEB):** In this system we select only the *most popular* (i.e. frequency) images from *websites* referenced within Tweets related to events. These images are selected using the techniques described in Section 7.4.1. We hypothesise that popular images crawled from related websites will be of *high quality*, as photographs found on news websites & blogs tend to be curated for some basic quality metrics (e.g. image size *etc*) by the website owner/author. Further, by eliminating adverts, thumbnails, computer generate media and images found towards then end of web pages, we hypothesise that only the most relevant images will be

selected.

7.5 Image Ranking

In the following we discuss methods for ranking on a subset of images I_e which have been filtered using the various selection methods proposed in the last section. The overall goal in our work is to rank the most *relevant* images in the top positions whilst maintaining their visual *diversity*. In this work, we exploit the wisdom of the crowd in order to indicate which images are most *relevant* for a given event by considering the most retweeted images and websites, as discussed in section 7.5.1. Depending solely on popularity can result in poor summarisation performance, however, as often users prominently post about the most interesting aspect of an overall event, thus ignoring many of its subtler sub-topics. For example, consider generating a summary of a football match between Real Madrid and Barcelona where Lionel Messi scores a freekick from 30 yards. Due to the amount of interest that would be generated for this single moment, we would expect different angles and aspects of the freekick to dominate the images posted on social media. Therefore, by selecting the most popular images, even though they may all be relevant, we may exclude many other sub-topics (e.g. other goals, free kicks, red cards *etc*) in the event's summary. We believe that an event summarisation model should depict a diverse representation of the entire event instead of focusing solely on one popular moment. Therefore, in this work we also employ semantic clustering techniques in order to maximise the diversity of images ranked in the top positions, as detailed in section 7.5.2. Finally, we discuss our different ranking approaches in section 7.5.3.

7.5.1 Promoting Relevance

By exploiting the "wisdom of the crowd" we aim to rank images by descending relevance. Specifically, we approach this by ordering images by their computed *popularity* as we hypothesise that popular images, referenced within relevant tweets describing an event, are most likely to be relevant for summarisation purposes. As images are collected from both Twitter and URLs resources, we first define "popularity" for each:

1. **Twitter:** for images posted directly to Twitter, we model popularity as the number of times an image is posted or retweeted, a measure which has been previously used for computing relevance (Duan et al., 2010; Ravikumar et al., 2012). To overcome the discussed problem of image duplicates and near-duplicates, however, popularity measures are computed on clusters of duplicates (instead of individual images), summing their occurrence to compute their *popularity*.

2. **Websites:** similarly, we model popularity as the number of times an image appears on one of the websites related to the event in question. As before, popularity is com-

puted as the sum of occurrences within its *duplicate cluster*, instead of the individual image.

Our *first* approach considers only an image's popularity; in traditional information retrieval, this method is similar to ranking documents by term frequency (TF). One of the major problems with using this popularity based model is that it is easily broken by spam content, a major problem on microblogging websites (Benevenuto et al., 2010). There exist many spam bots which use sophisticated techniques in order to go unnoticed (e.g. multiple URL redirects in links referring to the same website/image, posting dynamic content related to trending Twitter topics *etc*); therefore, if an event detection method determines these spam tweets, referring to the same website/image, as relevant for multiple events, their content will be promoted in the top ranks. In order to combat this, we attempt to capture the significance of an image related to a given event by capturing its inverse document frequency (IDF), computed as:

$$IDF(I) = log(|E|/|E_i|) \tag{7.1}$$

where $|E|$ is the number of events in our collection and E_i is the number of events containing the given image I. IDF scores are computed for each visually unique image[8] with respect to the number of events it exists in; therefore, spam images (i.e. those existing in many events) will expect to have a low IDF score with those existing in few events achieving a high IDF score. In our *second* approach, we therefore compute an image's TF-IDF score, which considers its popularity, normalised by its IDF value.

7.5.2 Promoting Diversity

Ranking purely on popularity will encourage the high ranking of relevant content, however, it does not ensure these images are *diverse* (i.e. they cover multiple aspects of an event). By not considering diversity, images in the top ranks are more likely to be taken of a single sub-event, especially if this sub event is sufficiently popular. For example, consider the 2014 Oscars which produced the most retweeted image in Twitter history i.e. the Ellen DeGeneres Group Selfie[9]. Within minutes of this image being posted, a number of comical photoshopped versions[10] began flooding Twitter and were subsequently retweeted. Due to their subtle differences, these images would not be captured as duplicates and would potentially also appear in the top ranks alongside the original. For summarisation purposes, it is in interest of the summarisation model to cover as many different "moments" within the overall event as possible, rather than different angles or versions of the same sub-event. For the Oscars 2014 event, in an optimal

[8] A visually unique image refers to those which exist in the same pHash cluster, as described in section 7.4.2.

[9] https://twitter.com/TheEllenShow/status/440322224407314432 - last accessed on 18th July 2016.

[10] http://goo.gl/j4Xcc7 - last accessed on 18th July 2016.

case the Ellen DeGeneres Group Selfie would be summarised by a single photograph, with other sub-events, such as the award for Best Picture, also summarised by a single image.

In order to achieve this, we propose a *semantic clustering* approach which aims to maximise the diversity, or number of "moments", covered in the top rankings. In the collection used in this work (McMinn and Jose, 2015), tweets are not only clustered into high level events, but also *within* their event (referred to as sub-event clusters), which are computed as follows:

> "Clustering is commonly used in event detection, however it is also inherently slow for large numbers of documents. We address this using the premise that tweets discussing an event must describe at least one of the named entities involved in the event, and partition tweets based upon the entities they contain... For the purpose of clustering, this can be thought of as having a unique Inverted Index for each named entity. For each named entity e in tweet d, a list of tweets D is retrieved from the inverted index for e and the maximum $TF-IDF$ weight cosine similarity score is calculated between d and each tweet in D. If the maximum score is above a set threshold (usually in the range 0.45-0.55 (Petrović et al., 2010)), then d is added to the same cluster as its nearest neighbour. If the nearest neighbour does not already belong to a cluster, then a new cluster is created containing both tweets and assigned to entity e. The new tweet is then added to the inverted index for entity e."
> *(McMinn and Jose, 2015)*

We therefore exploit these *semantic* sub-clusters in order to maximise the diversity of ranked images. We achieve this by selecting the highest ranked image in each of the largest K sub-clusters by scoring images using the $TF-IDF$ model, where TF counts the number of times a visually identical image exists within an event and IDF demotes the popularity of images across *all* events. By selecting the deemed optimal image for various sub-clusters (i.e. moments), we attempt to maximise both *relevance* and *diversity* in the rankings.

7.5.3 Ranking Systems

In the following section we define our various approaches which rank images from Tweets and related websites, as well as a combination. Specifically, we use various techniques to maximise relevance and diversity in the top ranks, as described in the previous sections. We define the following ranking systems which are referenced in later sections of this chapter as follows:

1. **Normalised Popular Twitter Images (P-TWR):** In our first method, we rank images collected from *Twitter* based on their descending popularity (i.e. frequency)

using the $TF-IDF$ weighting scheme as described in Section 7.5.1. As the most popular images found on Twitter will be those which have been frequently retweeted and considered highly interesting by users, we consider this ranking method a *strong baseline*. Furthermore, this ranking technique could even be considered as an *industry strength* baseline as this method is used by Twitter in order to rank images on their search page[11].

2. **Normalised Popular Website Images (P-WEB):** Using the same notion, we also rank images from related *websites* based on their descending *popularity* (i.e. frequency) using the same $TF-IDF$ weighting scheme described in Section 7.5.1. We hypothesise that if a website is retweeted multiple times it is likely to contain both relevant textual and visual content, which is suitable for our purpose. Further, as the related URLs are often blogs or news websites, we hypothesise that we will increase visual diversity as journalistic websites often detail different aspects, or offer differing opinions, of a overall news story. We therefore also consider this method a *strong baseline*.

3. **Combined Normalised Images (P-COM):** We also rank images based on a combination of both the P-TWR and P-WEB systems. The combination strategy is as follows: the top 10 ranked images from each source are weighted as $1/p$, where p is their position in the ranked list. The two lists are merged, summing the weights if an image exists in both rankings. The resulting list is then ordered by this descending weight. We hypothesise that combining sources will be complementary due to the difference in media type found on Twitter and related websites.

4. **Semantically Clustered Twitter Images (S-TWR):** As detailed in the previous section, we employ semantic clustering techniques in order to attempt to maximise the visual and semantic diversity of images in the top ranks. In this system, we select the highest ranked *Twitter* image (using the discussed $TF-IDF$ scoring approach) in each of the largest semantic clusters, as described in Section 7.5.2. By gathering *popular* images from the *largest* (i.e. pseudo-relevant) semantic clusters, we hypothesise that images in the top ranks will cover a wider range of sub-topics.

5. **Semantically Clustered Website Images (S-WEB):** Following this notion, we also semantically cluster images collected from websites by selecting the highest ranked *website* image (using the $TF-IDF$ weighting scheme) in each of the largest semantic clusters, as described in Section 7.5.2.

6. **Combined Semantically Clustered Images (S-COM):** Finally, in this system we combine the S-TWR and S-WEB rankings methods using the same merging scheme as used in P-COM, for which we hypothesise combining sources will be complementary.

[11] https://twitter.com/search?q=%22Derrick%20Rose%22&src=tren&data_id=tweet%3A636717627080069121 - last accessed on 18th July 2016.

7.6 Event Summary Presentation

For event summarisation presentation, the overall goal is to describe an event in the most succinct and descriptive way such that the user is able to quickly understand it and its entities (e.g. people, location *etc*) whilst engaging their interest.

Existing event summarisation methods (Marcus et al., 2011; Nowak and Dunker, 2010; Sharifi et al., 2013) create summaries by selecting the most representative tweet or sentence within a tweet cluster. Although this achieves descriptive succinctness, it often fails to capture the people, location and "story" of the event. Wordclouds alleviate this problem, however, by listing the most significant terms within a body of text. As a result, wordclouds have been used to summarise web search results (Kuo et al., 2007) and maps (Wood et al., 2007) in previous work where there exists a similar information overload problem as in event summarisation. In this work, we therefore employ tag wordclouds, as a baseline, in order to summarise events.

Despite succinctly describing significant entities, wordclouds fail to "set the visual scene" of an event. We hypothesise that by including images within event summaries, we will be able to both capture key entities whilst "setting the scene" for the user.

7.6.1 Presentation Systems

We define a number of presentation approaches which are referenced in later sections of this chapter:

1. **Title and Tweet (TTW):** we present the user with a single sentence title (extracted using the state-of-the-art title summarisation approach described by Sharifi et al. (2013)) and the most re-tweeted tweet describing the event (see (a) of Figure 7.2 for an example). We consider this approach as our summarisation *baseline*.

2. **Title and word cloud (TWC):** we present the user with the same single sentence title (as used in TTW) as well as a wordcloud describing the most significant terms contained within tweets describing the event (see (b) of Figure 7.2 for an example). The wordcloud includes the top 20 terms for the event in question, ordered by descending $TF-IDF$ score (where TF is the occurrence of a term within an event and IDF is its inverse document frequency across the entire collection). We consider this approach as a stronger summarisation *baseline*.

3. **Title and image (TIM):** in our experimental approach, we present the same single sentence title (as used in TTW & TWC) as well as the top ranked image, computed using the S-COM ranking strategy, in the summarisation interface (see (c) of Figure 7.2 for an example).

Designing interfaces which are considered "fair" for the evaluation of event summaries is a difficult problem due to the near infinite number of designs possibilities and content combinations. Before we created the interface, we first considered the

(a) Title and Tweet (b) Title and Wordcloud (c) Title and Image

Fig. 7.2 Event Summarisation Presentation Experiment

types of content which can be used to summarise an article, with the methods previously discussed being adopted (i.e. sub-title, word cloud & image). From this, we aimed to create interfaces that were as clean & consistent as possible by maintaining font sizes, colours, dimensions *etc*, whilst being familiar to the user; many websites employ a title & sub-title/image combination[12] with word clouds equally as popular online[13]. Therefore, by selecting these popular interface modalities and keeping the styling simple & consistent, we hoped to fairly evaluate the value of each content *type* for summarisation purposes.

7.7 Experiments

In the following, we discuss details of the collection used in this study in section 7.7.1 as well as details of its extension by extracting images from URLs posted in tweets. Focus then shifts to constructing a test set for evaluation purposes in section 7.7.2 before describing the metrics used and evaluation procedure in sections 7.7.3 & 7.7.4.

7.7.1 Collection

In this work, we use the Twitter collection introduced by McMinn et al. (2013) which was originally built for event *detection* purposes, opposed to event summarisation. At the time of download, the collection contained details of over 500 automatically detected events using state-of-the-art techniques. In our work, we consider the 50 largest events (in terms of tweets) for visual summarisation as we hypothesise that visual event summarisation is best suited for the most popular events where there exists sufficient photographs as well as multiple sub-topics (e.g. various bands playing at a festival). The collection used is as follows:

[12]For example, https://news.google.com adopts both interface styles - last accessed on 18th July 2016.

[13]http://goo.gl/eksG2n - last accessed on 18th July 2016.

1. **Tweets:** our collection contains 365k tweets posted by 220k different users over a 1 month period from 11th February 2014 till 11th March 2014. Of these tweets, 4% contain uploaded images and 44.6% contain a website URL. The low percentage of tweets containing images motivates our need to collect additional pictures from the large number of websites referenced in these tweets.

2. **Events:** these 365k tweets describe 50 distinct events as automatically identified by the model described by McMinn et al. (2013). The event detection method clusters on highly filtered tweets which contain at least a single *entity* (identified by an extension of the Stanford Parser (De Marneffe et al., 2006)). Each event contains on average 7,280 tweets which are further grouped into 135.6 sub-clusters, on average.

We extend this collection[14] by extracting images from websites referenced within tweets. Table 7.2 describes the images collected from each source and their characteristics. As can be observed, images collected from tweets and websites have very different characteristics with respect to size, visual category (i.e. synthetic vs photographs) and number of duplicates.

	Twitter	**Website**	**Overall**
Images	13k	534k	547k
Unique images	7.2k	60k	67.2k
% Duplicates	45.3%	88.9%	87.8%
% Synthetic	35.6%	45.1%	46.0%
% Photographs	64.4%	54.9%	54.0%
% Memes	<0.1%	<0.1%	<0.1%
% Images < 200px wide/tall	2.8%	71.4%	70.4%
Average image height	554px	168px	177.3px
Average image width	548.5px	280.6px	287.1px
% Pass Filter	63.4%	15.6%	16.8%

Table 7.2 Images collected from tweets and URLs in tweets

7.7.2 Building a Test Set

In order to determine the selection & ranking effectiveness of the various systems, described in Sections 7.4.4 and 7.5.3, we use judgements from crowdsourced users in order to create a test set of *relevant images* and allowing for the computation of traditional IR metrics to be measured. The details of this crowdsourced experiment are explained in the following paragraphs:

For this work, we use the paid crowdsourcing service CrowdFlower[15] (CF) for evaluation purposes. CrowdFlower provides a number of advantages over platforms

[14] Available at http://mir.dcs.gla.ac.uk/resources/ - last accessed on 18th July 2016.

[15] http://www.crowdflower.com - last accessed on 18th July 2016.

(e.g. Amazon Mechanical Turk[16]) such as the ability to post to multiple markets, a template editor and most importantly, quality control mechanisms. Due to these advantages, CrowdFlower has been used for the evaluation of many recent related studies (Finin et al., 2010; Hong and Baker, 2011).

In our evaluation, we attempt to determine the relevance (with respect to its event), image quality and category of images selected in each of our 8 systems. Therefore, for the images in the top 5 ranks of each system, we ask workers a number of questions to create this ground truth. If a worker accepts our experiment, or Human Intelligence Task (HIT), they are presented with the following instructions:

Task Description

> *You are presented with an image and an event title (describing a "trending topic" on Twitter). For each image & event title, you are asked to answer the following 3 questions:*

Q1. Is this image *relevant* for the event?
The event is a trending topic on Twitter. Please tell us if you think the image is relevant for the given event or not.
Possible answers: Likert Scale from 1 (not relevant) to 4 (relevant).

Q2. What *category* would you describe the *image* as?
Please select a suitable category from the following (see Figure 7.3 for the examples shown to the user).
Possible answers: (i) High quality photograph, (ii) Average quality photograph, (iii) Low quality photograph, (iv) Computer generated image.

Q3. What *category* would you describe the *event* as?
Please select a suitable category from the following:
Possible Answers: (i) Business and Economy (ii) Law and Politics (iii) Science and Technology (iv) Arts, Culture and Entertainments (v) Sports (vi) Disasters and Accidents (vii) Armed Conflicts and Attacks (viii) Miscellaneous

Crowdsourced workers are then asked to complete a number of *test questions* in order to judge their understanding of the task. These test questions follow the same task description as detailed in the previous paragraph, however, we *manually* specify the "*correct answer ranges*", prior to the user's submission, in order to identify spamming users (i.e. those who simply click answers without understanding what is required of them). We set these "correct answer ranges" generously so that there is some margin for error or opinion. For example if the user was asked to determine the *visual quality* (i.e. Q2) of (a) in Figure 7.3, we would accept both "high quality photograph" or "average quality photograph". In our experiment, we ask users to answer the discussed three

[16]https://www.mturk.com/ - last accessed on 18th July 2016.

questions for 8 separate images, in which they must answer *all* within the predefined "correct answer ranges" in order to continue to the HIT. By doing so, we ensure the highest submission quality and that users have understood the task required of them.

Further, CrowdFlower rates the trustworthiness of workers based on their previous work; we require that workers have at least a "Level 1 Contributor[17]" status. Also, as the gathered tweets are in English, we only allow users from countries which have an English speaking majority to take part. Finally, in order to ensure that no single user can have an overriding influence on our results, we limit judgements to 100 per user.

On passing this initial test, workers are presented with 5 images sequentially, which at each stage they are required to answer all 3 questions. They are allowed up to 15 minutes to complete all 15 questions otherwise their answers are ignored and they are not paid. Those users who successfully complete the HIT are paid $0.04 for their work. In total, 198 different workers accepted our HIT of which 114 (57.6%) passed the 8 test question phase making it to the evaluation stage, with 84 users (42.4%) failing. We hypothesise that the fairly low pass rate, which has been observed in other work (Sayeed et al., 2011), is in part due to the high number of test questions as well as the fact that the user requires some knowledge of current trends; nevertheless, conducting the test question phase ensures that we gather the most relevant judgements and has been shown to significantly increase the quality of crowdsourcing judgements (Oleson et al., 2011). For Q1, workers evaluated 1,695 images in total (with each judged by 3 users) for all 50 events. For these images, 606 were judged to be relevant for the event in question (i.e. those selected with an average Likert scale score of greater than 2.5), with workers judging image relevance with an average variance of 0.43 (i.e. less than half an option difference on a 4 point Likert scale). On average, 12.6 images were deemed relevant for each event, with 48 out of 50 events containing at least 1 relevant image.

For Q2, images were categorised as 27% high quality, 29% average quality, 18% low quality and 26% as computer generated with a *Crowdflower confidence score* of 0.75. This confidence measure, as computed by CrowdFlower[18] and used throughout this chapter, describes "the level of agreement between multiple contributors (weighted by the contributors' trust scores), and indicates our confidence in the validity of the result". Finally, for Q3 events were categorised with the following frequencies: Sport (20), Law & Politics (11), Arts Culture and Entertainments (9), Armed Conflicts and Attacks (4), Disasters and Accidents (2), Miscellaneous (3) and Science and Technology (1) where workers categorised with a Crowdflower confidence score of 0.87.

[17]"High performance contributors who account for 60% of monthly judgements and maintain a high level of accuracy across a basket of CF jobs."

[18]This confidence score ranges between 0 (no agreement) and 1 (complete agreement) - for more information, see: https://goo.gl/xxnPwy - last accessed on 18th July 2016.

High quality photograph (a)

Average quality photograph (b)

Low quality photograph (c)

Computer generated image (d)

Fig. 7.3 Image categories for Q2 in our crowdsourced experiment

7.7.3 Metrics

To evaluate our image selection and ranking methods, we use the following traditional information metrics, related to those defined in Chapter 2.3.2, comparing those images in the top 5 ranks against those deemed relevant by users in our test set. These metrics are as follows:

1. **Precision (P@N)**: The percentage of relevant images amongst the top N, averaged over all runs.
2. **Success (S@N)**: The percentage of runs, where there exists at least one relevant image amongst the top N returned.
3. **Mean Reciprocal Rank (MRR)**: Computed as $1/r$ where r is the rank of the first relevant image returned, averaged over all runs.

Due to the problems of image duplicity, as described in Section 7.4.2, we also use intent-aware metrics proposed in text based diversification. In our work, we use diversification metrics to discount those systems which promote duplicate images in the top ranks. Therefore, by "grouping" duplicate images into clusters (using the technique described in Section 7.4.2), or "sub-topics", we are able to apply diversification metrics to measure the "coverage", or diversity, of a system's rankings. These metrics are as follows:

1. **α-Normalised Discounted Cumulative Gain (α-nDCG@5):** this metric computes the usefulness, or gain, of an image based on its position in the ranked list. The parameter α balances the importance of relevance and diversity. We compute following common practice (Clarke et al., 2009) where α-nDCG is computed with $\alpha = 0.5$, in order to give equal weights.
2. **Intent-aware Expected Reciprocal Rank (ERR-IA@5)**: for this metric, the contribution of each image is based on the relevance of images ranked above it, by computing the ERR for each sub-topic, with a weighted average computed over sub-topics (Clarke et al., 2009).

7.7.4 Evaluating Summary Presentations

In order to compare the effectiveness of our event summary presentations, discussed in Section 7.6.1, we carry out a second crowdsourced experiment. This evaluation takes the form of a short survey asking users to select their interface preference for a number of criteria. If a worker accepts our experiment, they are presented with the following instructions:

Task Description

> *You are presented with 3 different interfaces describing an event (or trending topic on Twitter). Please answer the following questions regarding the interfaces:*

Q1. Which interface most effectively *summarises* the event?
Possible answers: Left, Centre, Right

Q2. Which interface most quickly helps you understand the involved *people*, *events* and *location*?
Possible answers: Left, Centre, Right

Q3. If these were tiles on a news website, which would you most likely click on?
Possible answers: Left, Centre, Right

Q4. Which of the 3 was your eye initially attracted to?
Possible answers: Left, Centre, Right

Workers are presented with the 3 interfaces shown in Figure 7.2 describing a single event. They are then required to answer all 3 presented questions. They are allowed up to 4 minutes to complete all 4 questions otherwise their answers are ignored and they are not paid. Those users who successfully complete the HIT are paid $0.02 for their work.

As this experiment captures user opinion, it is more difficult to capture spammers through traditional test questions and honeypot (Eickhoff and Vries, 2013) tests. Therefore, we instead focus on collecting as much high quality data from as many different users as possible. We achieve this by requiring each event to be judged by 5 different workers whom are categorised with the highest CrowdFlower rating (i.e. Level 3 Contributors[19]). Additionally, workers are only able to take our survey once in which they judge only a single event; by doing so we ensure opinion from as many users as possible, putting faith in the "wisdom of the crowd". Finally, as before we only accept HITs from users in English speaking countries.

[19]"Highest performance contributors who account for 7% of monthly judgements and maintain the highest level of accuracy across an even larger basket of CrowdFlower jobs."

7.8 Results

In the following we compare the effectiveness of our image selection & ranking approaches in section 7.8.1 before detailing the results of our second crowdsourced experiment which evaluated our 3 presentation approaches in section 7.8.2.

7.8.1 Image Selection & Ranking

Table 7.3 compares the selection and ranking effectiveness of our various systems. As some events were deemed by crowdsourced workers to have no relevant images (from those proposed in all 8 systems), we present two results tables: the top details evaluation metrics for all events, whereas the bottom table details those evaluation scores for events containing at least 1 relevant image.

For all events

System	P@1	P@5	S@5	MRR	α-nDCG	αnERR-IA
TWR	0.300	0.356	0.820	0.482	0.481	0.431
WEB	0.260	0.292	0.680	0.432	0.386	0.355
P-TWR	0.280	0.368	0.800	0.484	0.489	0.437
P-WEB	0.260	0.324	0.700	0.441	0.415	0.378
P-COM	0.300	0.336	0.720	0.462	0.455	0.408
S-TWR	0.400	0.420	0.820	0.565	0.560	0.518*
S-WEB	0.480*	0.500**	0.800	0.596*	0.547	0.521
S-COM	0.480*	0.480**	0.780	0.588	0.591*	0.555*

For events with at least 1 relevant image

System	P@1	P@5	S@5	MRR	α-nDCG	αnERR-IA
TWR	0.313	0.371	0.854	0.502	0.501	0.448
WEB	0.271	0.304	0.708	0.450	0.402	0.370
P-TWR	0.292	0.383	0.833	0.504	0.509	0.455
P-WEB	0.271	0.337	0.729	0.459	0.432	0.393
P-COM	0.313	0.350	0.750	0.482	0.474	0.425
S-TWR	0.417	0.437	0.854	0.589	0.583	0.539*
S-WEB	0.500*	0.521*	0.833	0.621*	0.570	0.543
S-COM	0.500*	0.500**	0.813	0.614	0.616*	0.578*

Table 7.3 Comparison of image selection & ranking approaches. Underlined values indicate the highest performing systems for the given metric. Statistical significance results against the best performing BL (TWR) are denoted as * being $p < 0.05$ & ** being $p < 0.01$.

Comparing Sources Images collected solely from Twitter are more relevant than those collected from related websites, achieving best relevance and diversity measures

when ranking *purely* on popularity (i.e TWR and P-TWR). This highlights the effectiveness of explicit user feedback (in the form of retweets) in promoting relevant content. Unlike Twitter, images on websites cannot be retweeted and therefore do not achieve the same selection and ranking performance when ordered purely on popularity (alternatively, retweets can imply relevance); for both WEB and P-WEB systems, lower performance is achieved for all metrics in comparison to TWR and P-TWR. This opens the question of whether user feedback present on websites (e.g. social media shares, web page popularity/authority) can be exploited for the purposes of image selection and ranking in visual summarisation models.

Semantic Clustering Highest performance is achieved using semantic clustering approaches where gathering filtered images from both sources and selecting the most popular images from the largest semantic clusters achieves highest image selection/ranking accuracy (i.e. S-COM for both diversification metrics), thus addressing **RQ7.2**. In particular, S-WEB improves significantly (+55%) over the P-WEB method. We hypothesise that this is due to URLs existing in a high percentage of tweets (44.6%) meaning that there will exist many websites in the long-tail of this distribution which are either spam or irrelevant to the event. Therefore, by only selecting images from those websites which exist in the largest clusters, we collect images from only the most focused and relevant websites related to the event, thus reducing the influence of images collected from irrelevant websites in the long tail.

Combining Sources Although performance decreases in the P-COM approach, we achieve significant increases for both diversification measures when combining sources in S-COM. This highlights that images from Twitter and websites can be complementary and can increase the coverage of visual sub-topics in an event.

Image Quality Using the judgements made in our first crowdsourced experiment, we are able to compare the quality of images suggested in the top ranks by each system, as detailed in Table 7.4. We observe that images from websites are consistently of higher quality than those collected from tweets. Although the quality of image drops when combining from both Twitter and websites (i.e. P-COM & S-COM), these approaches still contain over 30% *more* high quality photographs in the top 5 ranks in comparison to those systems using only Twitter images (i.e. P-TWR & S-TWR).

Event Category Summarisation Suitability Table 7.5 compares the type of event and the number of images judged relevant by users in our crowdsourced experiment. Given the high percentage of relevant images judged for "Armed Conflicts and Attacks", "Sports" *etc* we can firstly infer that there exist more relevant images online documenting these types of events and therefore they are most suitable for visual event summarisation purposes, addressing **RQ7.4**. On the contrary, "Law & Politics" events

System	High Quality	Avg Quality	Low Quality	Computer
TWR	0.259	0.312	0.259	0.171
WEB	0.330	0.343	0.096	0.231
P-TWR	0.279	0.324	0.239	0.159
P-WEB	0.371	0.382	0.092	0.155
P-COM	0.316	0.376	0.144	0.164
S-TWR	0.203	0.264	0.339	0.195
S-WEB	0.399	0.321	0.129	0.151
S-COM	0.304	0.296	0.220	0.180

Table 7.4 Comparison of image quality in the top 5 ranks for each system. Bold values indicate the highest value for each criteria.

Category	# Relevant	# Judged	% Relevant
Armed Conflicts & Attacks	82	169	0.485
Arts & Entertainments	87	209	0.416
Disasters & Accidents	27	84	0.321
Law & Politics	86	301	0.286
Miscellaneous	13	171	0.076
Sports	296	726	0.408

Table 7.5 Event category vs the # of images judged relevant. "Business & Economy" and "Science & Technology" categories are omitted due to insufficient data.

are more likely to happen "behind closed doors" and therefore are more difficult to capture for amateur photographers.

7.8.2 Event Summary Presentation

For the survey results gathered from users judging the 3 event summary interfaces, as detailed in Section 7.7.4, we can conclude our initial hypothesis that images can benefit the summarisation of events in a number of aspects, as detailed in Figure 7.4. In particular, users are able to identify entities (Q2) and understand the content of events more easily when summarised using images (Q1), in comparison to both baselines (i.e. TTW and TWC), thus supporting **RQ7.1** and **RQ7.3**. Even when significant entities are listed in the form of a wordcloud, photographs are more effective for describing people and locations existing in events by allowing the user to "picture the scene". Finally, by embedding images, event summaries are more likely to catch the attention of the user (Q4) as well as engage their interest (Q3), in comparison to both baseline approaches.

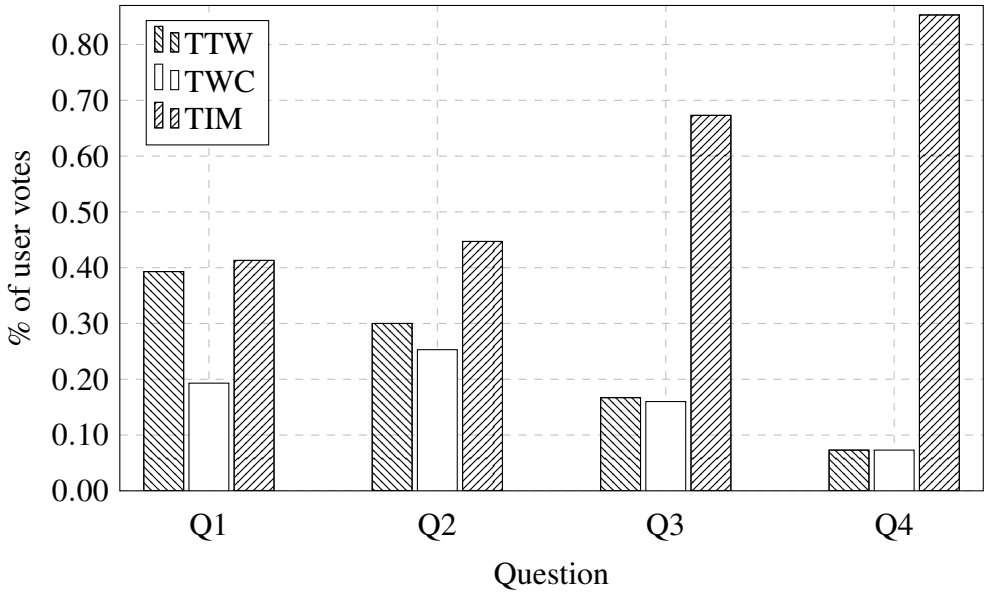

Fig. 7.4 Survey results for event summary presentation evaluation

7.9 Chapter Summary

In this chapter we proposed the automatic visual summarisation of events detected on noisy social media streams. In particular we aimed to address the following research questions:

RQ8.1 Can we use images to automatically summarise events detected on microblogging streams?

RQ8.2 How can we effectively *select* and *rank* images relevant to a social media event? How do we overcome the challenges of noisy and irrelevant images?

RQ8.3 Does adding images alongside text improve the summarisation effectiveness? Do images help the user to identify an event's topic and key entities (i.e. people, location *etc*)?

RQ8.4 Which event "type" (e.g. sports, politics) is best suited for visual summarisation?

For RQ8.2, we demonstrated that by combining filtered images from multiple sources (i.e Twitter and related URLs) and selecting the most popular images from the largest semantic clusters we were able to identify the most *relevant* and *diverse* images for summarisation purposes. For RQ8.3, the findings of our crowdsourced experiment showed that images both increased user engagement and helped users to understand the content, people & locations present in events. Bringing these results together, we conclude the we able to satisfy the requirement of automatic visual event summarisation as discussed in RQ8.1. Finally for RQ8.4, our crowdsourced experiment suggested that those events which are most accessible to the public (e.g. sports events) are those which are best suited for visual summarisation.

To the best of our knowledge, this work is the first in visual event summarisation and overcomes a number of the key problems associated with selecting/ranking images for real world events from social media. We have focused largely on creating techniques with low computational complexity (e.g. single-pass clustering) thus addressing scalability issues which is extremely important in the social media domain and should be considered at all stages of any future works. There exist many other avenues for improvement, however, for which we were not able to address within this work. For example, many tweets & images contain pornographic/adult content or have an extremely satirical tone which may be unsuitable for summarisation purposes. Additionally, we have not considered videos in this work which may offer a richer resource for summarisation purposes.

This work proposes the exploitation of context in an image *and* text environment (i.e. social media), thus allowing for textual features to be considered. In the follow chapter, we explore the value of contextual cues in the *absence of text*; specifically, we aim to employ context for a new "image popularity prediction" task, where there exists little or no textual evidence to draw upon.

Chapter 8

Image Popularity Prediction

The following chapter is published in the following conference:

- Philip J. McParlane, Yashar Moshfeghi, Joemon M. Jose (2014); "Nobody comes here anymore, it's too crowded"; Predicting Image Popularity on Flickr, *Proceedings of ACM International Conference on Multimedia Retrieval (ICMR) 2014, Glasgow, Scotland (UK).* ACM New York, NY, USA, pp385-392.

8.1 Introduction

Retrieval paradigms are constantly changing and evolving based on new user requirements and applications. Popularity prediction is one such paradigm which has been proposed in recent years in order to overcome the important information overload problem present in Web 2.0 applications. Given that a user cannot view all of the data uploaded to a social media website, recommending the most interesting[1] content is an important task. Additionally, predicting the popularity of a web object (i.e. document, image *etc*) also allows a company or website to make decisions more strategically, e.g. manage their resources (e.g. servers) better and target their advertisements more effectively. Overall, both the user benefits from a more enjoyable experience and the company benefits from a monetary gain.

Due to the importance of popularity prediction, many works have been proposed in bookmarking (Jamali and Rangwala, 2009; Lerman and Hogg, 2010), video (Szabo and Huberman, 2010) and social (Bandari et al., 2012; Hong et al., 2011) domains. More recently, work has attempted to predict the popularity of images on Flickr (Niu et al., 2012). However, this approach, as well as the others mentioned, rely on the interactions of users (e.g. clicks, ratings *etc*) to predict the popularity of a given item/document. As discussed throughout this thesis, the majority of images are only

[1] Although some users may not find popular content as the most interesting, we assume that popular content will satisfy the interests of the majority of users

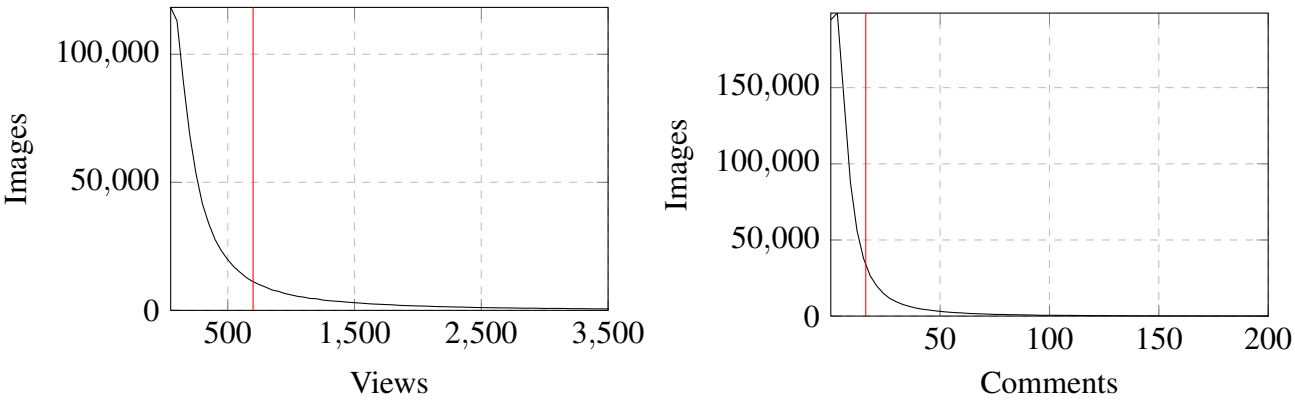

Fig. 8.1 View (left) and comments (right) distribution. Red line = top 20% popularity threshold

ever annotated with less than four tags (Sigurbjörnsson and van Zwol, 2008) and initially contain no interaction data. Therefore, an approach which requires no or little interaction data is desirable. This work is an attempt towards alleviating this problem as well as providing a different type of application than the previous chapter, where we are no longer able to rely on textual information. Due to this, the following chapter provides a more difficult scenario where we must rely more heavily on contextual features in the absence of text. Additionally, in order to further highlight the benefit of the features we proposed in Chapter 3, we apply them to this task of photo popularity prediction.

In this work we propose two measures of image popularity based on the number of *views* and *comments* an image has. We consider these two measures as popularity metrics as they reflect user interest in an image and cover both implicit (i.e. views) and explicit (i.e. comments) aspects of user feedback. Figure 8.1 shows that both these measures follow a power law distribution (MIR-FLICKR 1M collection (Huiskes et al., 2010)) where the *majority* of the images get little or no attention (i.e. the long tail of the distribution) and the minority of them receive a high level of attention (i.e. the head of the distribution). Understanding the underlying factors that result in an image falling into this minority (i.e. popular and therefore pseudo-interesting) can potentially be very beneficial to social networking websites in order to promote interesting, yet currently unpopular, content to their users as well as manage their resources & advertisements better. In order to do so, we qualitatively analyse each feature for predicting an image's popularity.

The contributions of this work are as follows:

1. We consider the exploitation of context, content and user features, which can be inferred in a cold start scenario, for the task of image popularity prediction.

2. We introduce measures of image "popularity" based on the number of views and comments an image has had.

3. We exhaustively combine and compare these features through extensive exper-

imentation, testing on the MIR-FLICKR 1M image collection (Huiskes et al., 2010).

The rest of this Chapter is organised as follows: firstly we detail works in document popularity prediction in Section 8.2. We define image popularity in Section 8.3 before introducing the collection and proposed features for popularity prediction in Section 8.4. Our experimental procedure and results are then detailed in Sections 8.5 and 8.6 before concluding in Section 8.7.

8.2 Background Work

In the following we detail related works in the field of text based popularity prediction before considering those in the multimedia domain. Finally, we motivate our reasoning for proposing this task within this thesis.

Popularity Prediction on the Web Document popularity prediction was first proposed for web documents where one attempted to predict the future popularity of a given article. Lerman and Hogg (2010) attempted to predict the popularity of such content based on early interaction data using a stochastic model. Testing on a Digg[2] collection, the authors showed that an article's popularity can be predicted from its initial user votes. Jamali and Rangwala (2009) exploited the implicit network behaviour on Digg to predict the popularity of submitted content. Specifically, the authors derive features from comments and social network data in a classification and regression framework. Chen and Zhang (2003) predicted the perceived popularity of web content with respect to a user in order to reduce loading time. Szabo and Huberman (2010) studied the popularity of YouTube videos and Digg posts showing that the early view patterns reflect long-term user interest. These works focus on the prediction of web documents and articles which are able to draw evidences from extensive textual data however. In our work, we instead consider the new area of *photo popularity prediction* which would allow image sharing websites to better manage their content & advertising as well as recommend "interesting" images to users in order to increase user engagement/retention. In our case, we focus on predicting image popularity in a cold start, where there exists few annotations - a common problem faced by image sharing websites (Sigurbjörnsson and van Zwol, 2008).

Popularity Prediction on Social Networks An important domain where interest has increased in recent years is that of user generated content (Web 2.0). Due to the overwhelming amount of content that is published every second online, popularity plays an important content filtering role. For example, the growth of Twitter has opened up a new area for forecasting document popularity. Hong et al. (2011) studied the importance of retweets for the prediction of popular messages on Twitter. Specifically, the

[2]http://www.digg.com/ - last accessed on 18th July 2016.

authors attempted to classify tweets based on a number of content, temporal and context features. Bandari et al. (2012) exploited properties of tweets, such as the source of the article, the category, subjectivity in the language and the entities mentioned, to predict their popularity. Similarly, some works have focused on the popularity of images on image sharing websites, such as Flickr. Niu et al. (2012) introduced a weighted bipartite graph model, called Incomplete Network-based Inference (INI), to predict image popularity based on network relationships. This work differs from ours in that the authors use a collaborative filtering approach which relies on (i) the previous interests of the users (ii) and the previous similarity of images/users. We do not depend on such information and instead focus on extracting evidences from an image's visual appearance and context, aspects which are available in a cold start.

Cha et al. (2009) studied how the popularity of pictures evolves over time, showing that even popular photos propagate slowly and that they do not spread widely. Valafar and Rejaie (2009) focused on indirect fan-owner interactions in photos on Flickr, showing that there exists no strong relationship between an image's age and popularity and that photos gain the majority of their fans in the first week in which they are uploaded. These works however focus only on network and interaction data thus failing in a cold start scenario i.e. where there exist no or little tag/interaction data. In our work we instead exploit an image's visual appearance, context and user context, which is readily available for all images, for the task of popularity prediction.

Aesthetic Prediction Related to image popularity, recent works have also considered an image's aesthetic value: Dhar et al. (2011) attempted to select images with the highest aesthetic value in a collection. In particular the authors studied the effect compositional, content and sky-illumination attributes have on perceived aesthetics, evaluated on 16,000 images in a crowdsourced experiment. Katti et al. (2008) considered aesthetics in images from a cognitive science perspective. Categories existing in image interestingness, such as colour and structure, are defined with their cognitive load computed through experimentation on 30,000 Flickr images. These works, however, consider an image's aesthetic value and how an image's visual appearance affect this. In our work, we instead focus on what makes an image *popular*, which differs from measuring aesthetic value, as aesthetically pleasing images are not always popular (e.g. an image of a horrible war scene is always aesthetically displeasing yet often popular).

Task Motivation As previously discussed, we propose this task of photo popularity prediction as it presents a more difficult scenario than the previous chapter due to the absence or lack of textual information. Based on this, we focus more heavily on exploiting context for this task and in particular employ our features proposed in Chapter 3 to further highlight their value in a scenario differing to that of photo tag recommendation.

8.3 Measuring Popularity

We define popular content as those items which achieve the highest user engagement; in order to determine an image's popularity, we consider two aspects of an image's interaction log for this purpose:

1. **Comments:** the number of user comments a given image has received contributes to its popularity. We use comment count as a measure of popularity due to its adoption by many other popularity prediction works (Jamali and Rangwala, 2009; Tatar et al., 2014, 2011). We consider this feature *"explicit"*, requiring effort from the user which we believe will make it a more reliable measure of popularity. We classify an image as having a *high* or *low* number of comments. The computation of these classifications is detailed in the following section.

2. **Views:** the number of views a given image has received also contributes to its "popularity". Again, we use view count as a measure of popularity due to its adoption by many other popularity prediction works (Arapakis et al., 2014; Niu et al., 2012). We consider this feature *"implicit"*, requiring little effort from the user and as a result consider it to be a less reliable notion of popularity (e.g. view count can more easily manipulated using page refreshes *etc*). We classify an image as having a *high* or *low* number of views. The computation of these classifications is also detailed in the following section.

In order to determine what constitutes "high" or "low" views/comments, we split our collection in two, using the Pareto Principle (or 80-20 rule) to compute thresholds, as used by existing work (Cha et al., 2007). The Pareto Principle is often used to describe the skewness in a distribution. In our work, we use this principle to select the threshold to split between images with *high* (20%) and *low* (80%) comments & views. An image is classified as having *high* views or comments if it exists in the top 20% of the given population. The thresholds for views and comments are shown in the long tail distributions in Figure 8.1 (i.e. $>$=700 views and $>$=16 comments).

In image popularity prediction, we aim to predict whether a new image will receive a *high* or *low* number of views and comments in the future. Therefore, the problem can be formalized as that of binary classification based on a number of features. In the following section we summarise a number of features representing an image which were proposed in Chapter 3 and have been adapted for our purposes.

8.4 Collection and Features

In our work, we classify images in the MIR-FLICKR 1M dataset (Huiskes et al., 2010) based on the features initially proposed in chapter 3 in order to further highlight their value in a new task other than photo tag recommendation. In the following subsections, we revisit these features by summarising the three different types, as proposed

in section 3.2. Finally, we propose a new feature category based on the tags present within the images, as used in our oracle approach. The features used in this chapter are as follows:

1. **Image Context (Section 3.2.2):** firstly we compute features from an image's context, focusing on *when* and *how* an image is taken. Specifically, we classify images based on their *time*, *day*, *season*, *device*, *orientation* and whether the *flash* fired. The technique used to categorise images for each class is described in section 3.2.2. Finally, in addition to these features, we also classify images based on their pixel size as we hypothesise that an image's original size could provide useful evidence when determining it's "quality" (i.e. "pseudo-popularity"): images are classified as *large*, *average* or *small* based on the size of the original image in pixels. We select thresholds so each set contains a similar number of images.

2. **Image Content (Section 3.2.3):** secondly, we compute features based on *what* an image looks like i.e. its appearance. Specifically, we classify images based on two visual *scenes* (referred to as scene #1 and scene #2), the number of *faces* present and the dominant *colour*. These classifications are identical to those defined in section 3.2.3.

3. **User Context (Section 3.2.4):** thirdly, we compute features based on *who* photographed an image. Specifically, we classify images based on the user's (i.e. uploader) *gender*, *account* type, number of *contacts* and number of *images* uploaded. The technique used to categorise images for each class is described in section 3.2.4.

4. **Tags:** finally, we compute features based on *what* an image is taken of. Specifically, we represent each image in our collection based on its user's annotations: for a given image, for a number of random tags (m) added by the user, we compute the TF-IDF vector (as described in Chapter 2.3.2) as a textual representation of a given image. In this work, we consider the cold start scenario testing with 1, 2 and 3 random tags from an image (i.e. $m = \{1,2,3\}$); when $m > 1$, we combine the TF-IDF vectors using a simple linear combination (i.e. averaging each element value). This feature is a 32,865 (i.e. the number of tags) dimensional vector of real values based on the output vector.

For the image context, content and user context features f, we define each as a c dimensional binary vector based on its classification, where c is the number of possible classes for feature f. For example, if an image is taken in the *evening* (i.e. the *time* feature), its binary representation would be $(0,0,1,0)$.

8.5 Experiments

In the following, we detail our evaluation in which we compare the value of our various features, as well as their combination, for the task of photo popularity prediction. Specifically, we discuss our evaluation procedure in section 8.5.1 before detailing our various experimental systems in section 8.5.2.

8.5.1 Evaluation Procedure

We investigate whether the number of views and number of comments can be predicted, given the image's context, content and user context. For this purpose, we classify images based on the described *TF-IDF* textual feature, the *seven* image context features, *four* content features, and the *four* user context features defined in Section 8.4. For the number of views and number of comments, we transform the values for each metric into a binary classification (+1/-1 or high/low), by using the method described in Section 8.3.

We learn a model to discriminate between the two classes using SVMs trained with a radial-basis function (RBF) kernel, which, based on our analysis, in the majority of cases, outperformed polynomial kernel SVMs. We also tried other models such as Bayesian logistic regression and decision trees but they underperformed with respect to the SVMs. We test on 1,000 randomly selected images from the MIR-FLICKR 1M collection using 10-fold cross validation. These images are taken by 784 users, viewed on average 519 times and commented on average 9.1 times.

8.5.2 Systems

Due to the novel nature of this work and research area, there exist no standard baselines to compare against; instead we focus on proposing and comparing the performance of various naive, oracle and experimental strategies. Firstly, we propose a *naïve random baseline* which is a popular approach used in other related multimedia works (Li et al., 2009) where there exist no standard benchmark. Secondly, we propose an *oracle approach* which we consider our "best case" performance when there exist sufficient evidences (i.e. textual content). As we instead aim to predict in a cold start scenario, we consider benchmarking against this approach as a difficult baseline. Our final two *experimental approaches* (EXP) are those proposed in this chapter which predict based on contextual features, tested individually as well as in combination. All of our systems are defined as follows:

BL **Naïve Baseline:** Due to the lack of work in image popularity prediction we compare against a naïve baseline which predicts image popularity with 50% accuracy, based on our test collection containing a 50/50 split of popular vs unpopular images. Benchmarking against a random baseline is a popular approach for

novel works where there exist no related studies (Li et al., 2009), thus motivating our adoption.

BL **Oracle Approach - Text(m):** In our oracle approach we classify based on textual evidence within the image's annotations; specifically, we classify images based on the TF-IDF representation of a number of random tags (m), testing with 1, 2 & 3 tags as previously discuss. As discussed we consider this the "oracle" approach as it predicts image popularity using interaction data (i.e. tag annotations) explicitly added by the user, a feature omitted from the majority of images on Flickr (Sigurbjörnsson and van Zwol, 2008).

EXP **Ind(f):** This system classifies images based on an individual feature, f, allowing for their effectiveness comparison. Specifically, we train/test our classifier based on the vectorial representation for a given feature.

EXP **Comb(l):** Finally, we also consider the combination of image and user features. This is achieved by concatenating the given vectors l, where $l = \{context, content, user\}$ representing the different feature types. For example, Comb(context) is the system with classifies based on the combination of all context features. We choose to concatenate the features due to its adopted by many other multimedia classification works (Ayache et al., 2007; Snoek et al., 2005).

8.6 Results

We believe it is important to be able to effectively predict whether an image will become popular in the future in order to be able to more accurately recommend *new* and *interesting* content to social media users, thus engage their interest and increasing website traffic & profits. In the following section we detail the results of our experiment for image popularity prediction. Firstly, we consider which of our metrics is most suited for popularity prediction purposes before considering the performance of each feature in isolation and combination.

Popularity Measure Effectiveness Figure 8.2 shows the classification accuracy averaged over all images in the test set. Firstly, it can be observed that we are able to classify an image's popularity, achieving accuracy of up to 76% when classifying comment count and 59% when classifying an image's view count. Therefore, we highlight that the number of comments is more highly correlated with the discussed features than image views. The explicit nature, and higher effort required, for commenting in comparison to viewing an image, makes comment count a more reliable measure of image popularity; viewing an image only requires a click, whereas commenting requires viewing an image, constructing an opinion and inputting a textual message. Further,

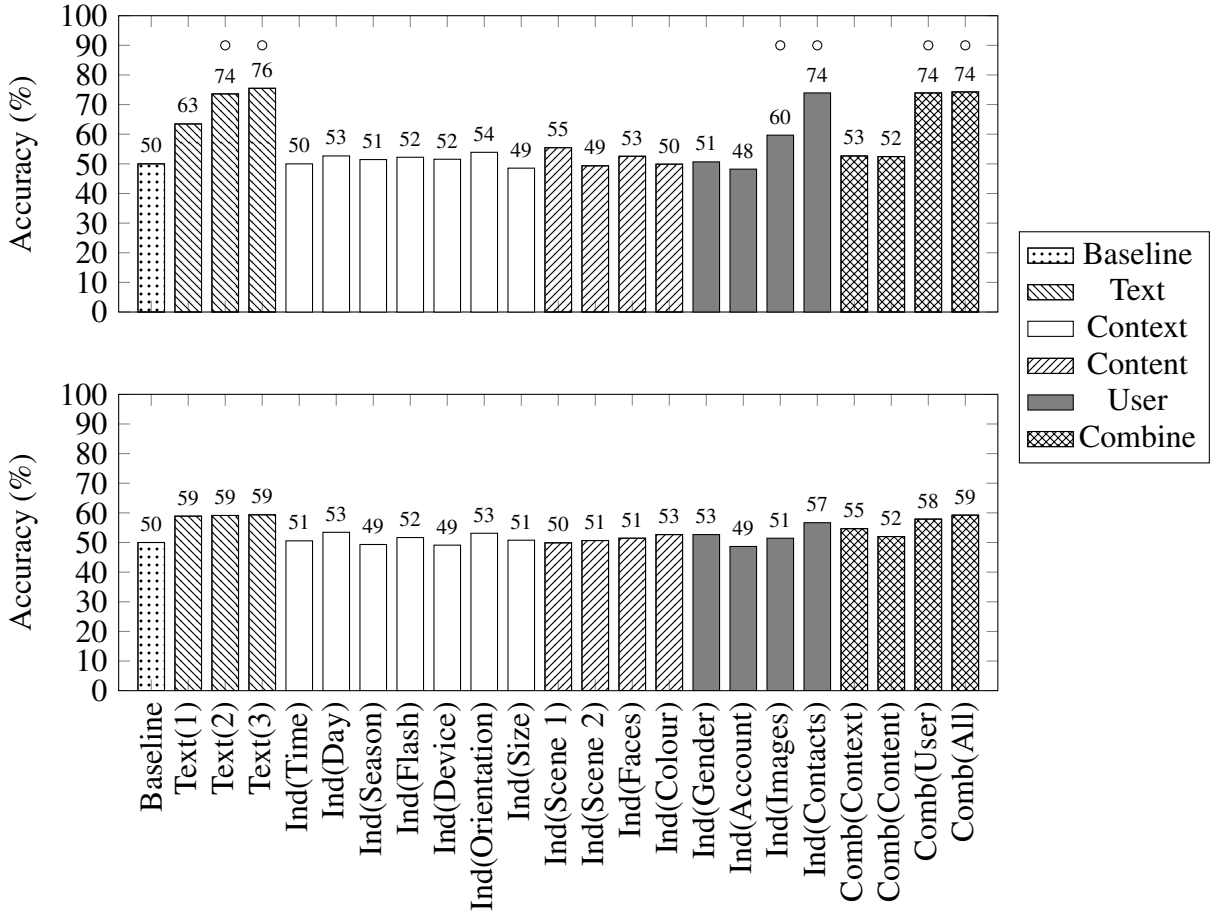

Fig. 8.2 SVM accuracy when classifying comments (top) and views (bottom). Statistical significance results against the baseline are denoted with an "o" above the bar for $p < 0.05$

this aspect is captured in the lack of statistically significant results when classifying for view count, motivating our notion that this metric is noisy and can be easily influenced by: users refreshing habits, bots/crawlers *etc*. From our results, we observe that an image's popularity most prominently depends on both the annotations (i.e. image "topic") as well as the activeness & social significance of the user.

Image Popularity Topic Analysis In Table 8.1, we compute the most significant tags (as employed in Chapter 3), sorted by descending order, for images with high/low comments and views. This is computed as the fraction of images tagged with tag t in images classified as x, minus the fraction of images tagged with t in all images, for the MIR-FLICKR 1M collection. By doing so, we identify the tags which occur significantly more in a subset of images sharing a common feature, than in the global set.

From table 8.1, we observe that highly viewed images tend to be images of people, especially women, (i.e. `girl, portait, woman`) in comparison to rarely viewed images which often depict `nature` scenes. Despite nature images being the lowest

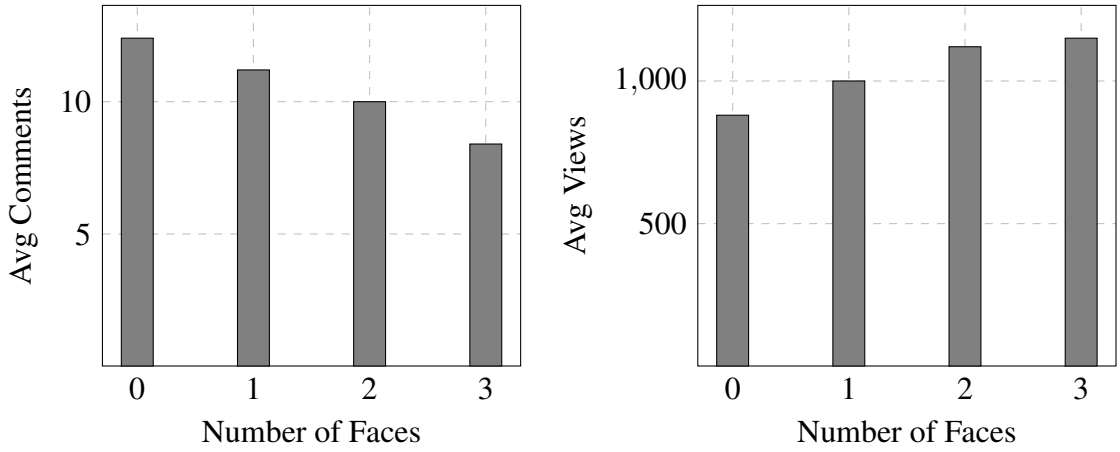

Fig. 8.3 Comments (left) and views (right) distribution vs the number of faces in an image.

	x	**Top Tags (t)**
Views	*High*	girl, portrait, woman, hdr, explore
	Low	flower, macro, cat, flowers, nature
Com	*High*	abigfave, aplusphoto, anawesomeshot, nature
	Low	2008, california, art, graffiti, sanfrancisco

Table 8.1 The most "significant" tags for high/low comments and views, where *significance* for an annotation is defined as the percentage of images tagged within a given subset (e.g. S_{male}), minus the percentage tagged in the full collection Due to space constraints, "com" = Comments.

viewed, they are commented the highest. This highlights that there exist many low quality nature photos (i.e. the long tail), yet few high quality nature photos which provoke much discussion and interest (i.e. the head of the distribution). This highlights the need to be able to predict where an image lies in the popularity distribution in order to be able to promote higher quality images to the user. Finally, our hypothesis that user comments are a more reliable measure of popularity is confirmed in Table 8.1 where the highest commented images contain many Flickr awards[3] (e.g. `abigfave, aplusphoto, anawesomeshot`).

Individual Feature Performance As observed in figure 8.2, for *context* the day type and orientation are the most discriminative features for both comments and view prediction; whereas for content, the features do not follow a trend for both metrics. All content features are able to offer some prediction improvement over our baseline, however, the improvement is minimal. We hypothesise that this is because images that grab the attention of the user are different, in some way, from the majority; therefore the visual appearance of images cannot be effectively used for popularity prediction. How-

[3]Flickr awards are given for images which have gained much attention (i.e. their image is deemed to be of the highest quality)

ever, we do identify a meaningful relationship between the number of faces present in an image and its popularity, as shown in Figure 8.3. We observe that images with less faces attract less views but have more comments; conversely those images with many faces (e.g. party photos) contain many views but few comments highlighting a high *browsing* and low *discussion* motivation for images with multiple people.

By far, the most important feature for image popularity prediction, however, concerns the user themselves where the number of contacts and images they have correlates highly with view and comment count. We observe that an image's popularity is linked closely to the user's popularity and activity, opposed to their contents and context. Further, by relying solely on the number of contacts a user has, we are able to achieve comparable popularity prediction performance in comparison to the case where multiple tags are present, overcoming the cold start scenario and highlighting the correlation between contact and comment count. Finally, the number of images a user has uploaded is also strongly correlated with image popularity, where by exploiting this feature we achieve 60% classification accuracy.

Combination Performance Combining evidences from an image's context, content and user context gives the best results in most cases. Specifically, combining all three gives the highest popularity prediction when comparing against all features individually, and the other combination approaches. Further, we observe that by combining all three evidences, we are able to match or outperform (except Text(3) for comments) the case where tags are present for popularity prediction, thus highlighting the merit of our approach where there lacks textual evidence.

8.7 Chapter Summary

Predicting the popularity of a web object has become an important task in recent years for social media websites in order to filter an ever expanding data set and maximise company profits. Instead of relying on interaction and textual data, however, as adopted by existing work, this work considered the challenging task of image popularity prediction in a *cold start scenario* i.e. where there exists no or little textual/interaction data. Instead we focused on exploiting our *contextual* evidences, originally proposed in section 3.2, for this new task of image popularity prediction. By doing so, we were able to highlight their value in a different paradigm, outside that of photo tag recommendation.

The findings of our experiments showed that we were able to predict, with up to 76% accuracy, whether an image will receive high or low user comments in the future. Specifically, the context surrounding a user, opposed to the *image* context, was seen to be the most effective feature set, where we showed that the popularity of an image is closely related to the user's popularity and activity level on Flickr, as well as its topical content (i.e. tags). Further, by combining evidences from the various feature sets (i.e. context, content and user) we achieved highest popularity prediction accuracy

highlighting that the features are complementary for this purpose and reliable in a cold start. Overall, our experiments demonstrated our initial hypothesis that non-textual evidences can indicate an image's future popularity as well as highlighting the value of textual annotations for popularity prediction

This work opens up a new area of prediction popularity for images as well as number of interesting research avenues which were unexplored due to time constraints, for example: (i) can "rules" from painting, art & design (e.g. composition, the "golden ratio", symmetry, patterns *etc*) be employed for predicting an image's popularity? (ii) can other, more elaborate visual analysis techniques (e.g. colour distribution analysis, contrast analysis *etc*) also be employed to more accurately predict a potentially "popular" image? Aside from exploring visual properties, the results of this work indicate that *the user* is the most important "feature" in determining an image's popularity. In particular, the number of contacts the user had on Flickr strongly correlated with an image's popularity; so instead future work should focus on predicting whether a user's network is likely to grow in the future, rather than predicting an individual image's future popularity.

In the previous two chapters we have proposed the exploitation of contextual image features for *large scale* social media collections where we are able to build models based on extensive training sets. In the following chapter, we instead explore the value of context on *much smaller*, personal lifelog collections; by doing so, we hope to demonstrate the value of our features in situations where there exists limited training data.

Chapter 9

Lifelog Summarisation

The following chapter is published in the following conference:

- Soumyadeb Chowdhury, Philip J. McParlane, Md. Sadek Ferdous, Joemon M. Jose (2015); "My Day in Review": Visually Summarising Noisy Lifelog Data, *Proceedings of ACM International Conference on Multimedia Retrieval (ICMR) 2015, Shanghai, China.* ACM New York, NY, USA, pp607-610.

9.1 Introduction

Not only has there been a rise in the number of images uploaded *online* in recent years, there has also been a new wave of personal image gathering, called Lifelogging, which presents new challenges for researchers. Lifelogging represents a way of digitally recording data (referred to as lifelogs) which capture a user's experiences, in varying amounts of detail, for a variety of purposes, using a lifelogging device. These devices capture important experiences in our daily routine without the need of explicit interaction due to their hands-free nature. Recently, a new wave of technological advancements and wearable devices such as those by Google Glass, Autographer, ParaShoot, and Narrative Clip has promoted lifelogging from a niche area of computer research to that of mainstream adoption. The growth in number of users capturing data using lifelogging devices has resulted in large personalized archives and has opened up a new area of multimedia retrieval research. It is challenging to manage, analyze, index and visualize streams of such multimodal information derived from lifelogging sensors, which can be noisy, error-prone and with gaps in continuity due to sensor calibration or failure (Gurrin et al., 2014; Melucci, 2012).

The advantages of hand free logging and photography has its disadvantages however; as no human interaction is involved in the capturing process, collections often consist almost entirely of noise. Providing abstractions over such data is an open research problem which becomes more challenging when the lifelogs are captured by

multiple sources. For example, in our scenario, images were captured using a wearable camera and locations captured using a GPS device requiring a prior stage of calibration to be taken out. Given that we can capture heterogeneous signals regarding a user's life (e.g. photographs, temporal-spatial streams *etc*), the challenge is whether we could combine such disparate lifelog streams to generate meaningful abstractions of such noisy data.

Most prior works (Doherty et al., 2013, 2012) have segmented unprocessed lifelog data into meaningful units called *events*, which are temporally related sequences of lifelog data over a period of time, with a defined beginning and end. Some studies have formed such meaningful units using GPS locations, in addition to temporal information obtained from lifelog images (Aizawa et al., 2004; Gurrin et al., 2013; Li et al., 2016b; Naaman et al., 2004). In this context, the most recent study reported by Kikhia et al. (2014) used location information to structure lifelogs based upon the notion of clustering interesting places, i.e. locations where the lifelogger has spent a significant amount of time. However, such a clustering technique may fail to generate key moments comprising of visually diverse images in a user's day in a scenario where the lifelogger has spent a significant time in a single location. To our knowledge, no previous study has extended the notion of temporal-spatial clustering to consider visual aspects for lifelogging data abstraction.

In this work, we combine visual scene evidences with time and location information in order to identify the most representative lifelog images as key moments in order to summarise a user's day. This feature is similar to Facebook's "Year in Review", where the objective for the model was to identify key moments in a user's year, offering an automatic visual summary based on their uploaded content. The feature, however, relied upon the number of likes (i.e. an explicit user feedback which is not present in traditional lifelogs) and did not consider if the images either belong to the user or they were publicly available content on the web (e.g. quotes, cartoons etc). In the context of lifelogs, we aim to generate summaries in the form of key moments, without any input from the lifelogger. Such a technique, which does not rely upon the lifelogger's input and provides a set of good quality, as well as visually diverse images, has to our knowledge not been reported in the existing literature. In the context of relevant use-cases, the automatic generation of key moments would be useful to lifeloggers, researchers interested in lifelogger's daily life experiences, community councils interested in community biographies of a sample of population *etc*.

The research presented in this chapter attempts to address the following overall research question: How can we effectively structure lifelog images to generate key moments, i.e. a review of a lifelogger's day? This question can be further partitioned in various sub problems:

- **RQ9.1**: How can we reduce the amount of noise in lifelogging collections? Can we exploit an image's context for this purpose?
- **RQ9.2**: Can we effectively use contextual information obtained from lifelogs to

recommend images in order to summarize the daily moments of one's life?

- **RQ9.3**: Can we combine visual scene features with temporal-spatial information to improve the summarization of daily moments?

The rest of this Chapter is structured as follows: In Section 9.2 we summarize the existing literature related to structuring lifelogs. In Section 9.3 we provide a brief overview of the devices used to capture images and corresponding locations, followed by the various approaches used to cluster lifelogs. Our experimental procedures and crowdsourced evaluation are then detailed in the Section 9.4, followed by the results in Section 9.5. Finally, we conclude in Section 9.6.

9.2 Background

In the existing research, lifelogs are often structured into activities or *events* (Doherty and Smeaton, 2008; Ellis and Lee, 2006). These events merge various sources of sensed data together into meaningful and logical units. However, such structuring would require sophisticated techniques for reasoning and inferring of activities from the lifelogs, which can differ in granularity as well as type. Once the events have been identified it is necessary to generate meaningful semantic information, which would minimize exhaustive browsing effort or simply relying on temporal information. Anguera et al. (2008) and many other lifelogging researchers have shown the potential of using meta information obtained from the lifelogs to automatically generate annotations. Wang and Smeaton (2012) attempted to categorise lifelogs into *daily activities* (e.g. *"taking a phonecall"*, *"cooking"*) using a density-based approach. Similarly, Hauptmann et al. (2007) used 10,000 concepts to effectively search and retrieve lifelogs. It is believed that inclusion of additional sources of evidence like GPS locations would help to improve the semantic annotation process. In this context, Lazer (2009) used GPS sensors in cell phones to annotate the lifelogs using their respective location. Gurrin et al. (2013) used WiFi sensors to further identify fine-grained locations of events. Li et al. (2016b) attempted to identify re-occurring events, or "motifs", in lifelogs through time series analysis. Specifically the authors applied Minimum Description Length (MDL) principles at various wavelet scales in order to identify motifs on the popular "All I have seen" Microsoft SenseCam dataset (Jojic et al., 2010). We refer the user to the review by Gurrin et al. (2014) for a full survey of lifelogging techniques, trends and challenges.

In the context of the research presented in this chapter, we further discuss two studies that have used location information to structure lifelogs. The most recent study by Kikhia et al. (2014) presented an approach to organise lifelogs based on places and activities obtained from GPS data. The lifelogs were structured using a combination of density-based clustering algorithms and convex hull construction. The lifelogs were

collected using SenseCam[1], whereas the GPS data was obtained from the smartphone. The system was evaluated with 12 users, who collected both the lifelog and GPS data for a single day, and used a prototype developed for the purpose of answering a survey. However, this work does not consider the elimination of noisy & duplicate images before clustering and therefore does not effectively address the information overload problem. One of the earliest works presented by Doherty and Smeaton (2010) used lifelog images with geographic data, to examine how the visual and location information might be useful to recall events from the past. In this context, 18 participants were asked to passively capture data using the SenseCam and a GPS device, over the period of two weeks. The lifelog images were presented as snaps, i.e. pictures which were presented in a temporal order from the beginning of a day, selected by the user. The GPS data showed the user tracks (daily routes) on a real-time custom Microsoft Virtual Earth map[2], which could be filtered by days. Finally, the GPS coordinates of lifelog images were combined and visualised on a map. It was concluded that lifelog images helped in recalling past activities, whereas location data supported inferential processes.

We hypothesise that structuring lifelog images into activities based on context (temporal-spatial information) may not be sufficient enough to support efficient retrieval or visualization, because of the vast amount of mundane data produced by a wearable camera. We hypothesize that the context, in the form of temporal and spatial information combined with the visual appearance of images will more effectively produce image clusters, thus decreasing the information overload problem of lifelogging systems.

9.3 Proposed Approach

In this following we give an overview of process & devices used to capture the lifelogs (images and location information) in section 9.3.1, followed by the techniques used to eliminate noisy lifelog images (i.e. blurred, duplicate) in section 9.3.2, and finally the ranking techniques employed to effectively generate key moments in section 9.3.3. A summary of the contents discussed in this Section is further illustrated in Figure 9.1.

9.3.1 Lifelog Capturing Phase

Due to the highly personal nature of lifelogging, conducting comparative studies on large scale collections is a very challenging task (Gurrin et al., 2014). In comparison to other areas of information retrieval which can evaluate on large *crawled collections*, lifelogging collections must instead be built *manually* by individuals wearing devices

[1]http://research.microsoft.com/en-us/um/cambridge/projects/sensecam/ - last accessed on 18th July 2016.

[2]http://www.bing.com/maps/ - last accessed on 18th July 2016.

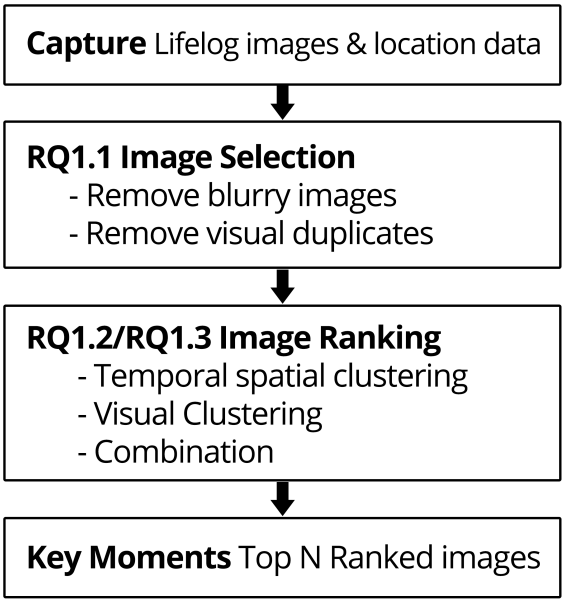

Fig. 9.1 Techniques used to structure lifelogs

for days or weeks on end. As a result, many works have conducted their experiments on small-scale collections (Gurrin et al., 2014). In recent years, however, some researchers have attempted to build larger collections with the view of creating datasets for use in comparative studies. For example, Bolaños and Radeva (2015) recently released their "Egocentric Dataset of the University of Barcelona" (EDUB) containing 4,912 daily images acquired by 4 users wearing the Narrative wearable device[3]. Most recently, Gurrin et al. (2016) proposed their collection used within the NCTIR lifelogging evaluation track[4]. This collection, which to our knowledge, is the largest of its kind containing 88k images taken by 3 users wearing the OMG Autographer device[5] over the course of a month. In addition to images, the collection has also been annotated with semantic details such as locations (e.g. work) as well as activity information (e.g. walking).

In our Computing Science School at the University of Glasgow, we are conducting a study, called the Glasgow Memory Server (GMS), which aims to bring together various different information streams specific to the city of Glasgow (e.g. social media posts, news articles *etc*)[6]. The most recent iteration of this study now considers lifelog data within the application. Therefore, within our experiments, we evaluate based on the first iteration of our lifelog collection, which is explained in the following paragraphs; additionally, this dataset has since been extended (Chowdhury et al., 2016).

The wearable sensors that are used to generate lifelogs depend on the variation of lifelogging that is performed. The array of sensors employed in lifelogging include:

[3]http://getnarrative.com/ - Last accessed on 19th July 2016.
[4]http://ntcir-lifelog.computing.dcu.ie/ - Last accessed on 19th July 2016
[5]http://www.autographer.com/ - Last accessed on 19th July 2016.
[6]This study is not currently publicly available.

(i) positional sensors such as GPS devices, Wi-Fi, Bluetooth (ii) activity tracking sensors like Fitbit "One", Fitbit "Zip" (iii) health sensors like heart rate, pulse rate monitor display watches (iv) wearable cameras/audio like Autographer, SenseCam, audio and video recorders. For the purpose of the research reported in this chapter, we have used two devices to collect the lifelogs as follows: (i) *Autographer*: a passive visual capture device, which can record more than 100 images per hour (ii) *GPS recorder*: recording location logs every 5 seconds. In the current scenario, location sensors in wearable cameras such as Autographer, take almost 5 to 6 minutes to capture a GPS fix, as opposed to 1 to 2 minutes claimed by the manufacturers. We believe, however, that this time lag will be resolved in future equipment. For this study, we therefore employ a separate, faster GPS recording device.

The lifelogs (i.e. images and corresponding GPS locations) for the research reported in this thesis were collected by a colleague within the department. The lifelogger carried both the devices for a period of 13 days, and transferred daily lifelogs from both the devices to the local hard drive of a stationary computer. The lifelogs were collected to form an exemplary collection, which would be used to generate key moments, using an array of approaches further discussed in this section. Our motivation is to develop a lifelogging visualization tool that can be used to support a number of real-time applications, which is further discussed in Section 9.6.

In this work the primary objective is to highlight the "key moments" in a user's day. From an information retrieval (IR) perspective, this can be formulated as *ranking* a subset of images from a large noisier collection (i.e. all images captured for an entire day), based upon some ranking score, whilst maintaining diversity in the top ranks.

9.3.2 Image Selection

One of the main problems in the automatic organization of lifelogs obtained from the wearable camera is managing noisy photographs, most predominantly: (i) blurry images (ii) visual duplicates. In the following subsections, we discuss the techniques employed for combating each problem, which also relates to the **RQ9.1** (as introduced in section 9.1).

Blurry Images As images captured by the lifeloggers are shot hands free and often whilst on the move, the image quality, in particular image sharpness, is often very low. In order to automatically identify blurry images, we adopt the technique proposed by Tong et al. (2004) due to its high reported accuracy (i.e. 98.6%) yet low computational complexity, which is important within the lifelogging domain where collections can rapidly grow over time due to the heads-free, automatic capture process. The method employs edge and sharpness analysis using a Haar wavelet transform. This technique computes a blur score B (where $0 \leqslant B \leqslant 1$) for a given image. Figure 9.2, highlights the blur distribution for all the images in our collection, where a lower score (i.e. blur

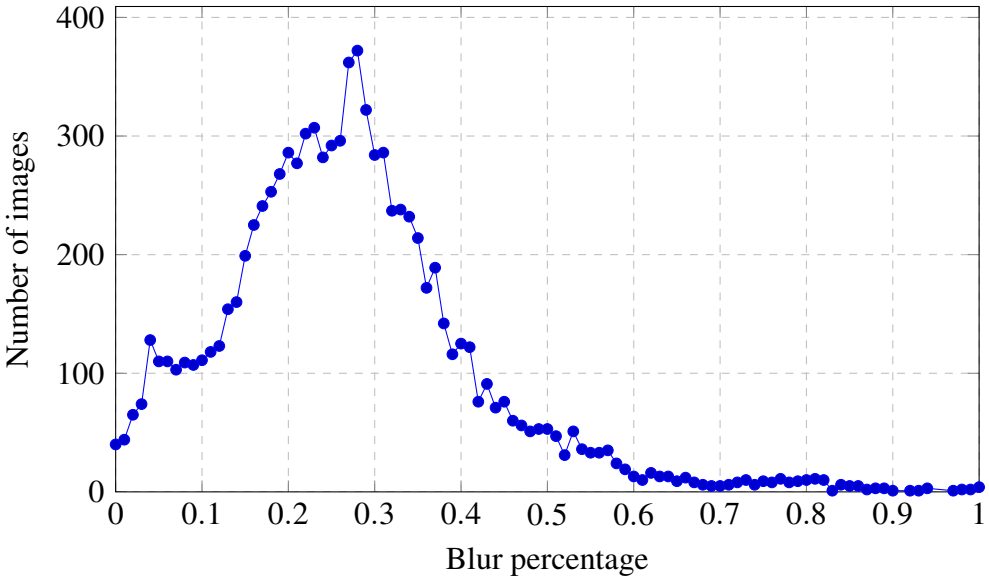

Fig. 9.2 Blur score distribution of the lifelog collection.

percentage) implies a sharper image.

For the purpose of the research reported in this chapter, we have selected only those images, where $B < 0.38$. This parameter is selected empirically, in order to remove the long tail of the distribution, which contains images which are blurrier than 80% of the overall population. We believe that blurred images will not be useful for providing high quality key moments in a user's day. However, such images (low quality) can be used for some human computational tasks like, elicitation based-surveys, where a social scientist may require all the lifelog images, irrespective of their quality. It is worth noting that all the lifelog images reported in this chapter are obtained from a single lifelogger. However, in the case of lifelogs obtained from multiple lifeloggers, the threshold value would most likely change for each lifelogger, because of the difference in use and the placement (round the neck, chest *etc*) of the wearable camera.

Duplicate Images As the lifeloggers may be stationary for longer periods of time (i.e. sitting at the office desk, travelling in a car/bus *etc*) and have limited control over as to when an image is captured, multiple image duplicates tend to exist within the lifelogs. These visual duplicates must be identified in order to avoid selecting duplicate images as key moments in a user's day, as well as eliminating duplicate information, which would further alleviate the information overload problem for our scenario. In our work, we detect the duplicate and near-duplicate images using the perceptual hashing technique (pHash) by Tang et al. (2012) due to its speed, low computational power and robustness in comparison to other similar image hashing functions (Zauner, 2010). The pHashing technique was shown to give high detection performance for resized, cropped and exposure compensated images; we hypothesise that detecting images which have these alterations will capture many of the duplicates taken by

lifelogging devices. We chose a hashing function to detect duplicate images in the lifelog collection due to its high performance, while maintaining low computational expense in extraction and matching between the images. By adopting the aforementioned method, we also ensure its scalability in large lifelogging collections.

To detect visual duplicates in our lifelog collection, we adopt the following procedure:

- **Step 1:** We first compute its pHash string (using the tools available at http://www.phash.org/).
- **Step 2:** Single pass clustering is then employed on all the images in our collection, using the hamming distance as a measure to compare between two pHash strings for a given pair of images.
- **Step 3:** Specifically, an image is added to an existing cluster if its hamming distance is small enough (T<8 as suggested by Chum et al. (2008)), otherwise the image is added to a new cluster.

Using the above method, duplicate images with the following alterations are grouped together: (i) brightness adjusted (ii) contrast adjusted (iii) gamma corrected (iv) 3x3 Gaussian low pass filtered (v) JPEG compressed (vi) watermark embedded (vii) resized and rotated slightly. For clusters containing multiple images (i.e. duplicates), we select only the sharpest image (i.e. smallest B).

9.3.3 Image Ranking

Once the blurred and duplicate images are removed, a second stage of ranking is carried out, in order to generate the key moments for a given day. The goal of image ranking in web search is to maximize the relevance of images in the top ranks with respect to a textual query. This differs from ranking lifelog images to generate automatic key moments, due to the absence of any query or similar user involvement, as well as textual annotations, which makes the ranking a non-trivial and ambiguous problem. As we are attempting to summarise a user's day, hence the focus is on maximising the visual diversity of images in the top ranks. We achieve the automatic generation of key moments by clustering images based on a number of visual and geospatial features, as follows.

Visual Clustering Our first approach attempts to cluster images based on their visual appearance in order to group images which have a similar *"visual scene"* (e.g. bedroom, office, street *etc*). We hypothesise that by selecting images from various scenes (i.e. different "moments") within a user's day we will improve the diversity, and therefore effectiveness, of the images selected in the top ranks. This scenario is similar to that of video shot boundary detection, where the goal is to split a continuous video into its constituent parts, for which the GIST feature (Oliva and Torralba, 2006)

has been shown to be very effective (Jhuang and Chikkerur, 2006). Furthermore, the GIST feature has also been effectively employed to classify between various similar, but different, traffic scenes (e.g. trafific, tunnel, highway *etc*) (Sikirić et al., 2013), an application which contains visually related images to those taken by a lifelogging device. Based on these motivations, we extract the GIST visual feature (Oliva and Torralba, 2006) from the lifelog images for visual clustering purposes. For each lifelog image taken within a single day, we execute the following approach:

- **Step 1:** Firstly, we only consider the images which have passed our selection process, i.e. all blurred and duplicate images are removed.
- **Step 2:** Secondly, we extract the normalised 512-D GIST feature for each of the images obtained in step 1.
- **Step 3:** Finally, we cluster using the Expectation-maximization (EM) algorithm. We employ this method over other popular clustering techniques (such as K-means), as the EM approach does not require an initial number of clusters to be set (i.e. K). This is important as we do not know the prior number of clusters or "moments" for any given day.

Temporal-spatial Clustering Time and location are crucial evidences for the purposes of segmenting images into various clusters or moments; images taken at a different time and place in a user's lifelog are likely to signify some change in activity or event (e.g. making breakfast at home *vs* going to the cinema at night). In this work, we therefore propose to cluster images based on both location (i.e. GPS co-ordinates) *and* time (i.e. image timestamp). In order to achieve this, we propose the following methodology:

Firstly, we model an image's time as the minute of the day in which it was captured normalised by the number of minutes in a day (e.g. 2:40pm is modelled as $880/3600 = 0.244$), referred to as T (where $0 \leqslant T \leqslant 1$). For time, we are able to extract this information from the image's timestamp; unfortunately, due to the time lag problem of capturing location information using the Autographer device (as described in Section 9.3.1) we have to capture GPS co-ordinates using a separate, faster, device. Specifically, we model an image's location using its *normalised* GPS co-ordinates; specifically we normalise the longitude (referred to as $L1$) and latitude (referred to as $L2$) values based on the maximum value recorded for each value, for each day. As the GPS equipment does not function well indoors, there are many images which do not contain GPS co-ordinates. In this case, we choose to set the longitude and latitude as the *mean* value for a given day, which we hypothesise in many cases would translate to some location between the user's home and work. We alternatively could have set these values to *zero*, the *minimum* or the *maximum* value for a given day, but this would have skewed the data set, thus degrading the clustering performance.

From this, we model an image as a 3D vector (i.e. $[T, L1, L2]$) of its time and location in order to cluster using the expectation-maximisation algorithm. This relates to **RQ9.2**, presented in Section 9.1.

Combining Visual and Temporal-spatial Clustering Finally, we combine the visual and temporal-spatial aspects by concatenating the two feature vectors, resulting in a 515-D vector (i.e. 512-D GIST feature + 3-D temporal-spatial feature), before clustering using the Expectation-maximisation approach. We hypothesise that visual appearance, time and location are all essential in the clustering process and will complement each other to summarise collections of lifelog images, by automatically generating key moments within a user's day. Such a combined clustering approach has not yet been studied to our knowledge in previous works and is beneficial in scenarios where location information is unavailable, either due to device constraints or other factors. This relates to **RQ9.3**, presented in Section 9.1.

9.4 Experiments

The overall goal of this work is to capture the key moments in a user's day, which we achieve by first removing noisy content before clustering images using visual and temporal-spatial techniques, as discussed in Sections 9.3. In this following, we discuss our lifelog collection in section 9.1, followed by various experimental systems in section 9.4.2 and finally our crowdsourced evaluation in section 9.4.3.

9.4.1 Data Collection

Table 9.1 shows the statistics related to our lifelog collection, which was captured by a colleague in the department of Computing Science at Glasgow University over a period of 13 days. The lifelogging devices (Autographer and GPS device) were used approximately from 8:30am in the morning to 6pm in the evening. However, the image capturing was stopped on a number of occasions, as required by the lifelogger due to personal or other reasons, which is not discussed further in this thesis. According to the statistics reported in Table 9.1, the number of images containing locations is only 14% because:

1. The GPS device did not log the location while the lifelogger went indoors, i.e. inside the university, supermarket, house and other buildings.

2. The timestamp in the GPS device was not synced with that of the wearable camera, due to a difference in sensor calibration. This could be done in future to increase the number of images containing the location information. A scaling factor might be also used to sync the timestamps.

Total Images	9,080
Number of days	13
Average # Images captured per day	698
% of blurred images	16.6%
% of images containing location information	14%

Table 9.1 Lifelog Image Collection Statistics

Despite there being a lack of location information gathered in our experiments, we believe that this will not be a problem in the future as the increased popularity of lifelogging will subsequently result in improved devices which will likely embed GPS sensors within. Hence there may not be a requirement to overcome the problem of syncing devices for the use cases identified in this thesis (mentioned in Section 9.1 of this chapter). However, in the current scenario, this does raise an essential limitation, i.e. temporal and spatial information cannot be solely relied upon to generate key moments. Finally, the lifelog data collected and evaluated for the research presented in this thesis is ethically approved through the lifelogger's informed consent, which is always a concern in this field, often making it difficult to evaluate the techniques.

9.4.2 Experimental Systems

In our experiment, we compare the top K images (where $K = 5$) for each day for each of the following five systems.

- **Random (S_{Random}):** Due to the lack of benchmarks in the field of lifelogging, especially in the context of generating the key moments of a user's day, we firstly propose ranking images randomly for each day, as a weak baseline. In this system (referred to as S_{Random}), we order images at random, using the MySQL $rand()$ function[7], and select the top K images for summarisation purposes.

- **Removing blurred images and duplicates (S_{Select}):** In this approach, we firstly attempt to remove the noisy images in the collection (i.e. blurred and visual duplicates) using the selection approaches presented in Section 9.3.2 before K random images are selected, as before. This system will be referred to as S_{Select}, and will be considered a stronger baseline compared to S_{Random} as we would expect more high quality images to appear in the top ranks due to the image filtering process.

- **Visual Clustering (S_{Visual}):** In this approach, we *visually* cluster images (using the EM clustering approach based on GIST features) which pass the filtering methods used in S_{Select}. We select the sharpest image (i.e. smallest B) from the K largest

[7]https://dev.mysql.com/doc/refman/5.0/en/mathematical-functions.html#function_rand - last accessed on 18th July 2016.

clusters as we hypothesise that the largest visual clusters will contain the most representative events in a user's day due to the length of time they spent observing a single scene; we use a round robin selection process in order to maximise the coverage of these events. In the case where less than K clusters exist, we select the 2nd sharpest image from each cluster and so on. This system is referred to as S_{Visual}.

- **Temporal-spatial clustering ($S_{Temp+Spatial}$):** In this approach, we cluster images (using the EM clustering approach based on the 3D temporal-spatial feature vector) which pass the filtering methods used in S_{Select}. We use the same image selection process as in S_{Visual} by selecting the sharpest images from the *largest* clusters as we believe these clusters will describe the major events in a user's day (or at least the events for which they spent the most time doing). We compare the effectiveness of clustering images based on time/location *vs* visual appearance in our evaluation. We refer to this system as $S_{Temp+Spatial}$.

- **Combined clustering ($S_{Combined}$):** Our final approach combines S_{Visual} and $S_{Temp+Spatial}$ to rank the images obtained after the selection process. Specifically, we concatenate the features output by these methods before clustering. Again we select the sharpest images from the largest K clusters. This system is referred to as $S_{Combined}$

9.4.3 Crowdsourced Evaluation

Benchmarking a system, which summarises a user's day based upon the lifelogs comprising of images captured by a wearable camera, is a non-trivial task due to the wealth of data collected even for a single day; for example, a device like Autographer can collect up to 200 images per hour, depending upon the settings used. However, such a setting was not used, as it drains the battery life within four hours. The device was set to capture 100 images per hour.

Most prior works in the field of lifelogging have evaluated their techniques with the lifelogger themselves, due to the context of the evaluation e.g. identifying events the user was involved in, annotating the key frames in the events, annotating the activities in the events *etc*. These evaluation techniques therefore relied on some amount of effort and explicit feedback from the user. In our work, we instead propose a crowdsourced evaluation which removes the lifelogger from the feedback loop and opens the door to much larger lifelogging experiments. As we are not evaluating events/activities which can be only identified by the lifelogger themselves, we can instead employ a taskforce of crowdsourced evaluators in order to judge the quality and diversity of the top ranked images for each experimental system presented in the previous section, thus increasing the speed and broadening the opinion of our evaluation. Moreover a crowdsourced evaluation will benefit use cases like generating community biographies for city councils and various forms of user experiences for social scientists, where the lifelog images are not necessarily used by the lifeloggers themselves but are instead submitted to external sources for a specific purpose, where quality and diversity are

essential dimensions.

Therefore, to evaluate this work we asked crowdsource evaluators to judge the quality and diversity of the images captured on a lifelogging device and ranked by our various approaches described in the previous section. We achieved this through two separate crowdsourced evaluations taken out on a popular crowdsourcing platform: CrowdFlower[8] (CF).

Evaluating Image Quality Our first crowdsourced evaluation attempts to judge the quality of the top K ranked images for each system. When the user accepts our HIT, they are presented with the following instructions:

You are presented with images taken from an individual's life logging device (a camera, worn around the user's neck, which takes 100 photographs per hour), as they go about their daily routine. In the following evaluation, you are asked 2 questions for 5 images taken on this device. The evaluators are asked the following questions with regards to a presented image:

- How clear is the photograph?
  Each evaluator is asked to rate on a Likert scale ranging from 0 (very blurry) to 5 (very sharp).

- How "interesting" is the photograph? Consider the scene and the objects, for example an image of a door or wall would be considered "uninteresting". An image of a street or depicting an activity (using a computer, eating etc.), would be considered "interesting".
  Each user is asked to rate on a Likert scale from 0 (very boring) to 5 (very interesting).

The first question aims to validate the blur detection part of our image selection process with the second attempting to gauge the noise reduction from an image interestingness perspective. Although image interesting is ill-defined and "in the eye of the beholder", we believe that by measuring perceived interest from a *wide spectrum* of crowdsourced evaluators, we will gain some insight into the visual attractiveness/engagement of the selected images.

As discussed in Chapter 5, one of the major problems with crowdsourcing is that evaluators often spam or try to complete the tasks with as little effort as possible, in order to maximize their profits (Vuurens et al., 2011) leading to poor quality evaluation. Test questions (i.e. those with known answers) are used to identify spamming evaluators. Specifically, we asked evaluators 4 test questions (i.e. 4 images), where the answers are clear with respect to the task description and questions. For these test questions, users are asked the two questions (as defined above) where the "correct" answers are pre-defined as a range on the Likert scale. For example, Figure 9.3 shows an example test image with the "correct" answers shown for demonstration purposes (the

[8]http://www.crowdflower.com/ - last accessed on 18th July 2016.

actual experiment hides the correct answers). In order to accept our HIT an evaluator must first answer all the test questions correctly within some reasonable range. By doing so, we ensure the highest submission quality and that evaluators have understood the task required of them. Further, CrowdFlower rates the trustworthiness of the evaluators based on their previous work; for the purpose of the evaluation reported in this thesis, we required that the evaluators must have at least a "Level 1 Contributor status" (i.e. high performance contributors who account for 60% of monthly judgments and maintain a high level of accuracy across a basket of CF jobs). Finally, as an additional measure to reduce spam, we used an even-point likert scale (i.e. 6 points) in all of our questions, thus removing the "middle option". This is a commonly used method which "forces" the user to make a decision (Allen and Seaman, 2007) rather than allowing them to "blindly" select the neutral point for each question. By employing this method, we hoped to increase user engagement and reduce the number of spam responses.

In our experiment, each evaluator was presented with five images from those selected by various systems detailed in Section 9.4.2. Each image was evaluated by three separate evaluators with the survey scores averaged. Each evaluator was paid $0.03 on completion of this task. For our test questions, 60 evaluators (out of 224) answered all questions correctly progressing to our experiment. For our two questions (both with six different options), we achieved a *Crowdflower confidence score*[9] of 0.66 and 0.56. The fairly low confidence scores achieved is mainly due to the high number of options available for each question (i.e. six) and the difficultly in gauging how relevant an image is for summarization purposes in the context of lifelogging. In order to alleviate this problem, we collect the opinion of three separate crowdsourced users and average their various scores.

Evaluating Diversity In order to evaluate the visual diversity of top ranked images, we conducted a second crowdsourced evaluation, where evaluators are asked to judge the "visual similarity" of image pairs as taken from the rankings of each of the systems. Users are presented with the following task description on acceptance: You will be presented with 10 pairs of images; your task is to judge how "visually similar" the two images are for each case. The user is asked the following question with regards to a presented image pair:

- How visually similar are these images?
  Each evaluator is asked to rate on a Likert scale ranging from 0 (completely different) to 5 (Identical).

As before, evaluators had to pass four test questions for the given task where the "correct" answers were pre-defined and images selected based on their clarity (i.e.

[9]This confidence score ranges between 0 (no agreement) and 1 (complete agreement) - for more information, see: https://goo.gl/xxnPwy - last accessed on 18th July 2016.

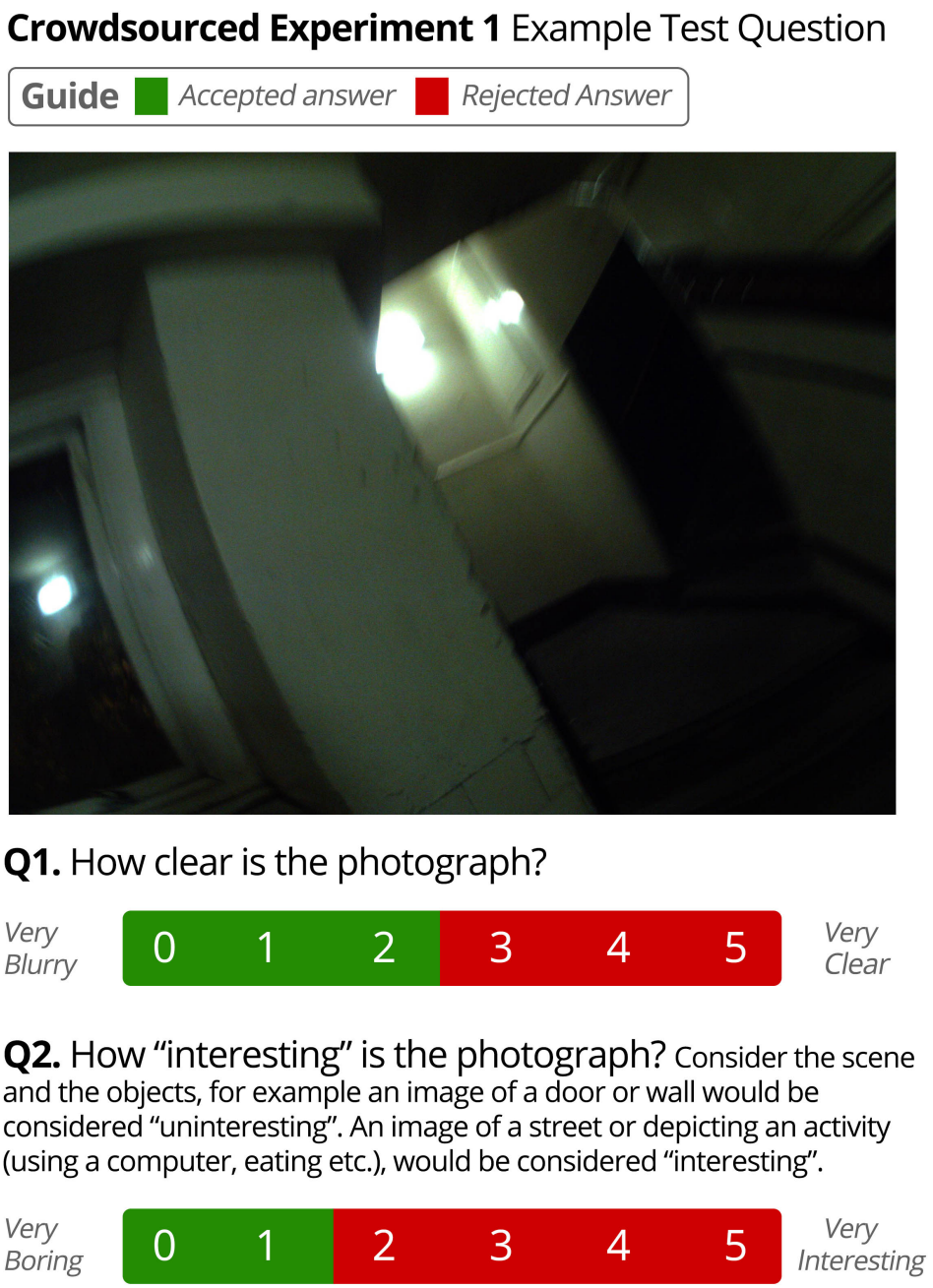

Fig. 9.3 An example test question for judging image quality where the "correct" answers are shown for demonstration purposes. The correct answers are not shown to the user in the actual experiment.

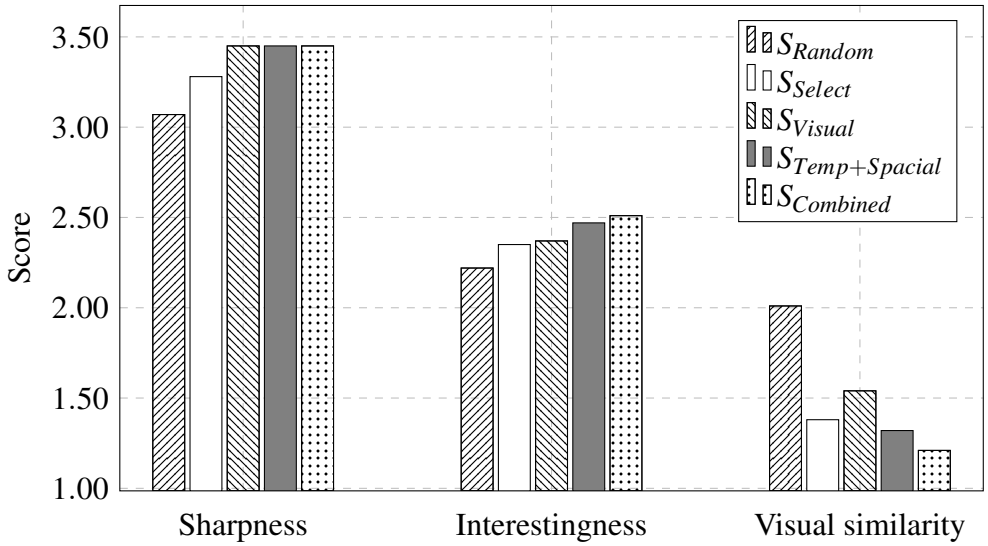

Fig. 9.4 Graphical results from crowdsourced evaluation. Sharpness (0=V Blurry…5=V Sharp), Interestingness (0=V Boring…5=V Interesting), Visual Similarity between top 5 (0=V Different…5=V Similar).

image pair were either clear duplicates, or completely different). Further, only CrowdFlower evaluators with a Level 1 contributor rating can accept our HIT. Further, the user must first answer four test questions correctly, where the answers are clear with respect to the task description, before they are allowed to take part in our experiment. In our experiment, users were paid $0.04 for judging 10 image pairs with each pair judged by three different users. 72 users (out of 201) answered all the test questions correctly progressing to take part in our experiment, for which they contributed 1,612 judgements in total. For our question judging the visual similarity (having six different options), we achieved a Crowdflower confidence score of 0.7.

9.5 Results

Based on the judgements made in our two crowdsourced evaluations, we are able to quantify the sharpness, interestingness and visual similarity in the top 5 ranked images for each system defined in Section 9.4.2. Table 9.2 shows the average statistics obtained from both the evaluations, and clearly demonstrates the effectiveness of the $S_{Combined}$ system (i.e. combining temporal, spatial and visual clustering); Further, Figure 9.4 shows this gain graphically.

> **RQ9.1:** How can we reduce the amount of noise present in lifelogging collections? Can we exploit an image's context for this purpose?

Firstly, considering the effectiveness of the selection methods, we observe from the scores (obtained from computing image blur using the HAAR wavelet method and

System	Sharpness	Interestingness	Visual similarity
S_{Random}	3.07	2.22	2.01
S_{Select}	3.28	2.35	1.38*
S_{Visual}	3.45**	2.37	1.54
$S_{Temp+Spatial}$	3.45***	2.47*	1.32*
$S_{Combined}$	3.45***	2.51*	1.21*

Table 9.2 Tabular results & statistical significance tests. T-tests against our S_{Random} baseline are denoted as * being p $<$ 0.05, ** being p $<$ 0.01 and *** being p<0.001. Sharpness (0=V blurry...5=V sharp), Interestingness (0=V boring...5=V interesting), Visual Similarity between top 5 (0=V different...5=V similar).

using the pHash method of duplicate detection) achieved by S_{Select} that two major problems faced by lifelogging devices (i.e. blurred and duplicate images) can be automatically alleviated resulting in a 7% increase to image sharpness and 31% decrease to visual duplicates in the top 5 ranks, compared to our S_{Random} baseline.

> **RQ9.2:** Can we effectively use contextual information obtained from lifelogs to recommend images in order to summarize the daily moments of one's life?

> **RQ9.3:** Can we combine visual scene features with temporal-spatial information to improve the summarization of daily moments?

In terms of effectiveness of the ranking/clustering methods, the statistics obtained from our evaluation demonstrates that the images which are clustered based upon the combination of all three aspects proposed within this chapter (i.e. visual appearance, temporal and spatial) received the highest average score (Table 9.2, last row) for each of the three aforementioned features thus addressing both **RQ9.2** and **RQ9.3**. The $S_{Combined}$ also achieved a 21% increase (on average) over the S_{Random} baseline for all the three features. The reduction of noise in the form of blurry images and visual duplicates significantly improves image interestingness implying a positive correlation between the two.

By improving the quality (i.e. sharpness and interestingness) and visual diversity of images taken by lifeloggers, we would expect to improve the user's search and retrieval experience when reviewing their images for a day. Firstly, we would expect our filtering approach to significantly reduce browsing time as almost 60% of images in our collection are considered noise (i.e. too blurry or a visual duplication). Secondly, by promoting visual diversity on this subset using our proposed clustering techniques, we would expect a user to more effectively review their lifelogging data by presenting them with 5 visually diverse images (out of possibly thousands of images).

9.6 Chapter Summary

This work extends the notion of structuring lifelog data by evaluating a number of techniques to automatically generate good quality and visually diverse images in order to form key moments, summarising a lifelogger's daily life. Two crowdsourced evaluations demonstrated that the most effective technique to generate such key moments would rely upon firstly eliminating (i) *blurred images*: using a Haar based edge and sharpness analysis technique (Tong et al., 2004), as well as (ii) *visually duplicate images*: using a pHash blur detection approach (Tang et al., 2012), thus addressing RQ9.1. In order to effectively rank images, we determined that clustering based on a *combination* of contextual (e.g. geo-temporal) as well as visual features (e.g. the GIST feature (Oliva and Torralba, 2006)) to be most effective for our purposes, thus addressing our research questions RQ9.2 and RQ9.3. As demonstrated in our evaluation, by doing so we were able to improve the sharpness, perceived interestingness and visual diversity of images in the top ranks. The techniques reported in this chapter also contribute to decreasing the information overload problem, one of the major problems posed by lifelogging devices, by eliminating blurred and visually duplicate images. Also, we acknowledge that automatic key moments generated through the technique reported in this chapter can be used with the existing event and activity detection techniques, to further provide meaningful insights in the context of lifelogging to address other use cases not considered in this chapter e.g. daily user activities – time spent walking, in front of computer, eating *etc*.

One of the drawbacks of this work, however, is that we were only able to evaluate on a small collection of images due to various issues for experiment subjects e.g. privacy, time & effort *etc*. Therefore, the major challenge for continued research in this area concerns open access to consented consumer lifelog data. We hypothesise that in order to overcome this challenge, researchers must first address the privacy concerns of lifeloggers and the general public, using techniques such as the automatic blurring of faces *etc*. From a quantity perspective, however, we believe that this problem will naturally alleviate itself with the rise in popularity of social media and online video broadcasting services such as Periscope[10]. Overall, it is an exciting time for lifelogging where the number of possible applications grows year on year, from surveillance & security, to healthcare applications for dementia suffers and beyond. Due to the breadth in applications, future research must focus on addressing each use-case individually, however, instead of proposing generic methods for the organisation of images, videos *etc* - for example, in our case, low quality/duplicate images are useless for our application and should be removed, however, they may prove to be crucial pieces of evidence within the *surveillance domain* and therefore left intact.

[10]https://www.periscope.tv/ - last accessed on 18th July 2016.

Part IV

Conclusion

In the final part of this thesis we summarise the various works presented in the previous chapters on photo tag recommendation and photo recommendation. In particular we focus on the high level "take away" messages which can be drawn from this work as well as dwell on new areas for future research. Finally, we also summarise the various research papers completed within this PhD which have subsequently be presented at various conferences.

Chapter 10

Conclusions

10.1 Introduction

This thesis investigated the role of *context* in the annotation and retrieval aspects of image search proposing that an image's pixels do not reveal the true extent of an image's meaning. For example, it is (nearly) impossible to identify an image's location or event based purely on its visual appearance. Therefore, building image retrieval systems indexed on (often incorrect) *high level* visual automatic annotations (e.g. `fish`) will only allow for the retrieval of simple objects or concepts. In order to allow for more complex queries (e.g. `T in the park 2014`) images must be annotated by people, however, users are not willing to spend the time and effort annotating their pictures and therefore rely on semi-automatic photo tag recommendation techniques in order to ease this process. In this area, existing works have considered only the relationships between tags and thus have ignored the context of an image, often resulting in irrelevant tag suggestions, especially for ambiguous queries (e.g. `NY`). In this thesis we therefore defined new evidences, which attempted to alleviate this ambiguity in a number of tag recommendation and retrieval tasks. In particular we aimed to address the following high-level research questions:

HL-RQ1. Can the *context* an image is taken in be exploited for photo tag recommendation purposes in order to complement existing textual evidences? Which contexts are most effective for this task?

HL-RQ2. Can the *context* an image is taken in be exploited for image recommendation and retrieval purposes? How can this context be used to alleviate the problems associated with retrieval on un-annotated images?

Firstly, Part II considered the *annotation* of images. The semantic gap has become somewhat of an irremovable constant from multimedia retrieval tasks despite over two decades of intense research (Datta et al., 2008; Hanbury, 2008; Smeulders et al., 2000); it was therefore our aim to focus on different directions within this domain. In particular, we identified 17 new features which either attempted to overcome, minimise, or

avoid, the effects of the semantic gap in order to gain some meaningful representation of an image. Using these features, we focused on the semi-automatic image annotation task of *photo tag recommendation* which, by exploiting a small amount of textual evidence (i.e. user annotations), we were able to achieve accuracy beyond that of the "state-of-the-art".

In Part III, we continued by exploiting the context of an image in three new image retrieval scenarios. The first work focused on retrieving pictures from social media, by exploiting an image's social context, which are most useful for event summarisation purposes. In the second we attempted to identify whether an image would become popular in the future based on various contextual features. Finally, we considered the new area of lifelogging by attempting to filter and cluster images, based on the visual appearance and context, from a user's lifelog in order to best summarise their day. In the following chapter, we conclude and summarise the high level research outcomes from this thesis starting with photo annotation in section 10.2 and photo retrieval in 10.3.

10.2 Photo Annotation

The majority of this thesis focused on the annotation of images. We proposed that, as an image's visual appearance does not reveal its entire meaning, and due to the often ineffective performance and high computational complexity of automatic image annotation techniques, the user must be employed in the annotation process. In this part, we therefore focused on semi-automatic photo tag recommendation techniques. In the following, we summarise the contributions of our many works within this task.

Internal Evidences for PTR Our first work attempted to identify new aspects surrounding an image which could be used to give extra clues as to its semantics (i.e. the user annotations). Firstly, we considered new aspects of an image's context: (i) its time taken (ii) location (iii) shooting device and settings (iv) and its social context (e.g. number of views/comments). Secondly, we considered new visual aspects for PTR, which attempted to build upon *high precision* techniques in (i) scene classification (ii) and face detection; each of these aspects attempted to classify images for only a few (<4) categories (e.g. indoor or outdoor, no face or one face *etc*) using state-of-the-art techniques, resulting in reliable visual evidences for photo tag recommendation purposes. Finally, we considered simple demographics surrounding the user whom uploaded a given image, such as (i) their gender, (ii) account "type" (iii) and social context (e.g. number of images/contacts).

Based on this work, we identified a number of conclusions:

- Firstly, we identified, through extensive analysis of 1M annotated images, that tags exhibit a number of *trends* and *tendencies*. For example, many tags, such as `party`, show strong daily, weekly and yearly temporal cycles whilst other tags, such as

`sunset`, demonstrate strong image orientation biases (e.g. where an image is far more likely to be annotated with `sunset` if taken in landscape, rather than portrait, mode). In fact, as is observed in Table 3.7, almost all of our contextual and visual classifications observed strong tagging correlations.

- Secondly, by building tag co-occurrence matrices which considered these evidences we were able to incorporate (and exploit) these trends and tendencies in the photo tag recommendation process, improving accuracy by up to 50% (for P@5) in comparison to our TF-IDF baseline, addressing our initial high level research question, **HL-RQ1**. In particular, we concluded that temporal, orientation, high level scene/colour and a user's online presence to be the most effective features for photo tag recommendation purposes.

- Thirdly, we explored two feature combination strategies for our purposes, achieving a 75% (for P@5) increase in recommendation accuracy over our TF-IDF baseline. From this, we concluded that evidences contain latent relationships which can also be exploited for PTR purposes. For example, consider two images annotated with the tag `sharks` (i) image #1, which contains 0 faces, is taken outdoors, on a summer afternoon in Australia (ii) image #2, contains 3+ faces, is taken indoors, on a winter evening in America. Which image would you most likely expect to be of the *animal*, or the *ice hockey team* (i.e. San Jose Sharks)? Obviously we would hypothesise that image #2 was taken at a San Jose Sharks ice hockey match. In isolation, each evidence may not provide enough discriminatory value to answer this question, however, when combined we are able to more accurately define the *context* of the image.

- Finally, we considered the application of the discussed features in a cold start recommendation scenario (i.e. no user tags, equivalent to an automatic image annotation setting). By exploiting all of the proposed features in combination, recommendation accuracy was improved by almost 50% in comparison to our weak baseline. Also, we further identified that the discussed features are best employed when the input tags are of *medium* or *low* frequency (i.e. unpopular tags).

External Evidences for PTR After some investigation, we identified a crucial "time lag" problem present on image sharing websites. Particularly, images are often uploaded long after they are taken. This causes problems for recommendation models as the image database, and therefore tag co-occurrence matrix, is often "out of date". For new events, recommendation models will not be able to make relevant suggestions due to the lack of "training data" caused by the delay between users *taking* and *uploading* photographs. To combat this problem, we focused on exploiting up to date social media streams and Wikipedia data for tag recommendation purposes. Based on this work, we identified a number of conclusions:

- Firstly we quantified the significance of this "time lag" problem by identifying that images taken at a major American music festival were uploaded, on average, 50 days after they were taken. Further, we identified that significantly more tweets were posted about the festival than Flickr photos: 2,750 Flickr images vs ~150,000 tweets (considering access to the Twitter "firehose" is available). A speed difference of many magnitude was also observed, where *no* ACL 2012 images were *uploaded* to Flickr during the first day of the event, whereas >10,000 tweets were posted.
- Secondly we were able to achieve recommendation accuracy comparable with a state-of-the-art model by making suggestions based on a combination of Twitter and Wikipedia content. We were able to achieve high quality recommendations by firstly reducing the noise present using natural language processing techniques, before selecting only popular nouns and entities from related Twitter streams and relevant Wikipedia articles, addressing our initial high level research question, **HL-RQ1**.

PTR Diversification Despite the improvement in tag recommendation accuracy achieved in Chapters 3 & 4, we observed that many of the suggestions were either synonyms of tags already assigned to the image, or synonyms of each other. To address the later, we attempted to diversify the tags suggested in the top ranks. This was achieved using diversification techniques from traditional information retrieval i.e. Maximal Marginal relevance (MMR). Based on this work, a number of observations were made:

- Firstly we addressed problems with the *evaluation* of existing photo tag recommendation approaches. Specifically we identified that images are often tagged with synonyms, therefore, a model which suggests multiple synonyms in the recommendation list will achieve higher evaluation scores (using relevance based metrics) despite the actual utility, or merit, of these suggestions being far lower than a model which promotes a diverse list of concepts.
- Based on this analysis, we re-ranked suggestions from a state-of-the-art recommendation model using the MMR notion of diversification, improving for intent aware metrics by up to 6.7% (for αn=DCG@5). Further, we identified optimal parameters for this model based on our experiments, which suggested an aggressive tag diversification strategy (i.e. $\lambda = 0.4$) based on the top 20 ranked tags.

AIA and PTR Evaluation Based on the analysis made in the previous chapter (i.e. that synonymous tag sets were used as ground truth in existing photo tag recommendation evaluations) we were interested in any other possible biases that may have been present in collections used to evaluate image annotation methods. To this effect, we studied *three* popular automatic image annotation and *two* photo tag recommendation testbeds. From this work, a number of conclusions were made:

- Firstly, we identified seven crucial flaws in these three existing AIA testbeds. Specifically, we concluded that due to various problems, such as synonymous ground truth

and testing on skewed testsets, misleading evaluation scores could be achieved. We identified in one collection (IAPR TC-12) that due to these problems, a model may "underperform" by up to 15%. Due to this, we built a new AIA testbed containing 312k images, called FLICKR-AIA, which attempted to reduce this bias as well as address six other identified issues.

- Secondly, we identified three major issues in two PTR testbeds. Aside from the discussed problems of tag synonymity, the main problem with these collections was that they were never publicly released, making any comparative studies impossible. To address the problem of tag synonymity, we conducted a crowdsourced experiment which grouped together the synonyms of 1,000 Flickr test images, allowing for the computation of intent aware diversification metrics. The resulting testset and training collection containing 2M images, called FLICKR-PTR, was created and publicly released to address the various identified issues.

10.3 Photo Retrieval

In the second part of this thesis, we focused on the retrieval of images for various new multimedia search tasks. In particular we focused on how *context* could benefit image retrieval and recommendation models, especially in a cold start. Specifically, we proposed three new scenarios, namely: (i) *visual event summarisation*, which covers the larger task of image search on social media streams where we attempted to retrieve the most relevant & diverse images for news events which had started to trend on Twitter (ii) *photo popularity prediction*, which covers the larger task of photo recommendation on image sharing websites where we attempted to identify images which would be popular in the future (iii) lifelog summarisation, which covers the larger task of retrieval within lifelogging collections where we attempted to retrieve the 5 most relevant and interesting images for a user's day. In the following sections we summarise the contributions of these works.

Visually Summarising Social Media Events With the huge amount of content shared today by users on social media streams, a number of works have attempted to exploit these resources for automatic news story detection purposes. Once microblog posts have been grouped into "candidate events", a phase of summarisation is taken out. How one summarises an event is still an open information retrieval research question, however. Elementary works in this field have focused on generating a *title* which encapsulates the topic of the story. We argued that due to the complex nature and informal language used on social media streams, attempting to do this automatically often resulted in substandard, or erroneous, titles. To this effect, we proposed to instead retrieve *images* from tweets which effectively summarised a given new story. Based on this work, we were able to make a number of conclusions and observations:

- Firstly, we identified a range of problems with using images posted to social media streams. These included: (i) *sparsity*, where only 4% of tweets contain an image (ii) *near-duplicate images*, where images which were taken of the same visual scene, except at a slightly different angle/time/camera or were cropped/watermarked (iii) *irrelevant images*, which could be categorised as memes, screenshots & reaction images and (iv) *spam/advertising*, where bots post image advertisements disguised as relevant tweets (by "piggybacking" onto trending Twitter topics).

- To address these issues we firstly filtered out noisy photographs using visual classification techniques. To address sparsity, we also collected additional images from websites referenced within tweets. To rank images from these sources, we firstly clustered tweets describing a single event into sub-topics and selected the most popular image from each cluster. In our most effective system, over half of the images in the top 5 ranks were deemed relevant in an extensive crowdsourced evaluation for its given event, addressing our initial high level research question **HL-RQ2**.

- In our crowdsourced experiment, we were also able to quantify the visual *quality* of the retrieved images. From this we observed that images collected from related websites (referenced within the tweets) were of higher quality than those collected directly from tweets.

- Finally, we studied the *suitability* of images for summarising each event type. Based on this analysis, we concluded that events which happen "behind closed doors" (e.g. Law/Politics) are least appropriate for this purpose, and that those which happen in the public domain (e.g. sports) are most suitable.

Image Popularity Prediction Due to the increase in uploaded content online, much research has also focused on identifying those items which are of high quality, or are likely to become popular in the future. In this work, we highlighted that the *minority* of images, receive the *majority* of public attention. We therefore proposed a new area of *photo popularity prediction* which attempted to identify these images (i.e. those which are likely to receive high levels of comments/views in the future) allowing image sharing websites to recommend, rank and organise their content more effectively. Due to the lack of annotations present in images, however, we focused on predicting popularity based on their *context* by exploiting the features presented in Chapter 3. Based on our work, a number of conclusion were made:

- In order to predict popularity, we firstly classified images based on 15 different *evidences* introduced in this thesis. Considering a combination of these classifications, we were able to categorise images as popular/unpopular (using a support vector machine) with 74% accuracy (for comments) which was in-line with our oracle approach, which classified based on 2 or 3 textual annotations, addressing our initial high level research question **HL-RQ2**.

- The most effective features for our purpose exploited the *user's context*, however, with poor accuracy achieved when considering only an image's visual appearance. This highlighted that an image's appearance is ineffective for popularity prediction purposes as popular images usually differ, in some way, from the majority.

- Finally, based on further analysis, we were able to identify a number of popularity trends, such as the number of faces in an image strongly correlates with number of views/comments it receives; particularly, images containing many faces are more likely to viewed but less likely to be commented on implying a different type of user interaction. We also highlighted that nature photographs are rarely viewed, but when they are, they generate a high number of comments. This further highlights our initial motivation to identify those high quality images which are likely to evoke much discussion, within a much larger set of mostly uninteresting content, for image recommendation purposes.

Lifelog Summarisation Aside from the increase in size of online image collections, the rise in popularity of lifelogging devices has resulted in individual users having *thousands* of images detailing every aspect of their life. In this work, we attempted to reduce this information overload by summarising their day into a few relevant images. Specifically, we attempted to rank the most relevant and interesting images based on their context and visual appearance. From the experiments taken out in this area a number of conclusions were made:

- Firstly, removing blurry images and identifying visual duplicates using a hashing technique we were able to increase image sharpness and reduce the number of duplicate images returned in the top ranks.

- Secondly, by clustering images based on their context (i.e. the time & location they were taken at), as well as their visual appearance, we were able to offer (i) sharper (ii) more visually diverse (iii) and more interesting images to the user thus addressing **HL-RQ1**.

Chapter 11

Future Work and Directions

11.1 Introduction

The purpose of undertaking a Doctor of Philosophy (Ph.D.) is to extend human knowledge regarding a very specific topic or problem (i.e. a research area) whilst a further success criteria for a Ph.D. should be its "opening of new avenues" for subsequent students. To this end, as of the 14th of July 2016, our 10 publications have been cited by authors on 27 occasions. In particular a number of *survey style* publications have included our works in their listings, such as those by: Li et al. (2016c), Li et al. (2016a), Levar and Castro (2013) & Andreadou et al. (2016). Most importantly, however, our work has been extended and benchmarked against by other work (Schinas et al., 2015, 2016). In the following I detail possible future research avenues which I did not have time to explore within the four years which confine a Ph.D, starting with those related to photo annotation in section 11.2 and photo retrieval in section 11.3.

11.2 Photo Annotation

Based on the research taken out over the last four years, most of the work focused on the semi-automatic annotation of images. The following sections propose future directions within our five major annotation works.

Internal Evidences for PTR Our work on internal evidences for photo tag recommendation opens up a range of possible research focuses, such as: (i) Where can new evidences exist for photo tag recommendation purposes? Could User interaction data (e.g. Flickr groups, favourited/shared images *etc*) or additional camera EXIF information (e.g. focal length, saturation, sharpness, contrast *etc*) be exploited for this purpose? (ii) How can we most effectively combine features for PTR purposes? Can we exploit topic modelling techniques, such as Latent Dirichlet allocation (LDA), to more effectively identify latent relationships between features? How do we most effectively weight these features? (iii) How do we personalise the proposed features for PTR pur-

poses? For example, an image containing many faces taken at night might imply a rock concert photograph for one user and an ice hockey match for another. Could we model this as a trade-off between *personal* and *global* co-occurrence matrices? (iv) Could pseudo-relevance feedback techniques, from traditional information retrieval, based on visually/semantically similar images be useful for our tagging purpose?

External Evidences for PTR Our work on external evidences for PTR was the first this area thus promoting many future directions, such as: (i) Can other textual sources be exploited for PTR purposes? For example could additional evidences be drawn from user comments, news/blog content or other social media sources (e.g. Foursquare, Instagram *etc*)? (ii) How can external knowledge databases and the relationships between entities be employed for our purpose (e.g. DBPedia (Lehmann et al., 2014))? (iii) As tags are inherently sparse, could these external textual resources be employed to reduce this sparsity (e.g. co-occurrence of terms on Wikipedia)?

PTR Diversification Our work on tag diversification was only ever submitted as a short paper, thus restricting the amount of analysis and experiments which could be taken out. Therefore, we propose a number of future directions: (i) In our work we only explored an *implicit* model of diversification (i.e. one which assumes similar documents, or tags, cover similar aspects). Recently, more elaborate *explicit* models of diversification have been proposed which attempt to decompose the various sub-topics using external evidences (Agrawal et al., 2009; Santos et al., 2011). Can these explicit diversification models further improve sub-topic coverage for photo tag recommendation purposes? (ii) How do we rank the importance of various sub-topics? For example, are *location* sub-topics more "useful" for annotation purposes than *object* sub-topics? (iii) Instead of employing diversification models, can we automatically infer synonym tag sets from historical Flickr images? By doing so we can recommend no more than one tag from each "cluster"?

AIA and PTR Evaluation One of the objectives of this thesis was to build fair and openly available testbeds for PTR and AIA purposes. In this line, we propose a number of future directions: (i) Due to the extensive computational expense and variations in code/parameters, state-of-the-art image features should also be included in collections to ensure comparable experiments. (ii) Additionally, a number of high level image classifications should also be included within these collections in order to promote research focuses in this line. (iii) Photograph timestamps are often wrong due to incorrect configurations on user cameras. Could these timestamps be automatically corrected based on the timezone, image context & content *etc*?

11.3 Photo Retrieval

In this thesis we also focused on new paradigms for image retrieval which exploited an image's context in a cold start scenario. For these new areas of research, we propose a number of future directions:

Visually Summarising Social Media Events Firstly, for our work on visual event summarisation we propose the following research questions: (i) How can other image evidences be exploited for image ranking & selection purposes? Can filenames or explicit & implicit user feedback (e.g. image "likes") be exploited for this purpose? (ii) Can images be extracted from other social media streams (e.g. Flickr, Instagram *etc*) in order to create visual event summaries? How does this content compare, in terms of speed and quality, with those collected from tweets and related websites? (iii) Can other media types be employed for summarisation purposes (e.g. videos, maps, audio *etc*)? (iv) How does one most effectively present images in a more elaborate interface? How many images is optimal for summarisation purposes and is this parameter event dependant?

Image Popularity Prediction In our second photo retrieval work we proposed the paradigm of image popularity prediction where we attempted to predict whether an image would gain much attention (i.e. comments & views) in the future based on various contextual and context-based features. Based on this work, we propose the following new research questions for future work: (i) Can "rules" from painting, art & design (e.g. composition, the "golden ratio", symmetry, patterns *etc*) be employed for predicting an image's popularity? (ii) Can other, more elaborate visual analysis techniques (e.g. colour distribution analysis, contrast analysis *etc*) also be employed to more accurately predict a potentially "popular" image?

Lifelog Summarisation In our final chapter we attempted to summarise a user's day from their lifelogging images. For this work, we propose the following future directions: (i) Firstly, as we only evaluated our approach on a small collection, we plan to extend this work by building a much larger test collection. In particular, we are collating a lifelogging collection containing 100 people using the device on a daily basis for a period of 12 months. (ii) Aside from time, location and high level scene, what other features can be used in order to effectively cluster images into different "moments"? Could we use face detection techniques to identify different people in the collection in order to better infer clusters of "events"? (iii) What makes a lifelogging image interesting and which features could capture this dimension? Could we build on techniques from works attempted to predict an image's aesthetic value (Jiang et al., 2010)?

References

(2007). *TVS '07: Proceedings of the International Workshop on TRECVID Video Summarization*, New York, NY, USA. ACM. 433077.

Aggarwal, C. C. and Subbian, K. (2012). Event detection in social streams. In *Proc. of SDM Conference*.

Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA. ACM.

Ahn, J.-w., Brusilovsky, P., Grady, J., He, D., and Syn, S. Y. (2007). Open user profiles for adaptive news systems: help or harm? In *Proceedings of the 16th international conference on World Wide Web*, pages 11–20. ACM.

Aizawa, K., Tancharoen, D., Kawasaki, S., and Yamasaki, T. (2004). Efficient retrieval of life log based on context and content. In *Proceedings of the the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, CARPE'04, pages 22–31, New York, NY, USA. ACM.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y., Umass, J. A., Cmu, B. A., Cmu, D. B., Cmu, A. B., Cmu, R. B., Dragon, I. C., Darpa, G. D., Cmu, A. H., Cmu, J. L., Umass, V. L., Cmu, X. L., Dragon, S. L., Dragon, P. V. M., Umass, R. P., Cmu, T. P., Umass, J. P., and Umass, M. S. (1998). Topic detection and tracking pilot study final report. In *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218.

Allen, I. E. and Seaman, C. A. (2007). Likert scales and data analyses. *Quality progress*, 40(7):64.

Alonso, O., Strötgen, J., Baeza-Yates, R. A., and Gertz, M. (2011). Temporal information retrieval: Challenges and opportunities. *TWAW*, 11:1–8.

Andreadou, K., Papadopoulos, S., Zampoglou, M., Andreadou, K., Papadopoulos, S., and Zampoglou, M. (2016). of deliverable: Multimedia linking and mining.

Anguera, X., Xu, J., and Oliver, N. (2008). Multimodal photo annotation and retrieval on a mobile phone. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, MIR '08, pages 188–194, New York, NY, USA. ACM.

Arapakis, I., Cambazoglu, B. B., and Lalmas, M. (2014). On the feasibility of predicting news popularity at cold start. In *International Conference on Social Informatics*, pages 290–299. Springer.

Arapakis, I., Jose, J. M., and Gray, P. D. (2008). Affective feedback: an investigation into the role of emotions in the information seeking process. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 395–402. ACM.

Arguello, J., Diaz, F., Callan, J., and Crespo, J.-F. (2009). Sources of evidence for vertical selection. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 315–322, New York, NY, USA. ACM.

Athanasakos, K., Stathopoulos, V., and Jose, J. M. (2010). A framework for evaluating automatic image annotation algorithms. In *ECIR '10 Milton Keynes, UK*.

Ayache, S., Quénot, G., and Gensel, J. (2007). Classifier fusion for svm-based multimedia semantic indexing. In *European Conference on Information Retrieval*, pages 494–504. Springer.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Bai, J., Nie, J.-Y., Cao, G., and Bouchard, H. (2007). Using query contexts in information retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 15–22, New York, NY, USA. ACM.

Baldauf, M., Dustdar, S., and Rosenberg, F. (2007). A survey on context-aware systems. *International Journal of Ad Hoc and Ubiquitous Computing*, 2(4):263–277.

Bandari, R., Asur, S., and Huberman, B. A. (2012). The pulse of news in social media: Forecasting popularity. In *ICWSM*.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359.

Belkin, N. J. and Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12):29–38.

Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. (2010). Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Bolaños, M. and Radeva, P. (2015). Ego-object discovery. *arXiv preprint arXiv:1504.01639*.

Boureau, Y.-L., Bach, F., LeCun, Y., and Ponce, J. (2010). Learning mid-level features for recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2559–2566. IEEE.

Bredin, H., Byrne, D., Lee, H., O'Connor, N. E., and Jones, G. J. (2008). Dublin city university at the trecvid 2008 bbc rushes summarisation task. In *Proceedings of the 2Nd ACM TRECVid Video Summarization Workshop*, TVS '08, pages 45–49, New York, NY, USA. ACM.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics.

Burke, R. (2007). Hybrid web recommender systems. In *The adaptive web*, pages 377–408. Springer.

Cai, D., He, X., Li, Z., Ma, W.-Y., and Wen, J.-R. (2004). Hierarchical clustering of www image search results using visual, textual and link information. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, pages 952–959, New York, NY, USA. ACM.

Caputo, B., Müller, H., Martinez-Gomez, J., Villegas, M., Acar, B., Patricia, N., Marvasti, N., Üsküdarlı, S., Paredes, R., Cazorla, M., et al. (2014). Imageclef 2014: Overview and analysis of the results. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, pages 192–211. Springer.

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *In Research and Development in Information Retrieval*, pages 335–336.

Carneiro, G., Chan, A. B., Moreno, P. J., and Vasconcelos, N. (2007). Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:2007.

Carvalho, V. R., Lease, M., and Yilmaz, E. (2011). Crowdsourcing for search evaluation. In *ACM Sigir forum*, volume 44, pages 17–22. ACM.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. (2007). I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, pages 1–14, New York, NY, USA. ACM.

Cha, M., Mislove, A., and Gummadi, K. P. (2009). A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 721–730, New York, NY, USA. ACM.

Chakrabarti, D. and Punera, K. (2011). Event summarization using tweets. In *ICWSM*.

Chan, T.-H., Jia, K., Gao, S., Lu, J., Zeng, Z., and Ma, Y. (2015). Pcanet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing*, 24(12):5017–5032.

Chang, C. H., Kayed, M., Girgis, M., and Shaalan, K. (2006). A survey of web information extraction systems. *Knowledge and Data Engineering, IEEE Transactions on*, 18(10):1411–1428.

Chapelle, O. and Chang, Y. (2011). Yahoo! learning to rank challenge overview. In *Yahoo! Learning to Rank Challenge*, pages 1–24.

Chen, H.-M., Chang, M.-H., Chang, P.-C., Tien, M.-C., Hsu, W. H., and Wu, J.-L. (2008). Sheepdog: group and tag recommendation for flickr photos by automatic search-based learning. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, pages 737–740, New York, NY, USA. ACM.

Chen, W.-Y., Chu, J.-C., Luan, J., Bai, H., Wang, Y., and Chang, E. Y. (2009). Collaborative filtering for orkut communities: discovery of user latent behavior. In *Proceedings of the 18th international conference on World wide web*, pages 681–690. ACM.

Chen, X. and Zhang, X. (2003). A popularity-based prediction model for web prefetching. *Computer*, 36(3):63–70.

Chen, Y., Wang, J., and Krovetz, R. (2005). Clue: cluster-based retrieval of images by unsupervised learning. *Image Processing, IEEE Transactions on*, 14(8):1187–1201.

Cheng, E. (2006). Scalable relevance feedback using click-through data for web image retrieval.

Cheng, E., Jing, F., Zhang, L., and Jin, H. (2006). Scalable relevance feedback using click-through data for web image retrieval. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 173–176. ACM.

Chirita, P. A., Costache, S., Handschuh, S., and Nejdl, W. (2007). PTAG: Large Scale Automatic Generation of Personalized Annotation TAGs for the Web.

Cho, Y. H., Kim, J. K., and Kim, S. H. (2002). A personalized recommender system based on web usage mining and decision tree induction. *Expert systems with Applications*, 23(3):329–342.

Chowdhury, S., Ferdous, M. S., and Jose, J. M. (2016). A user-study examining visualization of lifelogs. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE.

Christel, M. G., Hauptmann, A. G., Lin, W.-H., Chen, M.-Y., Yang, J., Maher, B., and Baron, R. V. (2008). Exploring the utility of fast-forward surrogates for bbc rushes. In *Proceedings of the 2Nd ACM TRECVid Video Summarization Workshop*, TVS '08, pages 35–39, New York, NY, USA. ACM.

Chum, O., Philbin, J., and Zisserman, A. (2008). Near duplicate image detection: min-hash and tf-idf weighting. In *British Machine Vision Conference*.

Cilibrasi, R. L. and Vitanyi, P. M. B. (2007). The google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383.

Clarke, C., Craswell, N., and Soboroff, I. (2009). Preliminary report on the trec 2009 Web track. *TREC 2009 Notebook*.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., and Sartin, M. (1999). Combining content-based and collaborative filters in an online newspaper.

Cleverdon, C. (1967). The cranfield tests on index language devices. *Aslib Proceedings*, 19(6):173–194.

Cleverdon, C., Mills, J., Project, A. C. R., and Keen, M. (1966). *Factors Determining the Performance of Indexing Systems: Design*. Factors Determining the Performance of Indexing Systems. College of Aeronautics.

Clough, P., Joho, H., and Sanderson, M. (2005). Automatically organising images using concept hierarchies. In *proceedings of the Multimedia Workshop running at ACM SIGIR conference*. Sheffield.

Clough, P. and Sanderson, M. (2013). Evaluating the performance of information retrieval systems using test collections. *Information Research*, 18(2).

Cutrell, E. and Guan, Z. (2007). What are you looking for?: An eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 407–416, New York, NY, USA. ACM.

Das, D. and Martins, A. F. (2007). A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4:192–195.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):5:1–5:60.

De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

Del Fabro, M. and Boszormenyi, L. (2012). Summarization and presentation of real-life events using community-contributed content. In Schoeffmann, K., Merialdo, B., Hauptmann, A., Ngo, C.-W., Andreopoulos, Y., and Breiteneder, C., editors, *Advances in Multimedia Modeling*, volume 7131 of *Lecture Notes in Computer Science*, pages 630–632. Springer Berlin Heidelberg.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR09*.

Deng, Y., Manjunath, B., Kenney, C., Moore, M., and Shin, H. (2001). An efficient color representation for image retrieval. *Image Processing, IEEE Transactions on*, 10(1):140–147.

Deselaers, T., Keysers, D., and Ney, H. (2008). Features for image retrieval: an experimental comparison. *Information Retrieval*, 11(2):77–107.

Dhar, S., Ordonez, V., and Berg, T. (2011). High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664.

Doherty, A., Kelly, P., and Foster, C. (2013). Wearable cameras: Identifying healthy transportation choices. *Pervasive Computing, IEEE*, 12(1):44–47.

Doherty, A. R., Pauly-Takacs, K., Caprani, N., Gurrin, C., Moulin, C. J., O'Connor, N. E., and Smeaton, A. F. (2012). Experiences of aiding autobiographical memory using the sensecam. *Human–Computer Interaction*, 27(1-2):151–174.

Doherty, A. R. and Smeaton, A. F. (2008). Automatically segmenting lifelog data into events. In *Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, WIAMIS '08, pages 20–23, Washington, DC, USA. IEEE Computer Society.

Doherty, A. R. and Smeaton, A. F. (2010). Automatically augmenting lifelog events using pervasively generated content from millions of people. *Sensors*, 10(3):1423.

Drosou, M. and Pitoura, E. (2010). Search result diversification. *SIGMOD Rec.*, 39(1):41–47.

Du, L., Ren, L., Dunson, D. B., and Carin, L. (2009). A Bayesian model for simultaneous image clustering, annotation and object segmentation.

Duan, Y., Jiang, L., Qin, T., Zhou, M., and Shum, H.-Y. (2010). An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 295–303, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., and Tomkins, A. (2007). Visualizing tags over time. *ACM Transactions on the Web (TWEB)*, 1(2):7.

Duhan, N., Sharma, A., and Bhatia, K. (2009). Page ranking algorithms: A survey. In *Advance Computing Conference, 2009. IACC 2009. IEEE International*, pages 1530–1537.

Dumais, S. T. (2005). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230.

Duygulu, P., Barnard, K., De Freitas, J., and Forsyth, D. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *ECCV '02*.

Eickhoff, C. and Vries, A. (2013). Increasing cheat robustness of crowdsourcing tasks. *Inf. Retr.*, 16(2):121–137.

Elahi, N., Karlsen, R., and Holsbø, E. J. (2013). Personalized photo recommendation by leveraging user modeling on social network. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services*, IIWAS '13, pages 68:68–68:71, New York, NY, USA. ACM.

Ellis, D. P. and Lee, K. (2006). Accessing minimal-impact personal audio archives. *IEEE Multimedia*, 13(4):30–38.

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open information extraction from the web. *Commun. ACM*, 51(12):68–74.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., and Pappas, G. (2008). Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. *The FASEB journal*, 22(2):338–342.

Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., and Equitz, W. (1994). Efficient and effective querying by image content. *J. Intell. Inf. Syst.*, 3(3-4):231–262.

Fan, J., Gao, Y., and Luo, H. (2007). Hierarchical classification for automatic image annotation. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 111–118, New York, NY, USA. ACM.

Fei-Fei, L., Fergus, R., and Perona, P. (2007). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70.

Feng, S., Manmatha, R., and Lavrenko, V. (2004). Multiple bernoulli relevance models for image and video annotation. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–1002 – II–1009 Vol.2.

Feng, Y. and Lapata, M. (2010). Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 831–839, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fergus, R., Fei-Fei, L., Perona, P., and Zisserman, A. (2005). Learning object categories from google's image search. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1816–1823. IEEE.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M. (2010). Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 80–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P. (1995). Query by image and video content: The qbic system. *Computer*, 28(9):23–32.

Foo, J. J. and Sinha, R. (2007). Pruning sift for scalable near-duplicate image matching. In *Proceedings of the Eighteenth Conference on Australasian Database - Volume 63*, ADC '07, pages 63–71, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

Foo, J. J., Zobel, J., Sinha, R., and Tahaghoghi, S. M. M. (2007). Detection of near-duplicate images for web search. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, CIVR '07, pages 557–564, New York, NY, USA. ACM.

Gao, Y., Wang, M., Tao, D., Ji, R., and Dai, Q. (2012). 3-d object retrieval and recognition with hypergraph analysis. *Image Processing, IEEE Transactions on*, 21(9):4290–4303.

Gao, Y., Wang, M., Zha, Z.-J., Shen, J., Li, X., and Wu, X. (2013). Visual-textual joint relevance learning for tag-based social image search. *Image Processing, IEEE Transactions on*, 22(1):363–376.

Garg, N. and Weber, I. (2008). Personalized, interactive tag recommendation for flickr. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 67–74.

Gilbert, A., Piras, L., Wang, J., Yan, F., Dellandrea, E., Gaizauskas, R., Villegas, M., and Mikolajczyk, K. (2015). Overview of the imageclef 2015 scalable image annotation, localization and sentence generation task. In *CLEF2015 Working Notes. CEUR Workshop Proceedings, CEUR-WS. org, Toulouse, France (September 8-11 2015)*.

Goh, K.-S., Chang, E., and Li, B. (2005). Using one-class and two-class svms for multiclass image annotation. *Knowledge and Data Engineering, IEEE Transactions on*, 17(10):1333–1346.

Grubinger, M., Clough, P., Muller, H., and Deselaers, T. (2006). The iapr tc-12 benchmark – a new evaluation resource for visual information systems.

Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., and Albatal, R. (2016). Ntcir lifelog: The first test collection for lifelog research. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 705–708, New York, NY, USA. ACM.

Gurrin, C., Qiu, Z., Hughes, M., Caprani, N., Doherty, A., Hodges, S., and Smeaton, A. (2013). The smartphone as a platform for wearable cameras in health research. *American Journal of Preventive Medicine*.

Gurrin, C., Smeaton, A. F., and Doherty, A. R. (2014). Lifelogging: Personal big data. *Foundations and trends in information retrieval*, 8(1):1–125.

Hanbury, A. (2008). A survey of methods for image annotation. *J. Vis. Lang. Comput.*, 19(5):617–627.

Harman, D. (2011). *Information Retrieval Evaluation*. Morgan & Claypool Publishers, 1st edition.

Harman, D. K. (1992). The first text retrieval conference (trec-1) rockville, md, usa, 4–6 november, 1992. *Information Processing & Management*, 29(4):411–414.

Hauptmann, A., Yan, R., and Lin, W.-H. (2007). How many high-level concepts will fill the semantic gap in news video retrieval? In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, CIVR '07, pages 627–634, New York, NY, USA. ACM.

He, Q., Pei, J., Kifer, D., Mitra, P., and Giles, L. (2010). Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 421–430. ACM.

Hirth, M., Hoßfeld, T., and Tran-Gia, P. (2010). Cheat-detection mechanisms for crowdsourcing. Technical Report 474, University of Würzburg.

Hollink, L., Schreiber, A. T., Wielinga, B. J., and Worring, M. (2004). Classification of user image descriptions. *Int. J. Hum.-Comput. Stud.*, 61(5):601–626.

Hong, J. and Baker, C. F. (2011). How good is the crowd at "real" wsd? In *Proceedings of the 5th Linguistic Annotation Workshop*, LAW V '11, pages 30–37, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hong, L., Dan, O., and Davison, B. D. (2011). Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 57–58, New York, NY, USA. ACM.

Hou, J. and Pelillo, M. (2013). A simple feature combination method based on dominant sets. *Pattern Recognition*, 46(11):3129–3139.

Hsiao, Y.-S., Sanchez-Riera, J., Lim, T., Hua, K.-L., and Cheng, W.-H. (2014). Lared: A large rgb-d extensible hand gesture dataset. In *Proceedings of the 5th ACM Multimedia Systems Conference*, MMSys '14, pages 53–58, New York, NY, USA. ACM.

Huiskes, M. and Lew, M. (2008). The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA. ACM.

Huiskes, M., Thomee, B., and Lew, M. (2010). New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. In *MIR '10*, pages 527–536, New York, NY, USA. ACM.

Ingwersen, P. and Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Jaimes, A. and fu Chang, S. (2000). A conceptual framework for indexing visual information at multiple levels. In *IN PROCEEDINGS OF SPIE INTERNET IMAGING 2000*, pages 2–15.

Jamali, S. and Rangwala, H. (2009). Digging digg: Comment mining, popularity prediction, and social network analysis. In *Web Information Systems and Mining, 2009. WISM 2009. International Conference on*, pages 32–38.

Jansen, B. J., Spink, A., Bateman, J., and Saracevic, T. (1998). Real life information retrieval: A study of user queries on the web. *SIGIR Forum*, 32(1):5–17.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., and Stumme, G. (2007). Tag recommendations in folksonomies. In *Knowledge Discovery in Databases: PKDD 2007*, pages 506–514. Springer.

Jeon, J., Lavrenko, V., and Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR 2003*, SIGIR '03, pages 119–126, NY, USA.

Jhuang, H. and Chikkerur, S. (2006). Video shot boundary detection using gist.

Jiang, W., Loui, A. C., and Cerosaletti, C. D. (2010). Automatic aesthetic value assessment in photographic images. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 920–925. IEEE.

Jojic, N., Perina, A., and Murino, V. (2010). Structural epitome: a way to summarize one's visual experience. In *Advances in neural information processing systems*, pages 1027–1035.

Jones, G. J. and Brown, P. J. (2004). The role of context in information retrieval.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H., and Hidaka, S. (1999). Overview of ir tasks at the first ntcir workshop. In *Proceedings of the first NTCIR workshop on research in Japanese text retrieval and term recognition*, pages 11–44.

Kanoulas, E., Carterette, B., Clough, P. D., and Sanderson, M. (2011). Evaluating multi-query sessions. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1053–1062, New York, NY, USA. ACM.

Kato, T. (1992). Database architecture for content-based image retrieval.

Katti, H., Bin, K. Y., Chua, T. S., and Kankanhalli, M. (2008). Pre-attentive discrimination of interestingness in images. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 1433–1436.

Ke, Y. and Sukthankar, R. (2004). Pca-sift: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–506. IEEE.

Ke, Y., Sukthankar, R., Huston, L., Ke, Y., and Sukthankar, R. (2004). Efficient near-duplicate detection and sub-image retrieval. In *In ACM Multimedia*, pages 869–876.

Kekäläinen, J. and Järvelin, K. (1998). The impact of query structure and query expansion on retrieval performance. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 130–137, New York, NY, USA. ACM.

Kekäläinen, J. and Järvelin, K. (2002). Using graded relevance assessments in ir evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129.

Kelby, S. (2012). *The digital photography book*. Peachpit Press.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retr.*, 3(1—2):1–224.

Kikhia, B., Boytsov, A., Hallberg, J., ul Hussain Sani, Z., Jonsson, H., and Synnes, K. (2014). Structuring and presenting lifelogs based on location data. In Cipresso, P., Matic, A., and Lopez, G., editors, *Pervasive Computing Paradigms for Mental Health*, volume 100 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 133–144. Springer International Publishing.

Kilgarriff, A. (2001). Comparing corpora. *International journal of corpus linguistics*, 6(1):97–133.

Kim, S., Park, S., and Kim, M. (2004). Image classification into object / non-object classes. In Enser, P., Kompatsiaris, Y., O Connor, N., Smeaton, A., and Smeulders, A., editors, *Image and Video Retrieval*, volume 3115 of *Lecture Notes in Computer Science*, pages 393–400. Springer Berlin Heidelberg.

Kittur, A. and Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46. ACM.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 91–101, New York, NY, USA. ACM.

Kohavi, R. (2015). Online controlled experiments: Lessons from running a/b/n tests for 12 years. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1–1, New York, NY, USA. ACM.

Kong, S. G., Heo, J., Abidi, B. R., Paik, J., and Abidi, M. A. (2005). Recent advances in visual and infrared face recognition a review. *Computer Vision and Image Understanding*, 97(1):103–135.

Koren, Y. (2009). Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 447–456, New York, NY, USA. ACM.

Kozat, S., Venkatesan, R., and Mihcak, M. (2004). Robust perceptual image hashing via matrix invariants. In *Image Processing, 2004. ICIP '04. 2004 International Conference on*, volume 5, pages 3443–3446 Vol. 5.

Krestel, R., Fankhauser, P., and Nejdl, W. (2009). Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 61–68, New York, NY, USA. ACM.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In Bartlett, P., Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 25*, pages 1106–1114. NIPS.

Kuo, B. Y.-L., Hentrich, T., Good, B. M. ., and Wilkinson, M. D. (2007). Tag clouds for summarizing web search results. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 1203–1204, New York, NY, USA. ACM.

Larson, M. A., Ionescu, B., Sjöberg, M., Anguera, X., Poignant, J., Riegler, M., Eskevich, M., Hauff, C., Sutcliffe, R. F. E., Jones, G. J. F., Yang, Y., Soleymani, M., and Papadopoulos, S., editors (2015). *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15, 2015*, volume 1436 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Lau, T. and Horvitz, E. (1999). Patterns of search: Analyzing and modeling web query refinement. In Kay, J., editor, *UM99 User Modeling*, volume 407 of *CISM International Centre for Mechanical Sciences*, pages 119–128. Springer Vienna.

Lavrenko, V., Choquette, M., and Croft, W. B. (2002). Cross-lingual relevance models. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 175–182, New York, NY, USA. ACM.

Lavrenko, V., Manmatha, R., and Jeon, J. (2003). A model for learning the semantics of pictures. In *IN NIPS*. MIT Press.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2169–2178. IEEE.

Lazer, D. (2009). Life in the network: the coming age of computational social science. *PMC 2009 Sep 16*.

Lee, K. S., Croft, W. B., and Allan, J. (2008). A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 235–242, New York, NY, USA. ACM.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2014). DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*.

Lerman, K. and Hogg, T. (2010). Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 621–630, New York, NY, USA. ACM.

Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference*, KDD '09, pages 497–506, New York, NY, USA. ACM.

Levar, M. and Castro, N. T. (2013). New trends in ir: Sigir 2013. *SIGIR Forum*.

Li, C., Loui, A. C., and Chen, T. (2010a). Towards aesthetics: A photo quality assessment and photo selection system. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 827–830, New York, NY, USA. ACM.

Li, C.-T., Shan, M.-K., Jheng, S.-H., and Chou, K.-C. (2016a). Exploiting concept drift to predict popularity of social multimedia in microblogs. *Information Sciences*, 339:310–331.

Li, L.-J., Su, H., Fei-Fei, L., and Xing, E. P. (2010b). Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, pages 1378–1386.

Li, N., Crane, M., Gurrin, C., and Ruskin, H. J. (2016b). Finding motifs in large personal lifelogs. In *Proceedings of the 7th Augmented Human International Conference 2016*, AH '16, pages 9:1–9:8, New York, NY, USA. ACM.

Li, X., Snoek, C. G., and Worring, M. (2010c). Unsupervised multi-feature tag relevance learning for social image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 10–17. ACM.

Li, X., Uricchio, T., Ballan, L., Bertini, M., Snoek, C. G., and Bimbo, A. D. (2016c). Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Computing Surveys (CSUR)*, 49(1):14.

Li, Y., Crandall, D. J., and Huttenlocher, D. P. (2009). Landmark classification in large-scale image collections. In *ICCV*, pages 1957–1964.

Lim, K. H., Chan, J., Leckie, C., and Karunasekera, S. (2015). Improving location prediction using a social historical model with strict recency context. *Proc. of CaRR*, 15.

Liu, D., Hua, X., Yang, L., Wang, M., and Zhang, H. (2009). Tag ranking. In *Proceedings of the 18th international conference on World wide web*, pages 351–360. ACM.

Liu, D., Hua, X.-S., Wang, M., and Zhang, H.-J. (2010). Image retagging. In *Proceedings of the international conference on Multimedia*, MM '10, pages 491–500, New York, NY, USA. ACM.

Long, R., Wang, H., Chen, Y., Jin, O., and Yu, Y. (2011). Towards effective event detection, tracking and summarization on microblog data. In *Proceedings of the 12th International Conference on Web-age Information Management*, WAIM'11, pages 652–663, Berlin, Heidelberg. Springer-Verlag.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165.

Ma, H., Zhu, J., Lyu, M. R., and King, I. (2010). Bridging the semantic gap between image contents and tags. *IEEE Transactions on Multimedia*, 12(5):462–473.

Makadia, A., Pavlovic, V., and Kumar, S. (2010). Baselines for image annotation. *International Journal of Computer Vision*, 90(1):88–105.

Manjunath, B. S. (2002). *Introduction to MPEG-7, Multimedia Content Description Interface*. John Wiley and Sons, Ltd.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 171–189. Springer.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., and Miller, R. C. (2011). Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 227–236. ACM.

McCulloch, W. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

McMinn, A. J. and Jose, J. M. (2015). Real-time entity-based event detection for twitter. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 65–77. Springer International Publishing.

McMinn, A. J., Moshfeghi, Y., and Jose, J. M. (2013). Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, pages 409–418, New York, NY, USA. ACM.

Melucci, M. (2012). *Contextual Search*. Now Publishers Inc., Hanover, MA, USA.

Melville, P., Mooney, R. J., and Nagarajan, R. (2002). Content-boosted collaborative filtering for improved recommendations. In *Eighteenth National Conference on Artificial Intelligence*, pages 187–192, Menlo Park, CA, USA. American Association for Artificial Intelligence.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Monay, F. and Gatica-Perez, D. (2004). Plsa-based image auto-annotation: Constraining the latent space. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, pages 348–351, New York, NY, USA. ACM.

Moshfeghi, Y. and Jose, J. M. (2013a). An effective implicit relevance feedback technique using affective, physiological and behavioural features. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 133–142. ACM.

Moshfeghi, Y. and Jose, J. M. (2013b). On cognition, emotion, and interaction aspects of search tasks with different search intentions. In *Proceedings of the 22nd international conference on World Wide Web*, pages 931–942. International World Wide Web Conferences Steering Committee.

Mousselly-Sergieh, H., Watzinger, D., Huber, B., Döller, M., Egyed-Zsigmond, E., and Kosch, H. (2014). World-wide scale geotagged image dataset for automatic image annotation and reverse geotagging. In *Proceedings of the 5th ACM Multimedia Systems Conference*, pages 47–52. ACM.

Müller, H., Deselaers, T., Deserno, T., Clough, P., Kim, E., and Hersh, W. (2006). Overview of the imageclefmed 2006 medical retrieval and medical annotation tasks. In *Evaluation of Multilingual and Multi-modal Information Retrieval*, pages 595–608. Springer Berlin Heidelberg.

Muller, H., Marchand-Maillet, S., and Pun, T. (2002). The truth about corel - evaluation in image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, CIVR '02, pages 38–49, London, UK, UK. Springer-Verlag.

Murdock, V. (2014). Dynamic location models. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1231–1234. ACM.

Mylonas, P., Vallet, D., Castells, P., Fernández, M., and Avrithis, Y. (2008). Personalized information retrieval based on context and ontological knowledge. *The Knowledge Engineering Review*, 23(01):73–100.

Naaman, M., Harada, S., Wang, Q., Garcia-Molina, H., and Paepcke, A. (2004). Context data in geo-referenced digital photo collections. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, pages 196–203, New York, NY, USA. ACM.

Naci, S. U., Damnjanovic, U., Mansencal, B., Benois-Pineau, J., Kaes, C., Corvaglia, M., Rossi, E., and Aginako, N. (2008). The cost292 experimental framework for rushes summarization task in trecvid 2008. In *Proceedings of the 2Nd ACM TRECVid Video Summarization Workshop*, TVS '08, pages 40–44, New York, NY, USA. ACM.

Nathenson, I. S. (1998). Internet infoglut and invisible ink: Spamdexing search engines with meta tags. *Harvard Journal of Law and Technology*, 12(1).

Nicholas, I. S. C. and Nicholas, C. K. (1999). Combining content and collaboration in text filtering. In *In Proceedings of the IJCAI 99 Workshop on Machine Learning for Information Filtering*, pages 86–91.

Nichols, J., Mahmud, J., and Drews, C. (2012). Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 189–198. ACM.

Niu, X., Li, L., Mei, T., Shen, J., and Xu, K. (2012). Predicting image popularity in an incomplete social media community by a weighted bi-partite graph. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pages 735–740.

Nowak, S. and Dunker, P. (2010). Overview of the clef 2009 large-scale visual concept detection and annotation task. In Peters, C., Caputo, B., Gonzalo, J., Jones, G. J., Kalpathy-Cramer, J., Muller, H., and Tsikrika, T., editors, *Multilingual Information Access Evaluation II. Multimedia Experiments*, volume 6242 of *Lecture Notes in Computer Science*, pages 94–109. Springer Berlin Heidelberg.

Nowak, S., Nagel, K., and Liebetrau, J. (2011). The clef 2011 photo annotation and concept-based retrieval tasks. In *CLEF (Notebook Papers/Labs/Workshop)*, pages 1–25.

Nowak, S. and Rüger, S. (2010). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pages 557–566, New York, NY, USA. ACM.

Oleson, D., Sorokin, A., Laughlin, G. P., Hester, V., Le, J., and Biewald, L. (2011). Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. *Human computation*, 11(11).

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175.

Oliva, A. and Torralba, A. (2006). Chapter 2 building the gist of a scene: the role of global image features in recognition. In S. Martinez-Conde, S.L. Macknik, L.M. Martinez, J.-M. Alonso and P.U. Tse, editor, *Visual Perception Fundamentals of Awareness: Multi-Sensory Integration and High-Order Perception*, volume 155, Part B of *Progress in Brain Research*, pages 23–36. Elsevier.

Ondrej, M., Zboril Frantisek, V., and Martin, D. (2007). Algorithmic and mathematical principles of automatic number plate recognition systems. *BRNO University of technology*, page 10.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724.

Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., and Ounis, I. (2012). Bieber no more: First story detection using twitter and wikipedia. In *Proceedings of TAIA'12*.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics.

Pak, R. and Price, M. M. (2008). Designing an information search interface for younger and older adults. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(4):614–628.

Panofsky, E. (1972). *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*. Harper & Row.

Paramita, M., Sanderson, M., and Clough, P. (2010). Diversity in photo retrieval: Overview of the imageclef photo task 2009. In Peters, C., Caputo, B., Gonzalo, J., Jones, G. J., Kalpathy-Cramer, J., Muller, H., and Tsikrika, T., editors, *Multilingual Information Access Evaluation II. Multimedia Experiments*, volume 6242 of *Lecture Notes in Computer Science*, pages 45–59. Springer Berlin Heidelberg.

Park, S. B., Lee, J. W., and Kim, S. K. (2004). Content-based image classification using a neural network. *Pattern Recogn. Lett.*, 25(3):287–300.

Pass, G., Zabih, R., and Miller, J. (1996). Comparing images using color coherence vectors. In *In Proceedings of ACM Multimedia 96*, pages 65–73.

Pazzani, M. J. and Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web*, pages 325–341. Springer.

Petrović, S., Osborne, M., and Lavrenko, V. (2010). Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pettersen, S. A., Johansen, D., Johansen, H., Berg-Johansen, V., Gaddam, V. R., Mortensen, A., Langseth, R., Griwodz, C., Stensland, H. K., and Halvorsen, P. (2014). Soccer video and player position dataset. In *Proceedings of the 5th ACM Multimedia Systems Conference*, MMSys '14, pages 18–23, New York, NY, USA. ACM.

Pham, T.-T., Maillot, N. E., Lim, J.-H., and Chevallet, J.-P. (2007). Latent semantic fusion model for image retrieval and annotation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 439–444. ACM.

Picault, J., Ribiere, M., and Gaste, Y. (2013). Indexing video segments using microblogs. In *Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop on*, pages 155–160.

Rae, A., Sigurbjörnsson, B., and van Zwol, R. (2010). Improving tag recommendation using social networks. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, pages 92–99, Paris, France, France. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

Rattenbury, T., Good, N., and Naaman, M. (2007). Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110. ACM.

Ravikumar, S., Balakrishnan, R., and Kambhampati, S. (2012). Ranking tweets considering trust and relevance. In *Proceedings of the Ninth International Workshop on Information Integration on the Web*, IIWeb '12, pages 4:1–4:4, New York, NY, USA. ACM.

Rayson, P., Leech, G. N., and Hodges, M. (1997). Social differentiation in the use of english vocabulary: some analyses of the conversational component of the british national corpus. *International Journal of Corpus Linguistics*, 2(1):133–152.

Ricci, F., Rokach, L., and Shapira, B. (2011). *Introduction to recommender systems handbook*. Springer.

Rivest, R. (1992). The md5 message-digest algorithm. *RFC*.

Robertson, S. (2008). On the history of evaluation in ir. *Journal of Information Science*.

Robertson, S. E. and Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146.

Robertson, S. E. and Sparck Jones, K. (1988). Document retrieval systems. chapter Relevance Weighting of Search Terms, pages 143–160. Taylor Graham Publishing, London, UK, UK.

Rubens, N., Kaplan, D., and Sugiyama, M. (2011). Active learning in recommender systems. In *Recommender systems handbook*, pages 735–767. Springer.

Rui, Y. and Huang, T. (2000). Optimizing learning in image retrieval. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 236–243 vol.1.

Rui, Y., Huang, T., Ortega, M., and Mehrotra, S. (1998). Relevance feedback: a power tool for interactive content-based image retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 8(5):644–655.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2014). Imagenet large scale visual recognition challenge.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Sahuguet, M. and Huet, B. (2013). Socially motivated multimedia topic timeline summarization. In *Proceedings of the 2Nd International Workshop on Socially-aware Multimedia*, SAM '13, pages 19–24, New York, NY, USA. ACM.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA. ACM.

Salton, G., editor (1971). *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, Englewood, Cliffs, New Jersey.

Salton, G. and Buckley, C. (1997). Improving retrieval performance by relevance feedback. *Readings in information retrieval*, 24(5):355–363.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., and Sperling, J. (2009). Twitterstand: news in tweets. In *ACM SIGSPATIAL '09'*. ACM.

Santos, R. L., Macdonald, C., and Ounis, I. (2010). Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 881–890, New York, NY, USA. ACM.

Santos, R. L., Macdonald, C., and Ounis, I. (2011). Intent-aware search result diversification. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 595–604, New York, NY, USA. ACM.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA. ACM.

Sayeed, A. B., Rusk, B., Petrov, M., Nguyen, H. C., Meyer, T. J., and Weinberg, A. (2011). Crowdsourcing syntactic relatedness judgements for opinion mining in the study of information technology adoption. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 69–77. Association for Computational Linguistics.

Schettini, R., Ciocca, G., Zuffi, S., et al. (2001). A survey of methods for colour image indexing and retrieval in image databases. *Color Imaging Science: Exploiting Digital Media*, pages 183–211.

Schinas, M., Papadopoulos, S., Kompatsiaris, Y., and Mitkas, P. A. (2015). Visual event summarization on social media using topic modelling and graph-based ranking algorithms. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 203–210. ACM.

Schinas, M., Papadopoulos, S., Kompatsiaris, Y., and Mitkas, P. A. (2016). Mgraph: multimodal event summarization in social media using topic models and graph-based ranking. *International Journal of Multimedia Information Retrieval*, 5(1):51–69.

Semeraro, G., Basile, P., de Gemmis, M., and Lops, P. (2009). User profiles for personalizing digital libraries.

Shamma et al. (2010). Statler: Summarizing media through short-message services. In *In Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*.

Shane (2006). ZoneTag: Designing Context-Aware Mobile Media Capture to Increase Participation. In *In: Proceedings of the Pervasive Image Capture and Sharing: New Social Practices and Implications for Technology Workshop (PICS 2006) at the Eighth International Conference on Ubiquitous Computing (UbiComp 2006)*.

Sharifi, B. P., Inouye, D. I., and Kalita, J. K. (2013). Summarization of twitter microblogs. *The Computer Journal*.

Shatford, S. (1986). Analyzing the subject of a picture: A theoretical approach. *Cataloging & Classification Quarterly*, 6(3):39–62.

Shen, X., Tan, B., and Zhai, C. (2005). Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50. ACM.

Shen, Y. and Fan, J. (2010). Leveraging loosely-tagged images and inter-object correlations for tag recommendation. In *ACM Multimedia*, pages 5–14.

Shrager, J. and Johnson, M. (1995). Timing in the development of cortical function: A computational approach. *Maturational windows and adult cortical plasticity. New York: Addison-Wesley*.

Siersdorfer, S. and Sizov, S. (2009). Social recommender systems for web 2.0 folksonomies. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, HT '09, pages 261–270, New York, NY, USA. ACM.

Sigurbjörnsson, B. and van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*, pages 327–336.

Sikirić, I., Brkić, K., and Šegvić, S. (2013). Classifying traffic scenes using the gist image descriptor. *arXiv preprint arXiv:1310.0316*.

Silva, A. and Martins, B. (2011). Tag recommendation for georeferenced photos. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '11, pages 57–64, New York, NY, USA. ACM.

Singh, S., Gupta, A., and Efros, A. A. (2012). Unsupervised discovery of mid-level discriminative patches. In *Computer Vision–ECCV 2012*, pages 73–86. Springer.

Sivic, J. and Zisserman, A. (2003). Video google: a text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477 vol.2.

Smeulders, A., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *Pattern Analysis, IEEE*, 22(12):1349–1380.

Smith, J. R. and Chang, S.-F. (1996). Visualseek: A fully automated content-based image query system. In *Proceedings of the Fourth ACM International Conference on Multimedia*, MULTIMEDIA '96, pages 87–98, New York, NY, USA. ACM.

Snoek, C. G., Worring, M., and Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM.

Spärck-Jones, K., Robertson, S. E., and Sanderson, M. (2007). Ambiguous requests: implications for retrieval tests, systems and theories. *SIGIR Forum*, 41(2):8–17.

Spina, D., Meij, E., De Rijke, M., Oghina, A., Bui, M. T., and Breuss, M. (2012). Identifying entity aspects in microblog posts. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1089–1090. ACM.

Spink, A., Wolfram, D., Jansen, M. B. J., and Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234.

Srivastava, N. and Salakhutdinov, R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230.

Stern, D. H., Herbrich, R., and Graepel, T. (2009). Matchbox: large scale online bayesian recommendations. In *Proceedings of the 18th international conference on World wide web*, pages 111–120. ACM.

Stockman, G. and Shapiro, L. G. (2001). *Computer Vision*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

Stricker, M. and Orengo, M. (1995). Similarity of color images. In *Proceedings of the SPIE*, pages 381–392.

Su, X. and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009:4.

Szabo, G. and Huberman, B. A. (2010). Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88.

Szeliski, R. (2006). Image alignment and stitching: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 2(1):1–104.

Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

Takashita, T., Itokawa, T., Kitasuka, T., and Aritsugi, M. (2010). Tag recommendation for flickr using web browsing behavior. In *International Conference on Computational Science and Its Applications*, pages 412–421. Springer.

Tang, Z., Dai, Y., and Zhang, X. (2012). Perceptual hashing for color images using invariant moments. *Appl. Math*, 6(2S):643S–650S.

Tatar, A., Antoniadis, P., De Amorim, M. D., and Fdida, S. (2014). From popularity prediction to ranking online news. *Social Network Analysis and Mining*, 4(1):1–12.

Tatar, A., Leguay, J., Antoniadis, P., Limbourg, A., de Amorim, M. D., and Fdida, S. (2011). Predicting the popularity of online articles based on user comments. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 67:1–67:8, New York, NY, USA. ACM.

Thomee, B., Moreno, J. G., and Shamma, D. A. (2014). Who's time is it anyway?: Investigating the accuracy of camera timestamps. In *Proceedings of the ACM International Conference on Multimedia*, pages 909–912. ACM.

Thomee, B. and Popescu, A. (2012). Overview of the imageclef 2012 Flickr photo annotation and retrieval task. In *CLEF (Online Working Notes/Labs/Workshop)*, volume 53, pages 54–58.

Tombros, A. and Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 2–10, New York, NY, USA. ACM.

Tong, H., Li, M., Zhang, H., and Zhang, C. (2004). Blur detection for digital images using wavelet transform. In *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, volume 1, pages 17–20 Vol.1.

Trieschnigg, D., Kraaij, W., and de Jong, F. (2007). The influence of basic tokenization on biomedical document retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 803–804. ACM.

Utgoff, P. E. and Stracuzzi, D. J. (2002). Many-layered learning. *Neural Comput.*, 14(10):2497–2529.

Vailaya, A., Jain, A. K., and Zhang, H. (1998). On image classification: city images vs. landscapes. *Pattern Recognition*, 31(12):1921–1935.

Vailaya, A., Member, A., Figueiredo, M. A. T., Jain, A. K., Zhang, H.-J., and Member, S. (2001). Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10:117–130.

Valafar, M. and Rejaie, R. (2009). Beyond friendship graphs: A study of user interactions in flickr. In *In Proc. ACM SIGCOMM WOSN*.

Valieri, S. and Marin, N. (2012). The optimization of transactional emails in a marketing perspective: Incomedia case.

Vallet, D. and Castells, P. (2012). Personalized diversification of search results. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 841–850, New York, NY, USA. ACM.

van de Sande, K., Gevers, T., and Snoek, C. (2010). Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1582–1596.

van Leuken, R. H., Garcia, L., Olivares, X., and van Zwol, R. (2009). Visual diversification of image search results. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 341–350, New York, NY, USA. ACM.

van Rijsbergen, C. (1979). *Information Retrieval. 1979*. Butterworth.

Vargas, S., Castells, P., and Vallet, D. (2012). Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 75–84, New York, NY, USA. ACM.

Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *Information Processing & Management*, 40(4):677–691.

Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., de Herrera, A. G. S., Bromuri, S., Amin, M. A., Mohammed, M. K., et al. (2015). General overview of imageclef at the clef 2015 labs. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 444–461. Springer.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–511–I–518 vol.1.

Virga, P. and Duygulu, P. (2005). Systematic evaluation of machine translation methods for image and video annotation. In *Image and Video Retrieval*, pages 174–183. Springer.

Vlachos, M., Meek, C., Vagena, Z., and Gunopulos, D. (2004). Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, SIGMOD '04, pages 131–142, New York, NY, USA. ACM.

Von Ahn, L. (2009). Human computation. In *Design Automation Conference, 2009. DAC'09. 46th ACM/IEEE*, pages 418–419. IEEE.

von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 319–326, New York, NY, USA. ACM.

Vuurens, J., Vries, A., and Eickhoff, C. (2011). How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*.

Wan, Y. and Hu, B.-G. (2002). Hierarchical image classification using support vector machines. In *ACCV*.

Wang, F. and Kan, M. (2006). Npic: Hierarchical synthetic image classification using image search and generic features, civr 06. In *Proc. of Conf. on Image and Video Retrieval*, pages 473–482.

Wang, K., Li, X., and Gao, J. (2010a). Multi-style language model for web scale information retrieval. In *Proceedings of the 33rd Annual ACM SIGIR Conference (SIGIR'2010), 19-23 July 2010, Geneva, Switzerland*. Association for Computing Machinery, Inc.

Wang, M. and Hua, X.-S. (2011). Active learning in multimedia annotation and retrieval: A survey. *ACM Trans. Intell. Syst. Technol.*, 2(2):10:1–10:21.

Wang, M., Ni, B., Hua, X.-S., and Chua, T.-S. (2012). Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Computing Surveys (CSUR)*, 44(4):25.

Wang, M., Yang, K., Hua, X.-S., and Zhang, H.-J. (2010b). Towards a relevant and diverse search of social images. *Multimedia, IEEE Transactions on*, 12(8):829–842.

Wang, M., Zhou, X., and Chua, T.-S. (2008). Automatic image annotation via local multi-label classification. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, CIVR '08, pages 17–26, New York, NY, USA. ACM.

Wang, P. and Smeaton, A. (2012). Semantics-based selection of everyday concepts in visual lifelogging. *International Journal of Multimedia Information Retrieval*, 1(2):87–101.

Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics.

Weinberger, K. Q., Slaney, M., and Van Zwol, R. (2008). Resolving tag ambiguity. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, pages 111–120, New York, NY, USA. ACM.

Weng, J. and Lee, B.-S. (2011). Event detection in twitter. In *ICWSM*.

Westerveld, T. and de Vries, A. P. (2003). Experimental evaluation of a generative probabilistic image retrieval model on 'easy' data. In *SIGIR 2003*.

White, R. W., Jose, J. M., and Ruthven, I. (2003). A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing & Management*, 39(5):707–733.

Wood, J., Dykes, J., Slingsby, A., and Clarke, K. (2007). Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1176–1183.

Wu, Z., Ke, Q., Isard, M., and Sun, J. (2009). Bundling features for large scale partial-duplicate web image search. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 25–32.

Yang, B., Mei, T., Hua, X.-S., Yang, L., Yang, S.-Q., and Li, M. (2007). Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 73–80. ACM.

Yang, C., Dong, M., and Hua, J. (2006). Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2057–2063, Washington, DC, USA. IEEE Computer Society.

Zangerle, E., Gassler, W., and Specht, G. (2013). On the impact of text similarity functions on hashtag recommendations in microblogging environments. *Social Network Analysis and Mining*, 3(4):889–898.

Zauner, C. (2010). *Implementation and benchmarking of perceptual image hash functions*.

Zhang, D., Islam, M. M., and Lu, G. (2012a). A review on automatic image annotation techniques. *Pattern Recognition*, 45(1):346–362.

Zhang, H., Korayem, M., You, E., and Crandall, D. J. (2012b). Beyond co-occurrence: Discovering and visualizing tag relationships from geo-spatial and temporal similarities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 33–42, New York, NY, USA. ACM.

Zhang, S., Huang, J., Huang, Y., Yu, Y., Li, H., and Metaxas, D. N. (2010). Automatic image annotation using group sparsity. In *In CVPR*.

Zheng, N., Li, Q., Liao, S., and Zhang, L. (2010). Which photo groups should i choose? a comparative study of recommendation algorithms in flickr. *Journal of Information Science*, 36(6):733–750.

Zhou, K., Cummins, R., Lalmas, M., and Jose, J. M. (2012). Evaluating aggregated search pages. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 115–124, New York, NY, USA. ACM.

Zhou, X. S. and Huang, T. S. (2003). Relevance feedback in image retrieval: A comprehensive review. *Multimedia systems*, 8(6):536–544.

Zubiaga, A., Spina, D., Amigó, E., and Gonzalo, J. (2012). Towards real-time summarization of scheduled events from twitter streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 319–320. ACM.

Zuccon, G., Leelanupab, T., Whiting, S., Yilmaz, E., Jose, J. M., and Azzopardi, L. (2013). Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information retrieval*, 16(2):267–305.

Appendix A

Tag Recommendation Interfaces

Over the course of this PhD, two photo tag recommendation interfaces were created as a demonstrators for the LiMoSiNe project in order to showcase the research being undertaken. Specifically, in year 1 we created an interface which exploited *internal evidences* as proposed in Chapter 3 (i.e. contextual and visual classifiers) to aid the user in the photo tagging process by consider features such as (i) the time taken (ii) the number of faces *etc*. In year 3 we created an interface which exploited *external evidences*, as discussed in Chapter 4, in order to aid the user when annotating images taken at a large scale social event. The following sections briefly detail the methodology used to create these interfaces.

A.1 Internal Evidence Tagging Interface

In the internal evidence tagging interface the user is aided in the image annotation process with tags suggested based on those already added. As is described in Chapter 3, there exist many more image contexts which have not been explored to our knowledge for recommendation purposes. In this demo, the user is able to explore tag recommendations for a sample of 1,000 Flickr images from the FLICKR-COL collection as is described in the following sections.

A.1.1 Methodology

The opening page, shown in Figure A.1, shows a sample of the 1,000 images. The images can be reshuffled so that a new selection is shown using the *shuffle* link. If a user already knows an ID of an image they are interested in, they can enter this ID into the search box and click *GO*.

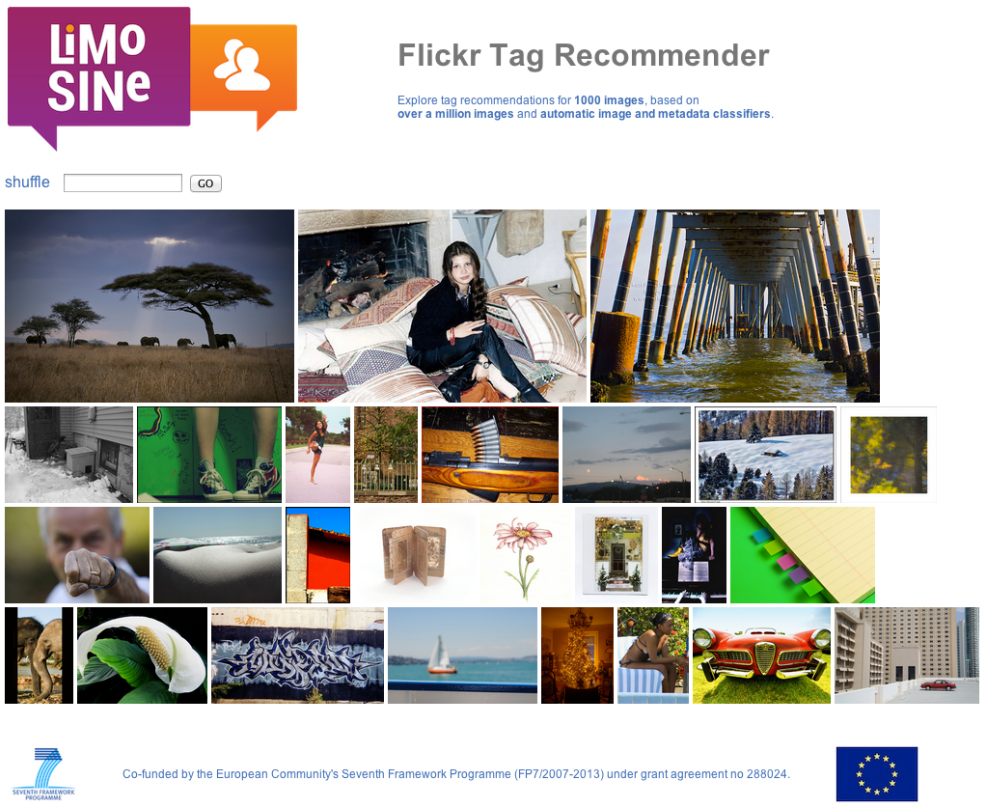

Fig. A.1 A screenshot of an opening page of the internal evidence tagging interface

Upon a selection of an image, the user sees a page such as the one shown in Figure A.2. Here, the image is displayed alongside tag recommendations and contextual/visual classifications. These contextual & visual classifications are as follows:

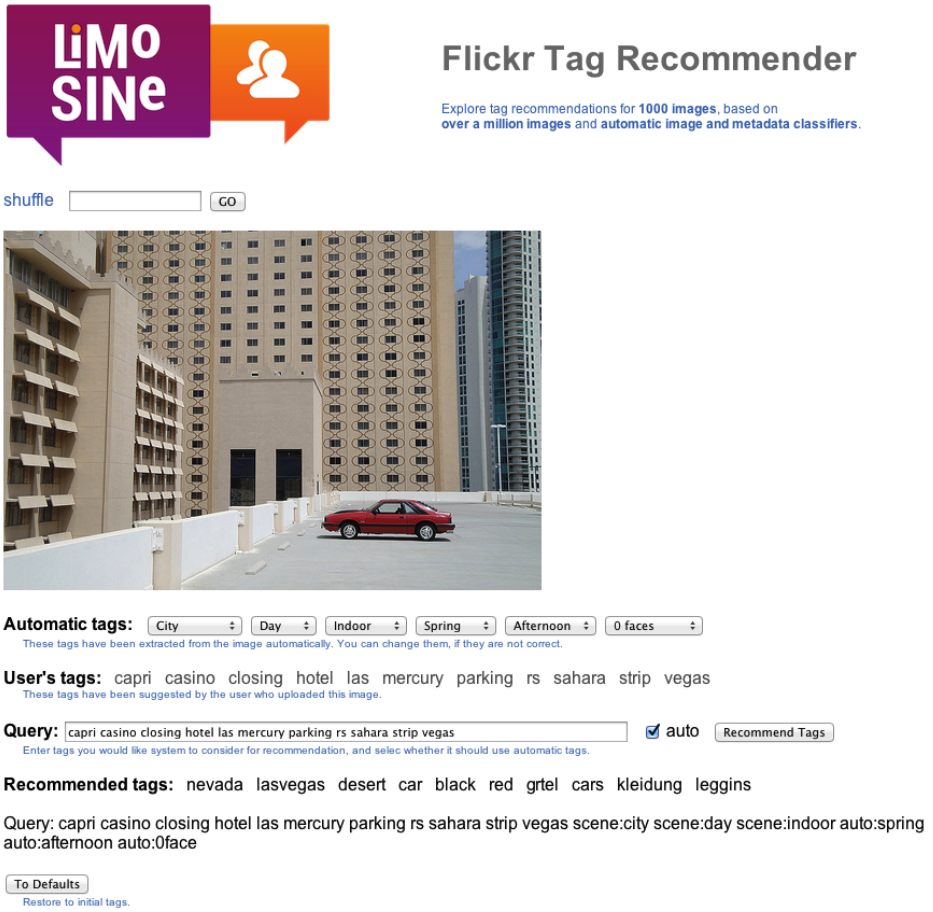

Fig. A.2 Image tag explorer

1. **Time of day**: images are classified as either taken in the *morning* (06:00 to 11:59), *afternoon* (12:00 to 17:59), *evening* (18:00 to 23:59) or *night* (00:00 to 05:59) based on their EXIF `time taken` attribute.

2. **Season**: images are classified as either taken in the *winter*, *spring*, *summer* or *autumn* based on the month from their EXIF `time taken` attribute.

3. **City/Landscape**: images are classified as *city* or *landscape* based on their visual appearance. We achieve this using popular city/landscape classification techniques (Vailaya et al., 1998) which trains a support vector machine (SVM) using the Edge Direction Coherence visual feature. Best classification accuracy (77.3%) is achieved using a Radial basis function (RBF) kernel where $C = 2^7$ and $\gamma = 2^{-3}$.

4. **Day/Night**: images are classified as being taken at *night* or during the *day* based on their visual appearance. We achieve this using popular day/night classification techniques (Wan and Hu, 2002) which train a SVM using the HSV histogram (Manjunath, 2002) visual feature. Best classification accuracy (88.3%)

is achieved using a linear kernel where $C = 2^{-1}$.

5. **Indoor/Outdoor**: images are also classified as being taken *indoors* or *outdoors* based on their visual appearance. We achieve this using popular indoor/outdoor classification techniques (Vailaya et al., 2001) by training a support vector machine (SVM) using the Colour Moments (Stricker and Orengo, 1995) visual feature. Best classification accuracy (71.1%) is achieved using an RBF kernel where $C = 2^5$ and $\gamma = 2^{-5}$.

6. **Number of Faces**: images are also classified based on the number of faces detected using a popular face detection method (Viola and Jones, 2001). Specifically, we classify images as having 0, 1 or 2+ faces.

Firstly, as observed in Figure A.2, both the *user tags* and the *automatically extracted tags* (i.e. contextual and visual classifications) are shown i.e. the user tags: `capri`, `casino` *etc*, and the automatically extracted tags: *City*, *Day etc*. The automatically identified tags are shown as drop down selections which can be modified if they are incorrectly classified, which in turn updates the tag recommendations to reflect this new evidence. The top ten tag suggestions are computed using our TF-IDF recommendation model using the methodology proposed in Chapter 3.

A.1.2 Architecture and Technical Specification

Figure A.3 describes the architecture used in the internal evidence demonstrator. As can be observed, there exists 3 major components:

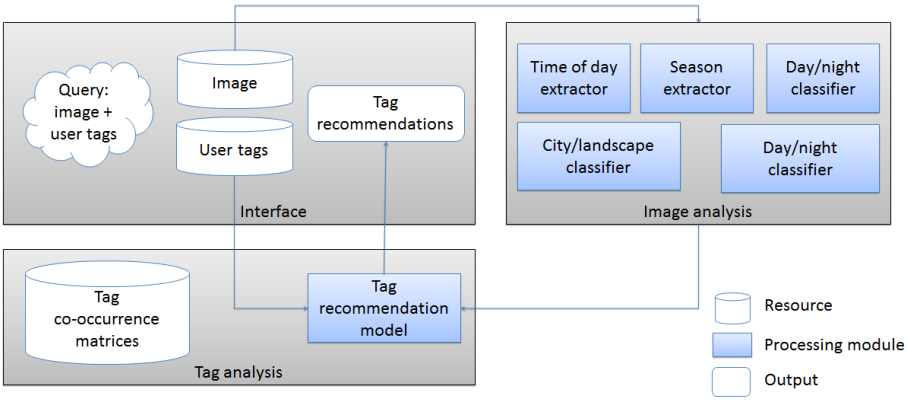

Fig. A.3 Internal evidence recommendation interface architecture

1. **Interface:** This demo uses standard web technologies and can be installed on most standard web servers. Specifically, the interface is built using standard HTML, CSS & Javascript web technologies with the processing built using PHP.

2. **Tag analysis:** given the tags annotated for an image, recommendations are computed using the TF-IDF strategy previously described. The co-occurrence vector for each tag is stored as a row in a MySQL database. Specifically, for each tag, six different co-occurrence vectors exist: one computed for all images in the training set, and six vectors computed for each of the six image classifiers.

3. **Image analysis:** the six image classifications are made offline using the techniques described in the previous section.

Fig. A.4 Opening page of the demo

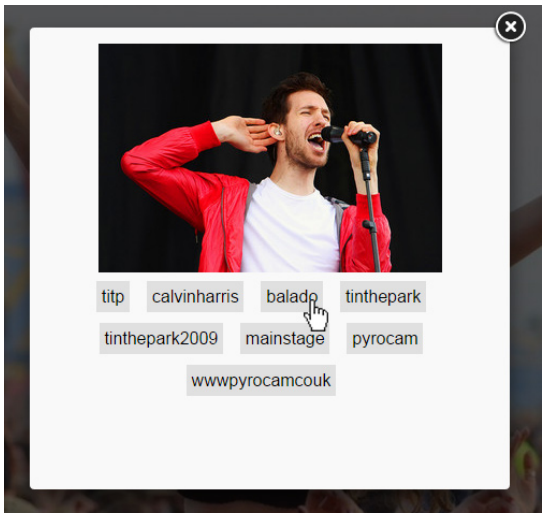

Fig. A.5 Adding tags in related images

A.2 External Evidence Tagging Interface

In the external evidence tagging interface the user is aided with various relevant textual sources for exploitation in the image annotation process. In particular we implement many of the techniques proposed in Chapter 4 by immersing the user in a wealth of textual content related to a predefined social event; we focus on exploiting social media streams and encyclopaedic resources for photo tag recommendation purposes. Social media streams, such as Twitter, offer distinct advantages over recommending tags based on historical image tagging trends, as discussed in Chapter 4; specifically, these resources are able to overcome the "time lag" issue associated with recommending based on historical content uploaded to image sharing websites. The exploitation of these sources is detailed in the following subsections.

A.2.1 Methodology

As can be observed in Figure A.4, initially the user is asked to upload a single image and the hashtag related to the social event it was taken at (e.g. a music festival). Based on this information, we present the user with an extensive image annotation interface, as seen in Figure A.6. In order to aid the user in the image annotation process, we offer suggestions from a number of sources as well as offer the user various tagging strategies as follows:

- **Flickr tag recommendation:** Firstly, we suggest tags based on historical image tagging trends. This is achieved by proposing tags which co-occur highly with the *hashtag* (with the hash symbol removed) in existing Flickr images. This method is described in Chapter 2.3.2 (TF-IDF). The recommendations made by this approach are shown in the red box in the top right of Figure A.6.

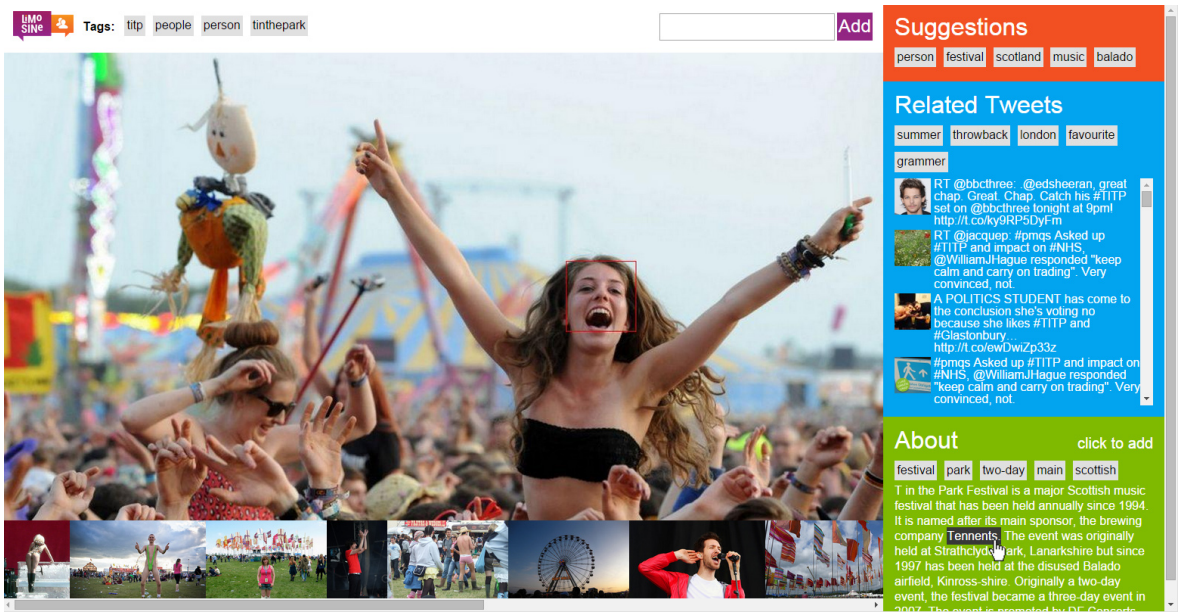

Fig. A.6 External evidence tagging interface

- **Twitter tag recommendation:** In order to offer suggestions from Twitter streams we first gather the top 1000 tweets (ranked chronologically) containing the given event hashtag (or description, removing whitespace). From this, we suggest those hashtags (removing the hash symbol) which occur most frequently in this tweet set to the user. The recommendations made by this approach are shown at the top of the blue box in Figure A.6. Aside from providing the user with Twitter based tag suggestions, the related Twitter stream is also presented to the user for which they can click any term within a tweet in order to use it to annotate their image.

- **Wikipedia tag recommendation:** We also suggest tags from the article related to the event; this approach assumes we are able to identify the relevant Wikipedia article for the event in question. We achieve this process automatically by querying the Wikipedia URL resolving API[1] against the relevant hashtag, as explained in Chapter 4. The text within a matched article is then processed, suggesting the most frequent *entities* (De Marneffe et al., 2006) as tag suggestions to the user. The recommendations made by this approach are shown at the top of the green box in Figure A.6. Aside from providing the user with Wikipedia based tag suggestions, the Wikipedia article "abstract" (i.e. first paragraph) is also presented to the user for which the user can click any term within the paragraph in order to use it to annotate their image.

- **Related image tag recommendation:** Images taken by other users at the same event are also likely to contain relevant tags; we therefore create an interface which allows the user to easily "steal" the annotations of these related images. This is achieved by searching the Flickr search API for photographs tagged with the given hashtag.

[1] https://www.mediawiki.org/wiki/API:Query#Resolving_redirects - accessed 10th July 2016.

These images are then ranked by descending view count and are presented to the user along the bottom of the screen. Once clicked, a popup revealing the selected image's annotations is displayed to the user (see Figure A.5) for which they can click in order to add any given tag to annotate their image.

- **Automatic image annotation:** Although not the focus of our work, the demonstrator also employs an industry strength automatic image annotation model[2] to first extract high level scenes/concepts (e.g. person, sport etc) from an image based on its visual appearance. These annotations are appended to the Flickr suggestions (in the red box for Figure A.6).

- **Face detection tag recommendation:** We also extract the number of faces from an image using an industry strength facial recognition model[3]. Based on this analysis, if at least one face is detected, the user is suggested with the following tags: `person`, `man`, `woman`, `girl`, `boy`. If more than one face is detected, the user is also suggested the tags `people` and `group`. These annotations are also appended to the Flickr suggestions (in the red box for Figure A.6).

- **Manual tag recommendation:** Finally, as offered by traditional tag recommendation models, the user is also able to add tags by manually entering them using the keyboard.

This interface allows the user to annotate images using their mouse predominately, reducing the need for manual annotation and the workload on the user in the image annotation process.

A.2.2 Architecture and Technical Specification

Figure A.7 describes the architecture used in the external evidence demonstrator. As can be observed, there exists 5 major components:

- **Interface:** the web interface, built using traditional web programming techniques (i.e. HTML, css, Javascript) allows the user to enter their hashtag and upload their image. Aside from interacting and providing feedback to the user, the interface also executes the necessary processing units in an AJAX asynchronous paradigm, meaning the page never refreshes.

- **Tag analysis:** given the hashtag, the relevant row from the tag co-occurrence matrix (stored as a sparse matrix in a MySQL database) is extracted for use in the TF-IDF model in order to suggest tags to the user.

[2]http://www.alchemyapi.com/products/features/image-tagging/
[3]http://rekognition.com/developer/face

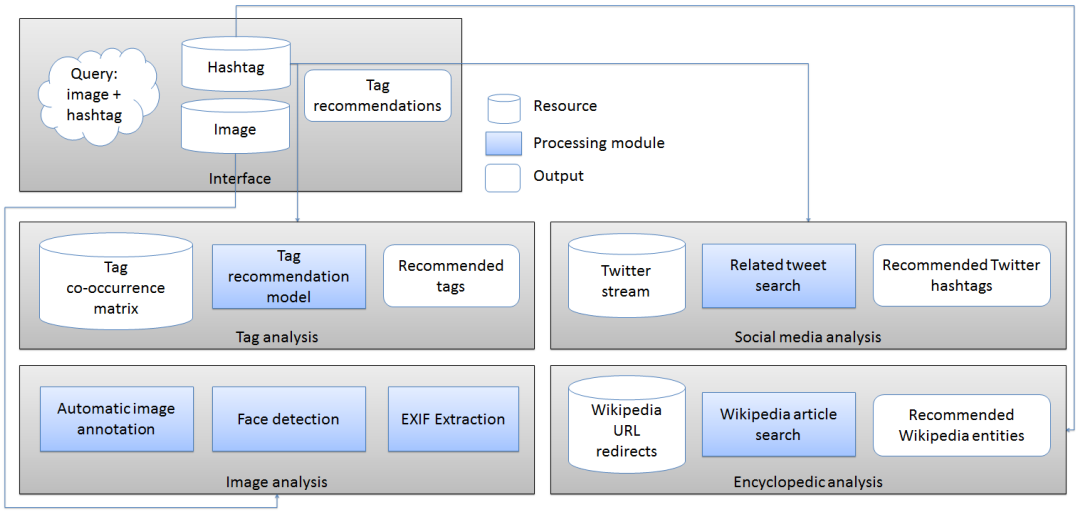

Fig. A.7 External evidence tagging interface architecture

- **Social media analysis:** given the hashtag, tweets are downloaded (using the Twitter API[4]) and processed "on the fly". This results in temporally significant tag suggestions for the user.
- **Encyclopaedic analysis:** as with tweets, the suggestions from Wikipedia are also computed on the fly, resulting in up to the date recommendations from this source. In order to identify the correct Wikipedia page article, the hashtag is queried against a MySQL database of Wikipedia URL redirects, as previously described.
- **Image analysis:** finally, the image analysis is implemented by calling APIs of two popular image annotation and face detection models as previously described. By doing so, if either model is improved, the demonstrator will also be improved.

[4]https://dev.twitter.com/