"Picture the scene..."; Visually Summarising Social Media Events

Philip J. McParlane*
School of Computing Science
University of Glasgow
Glasgow, UK
p.mcparlane.1@research.gla.ac.uk

Andrew J. McMinn
School of Computing Science
University of Glasgow
Glasgow, UK
a.mcminn.1@research.gla.ac.uk

Joemon M. Jose
School of Computing Science
University of Glasgow
Glasgow, UK
Joemon.Jose@glasgow.ac.uk

ABSTRACT

Due to the advent of social media and web 2.0, we are faced with a deluge of information; recently, research efforts have focused on filtering out noisy, irrelevant information items from social media streams and in particular have attempted to automatically identify and summarise *events*. However, due to the heterogeneous nature of such social media streams, these efforts have not reached fruition. In this paper, we investigate how images can be used as a source for summarising events. Existing approaches have considered only textual summaries which are often poorly written, in a different language and slow to digest. Alternatively, images are "worth 1,000 words" and are able to quickly & easily convey an idea or scene. Since images in social media can also be noisy, irrelevant & repetitive, we propose new techniques for their automatic *selection*, *ranking* and *presentation*. We evaluate our approach on a recently created social media event data set containing 365k tweets and 50 events, for which we extend by collecting 625k related images. By conducting two crowdsourced evaluations, we firstly show how our approach overcomes the problems of automatically collecting relevant and diverse images from noisy microblog data, before highlighting the advantages of multimedia summarisation over text based approaches.

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: Information Storage and Retrieval; I.4 [**Applications**]: Image Processing

Keywords

visual event summarisation; twitter; social media; image diversification; near duplicate detection

*This research was supported by the European Community's FP7 Programme under grant agreements nr 288024 (LiMoSINe)

CIKM '14, November 03–07 2014, Shanghai, China.

http://dx.doi.org/10.1145/2661829.2661923 .

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 [26, 39, 33]. With the wealth, coverage, speed, lack of censorship and unbias nature of microblog posts, they present many advantages over traditional journalistic media. In recent years, many works have attempted to textually summarise these events [36, 35, 1, 33], 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 propose to 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 (ii) naturally "multilingual" (iii) easier to digest (iv) & can show what words cannot explain. On the contrary, text is often: (i) slow and laborious to digest (ii) written in a different language (iii) & poorly written. We propose to use images in summaries to alleviate these problems.

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 for researchers 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 paper we also address this problem by considering which event types are most suitable for *visual* summarisation purposes.

In this work, we therefore attempt to integrate images in the summaries of events automatically detected on Twitter. We break the overall problem of visual event summarisation into 3 constituent parts: image *selection*, image *ranking* and summary *presentation*. Each component has its own inherit challenges which are addressed

[1]"An event is a significant thing that happens at some specific time and place" [26]

in Sections 4, 5 and 6. In particular, we attempt to address the following research questions (RQ):

RQ1. Can we use images to automatically summarise events detected on microblogging streams?

RQ2. 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?

RQ3. Are images present in summaries able to help the user to identify an event's topic and key entities (i.e. people, location *etc*)?

RQ4. Which event "type" (e.g. sports, politics) is best suited for visual summarisation?

The rest of this paper is as follows: Section 2 describes the related works in event detection and summarisation. Section 3 formulates the problem of visual event summarisation before further discussing image selection, ranking and presentation problems in Sections 4, 5 & 6, respectively. In Section 7 we describe our crowdsourced evaluation before discussing the results of this in Section 8. Finally, we conclude in Section 9 discussing the implications of this study and future work.

2. RELATED WORK

In the following section we detail the works in event detection and summarisation, discussing how they differ from our work.

2.1 Event Detection

Event detection has been a research focus since 1997 when the Topic Detection and Tracking (TDT) project [2] 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.* [35] proposed 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, Aggrawal *et al.* [1] used clustering and growth rate thresholding in order to detect events on Twitter. Finally, Sakaki *et al.* [33] attempted to detect tweets referencing natural disasters in order to issue early warning alerts. By using a simple technique of keyword filtering, 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 goal is to cluster related tweets into coherent events as they happen. In event *summarisation*, the objective is to *succinctly* describe these clusters of tweets in a way that covers as many of the five Ws (i.e. who, what, where, when, why) as possible. 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.

2.2 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. Chakrabarti *et al.* [4] 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.* [28] 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.* [41] 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.* [9] exploited microblog posts for the segmentation and summarisation of the 2008 US 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*) (iii) well defined hashtags. Therefore, the stream can be more easily parsed and understood with specific prior knowledge (e.g. when half time is *etc*). This paper attempts 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.* [36] developed a 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. We use this model in our work to compute textual summaries for baseline/experimental purposes. Marcus *et al.* [25] 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.* [23] also proposed a similar summarisation approach which attempted to cluster posts on the Sina microblogging website[2] by selecting topic words before building a graph based topic co-occurrence model for event tracking purposes. Summarisation is computed as the subset of posts which have the highest coverage, and therefore diversity, of the cluster topics.

Table 1 demonstrates the problems with using automatically generated textual summaries by comparing various summaries extracted using a state-of-the-art model [36], vs summaries written by a hu-

Table 1: Problems with SOTA [36] vs human summaries

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

[2] http://www.sina.com.cn/

man, when attempting to describe various events included in a recent large-scale social media event detection collection [26]. 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 this work we plan to alleviate these problems, as well as those already discussed, by introducing images, automatically collected 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 [26].

2.3 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. Fabro *et al.* [10] 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 *et al.* [32] 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.

In our work we instead consider the more challenging task of summarising *automatically detected events*, opposed to those hand selected, using images from *noisy* social media streams, opposed to querying the retrieval systems of popular image/video sharing websites. By selecting images from Twitter, instead of image sharing websites, we are able to create summaries which are more real time, as images uploaded to websites such as Flickr are often uploaded long after they are taken [27]. In the following section, we define the overall goal in the visual event summarisation task.

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.

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. Therefore, for small events, there are often insufficient images to use for visual summarisation.
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.
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*).

In the following sections, we discuss our methods and techniques for addressing these problems.

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; 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 [14, 5]; 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 in the HTML document for selection purposes. 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.

[3] `http://www.google.com/trends/`

[4] `http://commons.wikimedia.org/wiki/File:Standard_web_banner_ad_sizes.svg`

[5] `http://ogp.me/`

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, at a slightly different angle and time). 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 [37]. Hashing functions are used to generate fixed-length output strings which act as a shortened reference to its initial data. These functions were initially created for cryptographic purposes [31], however, in recent years, their application has been experimented for near-duplicate image detection. For example, Chum *et al.* [6] proposed two new image similarity measures using locality sensitive hashing (LSH). Foo *et al.* [17] compared the effectiveness of dynamic partial functions (DPF) and hash based counting techniques. In this work we used a related hashing method, called the Perceptual Hash (pHash), which has been shown to give high detection performance for resized, cropped and exposure compensated images [37]. We choose to detect using a hashing function due to its high performance while maintaining low computational expense in extraction and matching between images. 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 [37]), otherwise the image is added to a new cluster. Using this method, duplicate images with the following alterations are grouped together: (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.

4.3 Irrelevant Images

Aside from the problem of duplicate images, the major problem in the selection process is that of irrelevant images. Ranking based on popularity provides one such solution in order to reduce the number of irrelevant images selected (i.e. the wisdom of the crowd), however, there exist a wealth of images which are often popular (thus promoted in the rankings), but not suitable for event summarisation purposes. For example, in Figure 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 1 which refers to the 2014 conflict between Ukraine and Russia) but not suitable for event summarisation purposes.

2. **Screenshots:** there exist many screenshot images (e.g. (b) of Figure 1), which may contain tweets relevant to the event but are not useful for use in event summaries.

3. **Reaction images:** there are also a wealth of reaction images posted (e.g. (c) of Figure 1) which are used to evoke the user's emotion but are neither relevant or suitable.

[6]Using the tools available at `http://www.phash.org/`

(a) (b) (c)

Figure 1: Unsuitable images for event summarisation

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, to identify these images, we implement a synthetic image detection model introduced in [38]; the authors create a image classification collection[7] containing various *synthetic* and *natural* (i.e. real life photographs) images collected from Google image search. Of these images, we randomly select 600 synthetic and 600 natural photographs in order to train an Support Vector Machine (SVM) classifier. For each of these images, we extract the following high level features:

1. **Colour histogram (CH):** colour histograms are a popular method to describe the colour distribution of an image by "binning" the pixel frequency of each colour range. We compute this 64 bit histogram in both RGB and HSV colour spaces.

2. **Edge Histogram (EH):** the edge histogram captures the local edge distribution of an image. Edges are grouped into vertical, horizontal, 45 and 135 degree categories as detailed in [34].

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 in order 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 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 images.

Automatically detecting meme images (e.g. (a) of Figure 1) is more difficult, however, with the large diversity of image captions. Due to this, near-duplicate image detection methods fail. Instead we employ local feature matching which is able to handle changing image captions, whilst being able to effectively identify the background meme image. Specifically, we match all images in our collection against a crawled database of meme background pictures based on the popular Scale-invariant feature transform (SIFT) [24]. Originally introduced for matching between different views, angles and lighting conditions of an object or scene, we attempt to match the background image within a meme.

We first crawl a number of popular meme *background* images (i.e. containing no captions) from an online archive[8]; in total, 587 images are collected for which we extract SIFT keypoints. For each image in our collection, we also extract SIFT keypoints and attempt to match keypoints using an Affine Transform, against each of the 587 meme background pictures. We consider an image to be a meme, if it matches a certain percentage of its keypoints with a

[7]`http://wing.comp.nus.edu.sg/npic/`
[8]`https://imgflip.com/memetemplates`

meme background image. We select a 25% matching threshold due to previous work [16] 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.

4.4 Selection Systems

We define the following summarisation systems which are referenced in later sections of this paper:

1. **Filtered Website images (WEB):** we select only the *most popular* images from *websites* referenced in Tweets (based on occurrence) related to events. These images are selected using the techniques described in Section 4.1.

2. **Filtered Twitter images (TWR):** we select only the *most popular* images referenced in *Tweets* (based on re-tweets) related to events. These images are selected using the techniques described in Section 4.2 & 4.3.

5. IMAGE RANKING

In the following section we discuss methods of image ranking on a filtered set of images, I_e. The overall goal is to rank the most *relevant* images in the top positions whilst maintaining their *diversity*; showing *multiple* images which are of the same aspect of an event in an event summary is sub-optimal; the user should instead be presented with a visually diverse overview of images describing the event. In the following, we discuss various ranking methods which aim to achieve this goal.

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 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 relevance purposes [30, 12]. 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 computed 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 [3]. 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:

$$IDF(I) = log(|E|/|E_i|) \quad (1)$$

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 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.

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 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 [26], tweets are not only clustered into high level events, but are also clustered *within* their event. We therefore exploit these sub-clusters in order to maximise the diversity of ranked images. We achieve this by selecting the highest ranked image (using the $TF-IDF$ ranking model) in each of the largest K sub-clusters. By selecting the optimal image for various sub-clusters (i.e. moments), we attempt to maximise both *relevance* and *diversity* in the rankings.

5.3 Ranking Systems

We define the following ranking systems which are referenced in later sections of this paper:

1. **Normalised Popular Twitter Images (P-TWR):** images collected from *Twitter* are ranked in descending order using the $TF-IDF$ weighting scheme described in Section 5.1.

2. **Normalised Popular Website Images (P-WEB):** images collected from *websites* are ranked in descending order using the $TF-IDF$ weighting scheme described in Section 5.1.

[9] https://twitter.com/TheEllenShow/status/440322224407314432

[10] http://www.funnyordie.com/lists/5c274d5d70/best-examples-of-the-ellen-degeneres-oscar-selfie-meme

3. **Combined Normalised Images (P-COM):** we rank images based on a combination of both 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.

4. **Semantically Clustered Twitter Images (S-TWR):** we select the highest ranked *Twitter* image (using the $TF-IDF$ weighting scheme) in each of the largest semantic clusters, as described in Section 5.2.

5. **Semantically Clustered Website Images (S-WEB):** we select the highest ranked *website* image (using the $TF-IDF$ weighting scheme) in each of the largest semantic clusters, as described in Section 5.2.

6. **Combined Semantically Clustered Images (S-COM):** we combine the S-TWR and S-WEB rankings using the same merging scheme as used in P-COM.

6. EVENT SUMMARY PRESENTATION

For event summarisation presentation, the overall goal is to describe an event in the most succinctly 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 [36, 25, 23] 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 [22] and maps [40] in previous work where there exists a similar information overload problem as in event summarisation. In this work, we therefore employ tag wordclouds 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.

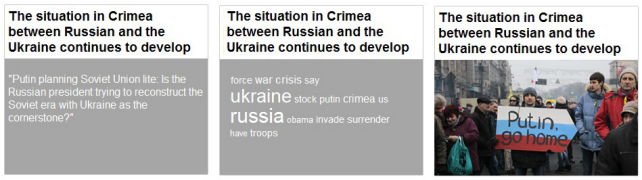
6.1 Presentation Systems

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

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 in [36]) and the most re-tweeted tweet describing the event (see (a) of Figure 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 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 second summarisation *baseline*.

3. **Title and image (TIM):** in our experimental approach, we present the same single sentence title (as used in TTW) as well as the top ranked image, computed using the S-COM ranking strategy, in the summarisation interface (see (c) of Figure 2 for an example).



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

Figure 2: Event Summarisation Presentation Experiment

For each interface, we aim to reduce bias by keeping as many of the design aspects as consistent as possible e.g. font sizes, colours, size *etc*. By doing so, we are able to evaluate the influence and value of using images in event summaries.

7. EXPERIMENTS

In the following sections, we discuss details of the collection used in this study 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 before describing the evaluation procedure and metrics used.

7.1 Collection

In this work, we use the collection introduced in [26] which was built with event detection evaluation in mind. This collection contains details of over 500 automatically detected events using state-of-the-art techniques. In our work, we consider the 50 largest (in terms of tweets) events for visual summarisation. We hypothesise that visual event summarisation is most effective for the largest events where there exists sufficient images. The collection used is as follows:

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 in [26]. The event detection method clusters on highly filtered tweets which contain at least a single *entity* (identified by an extension of the Stanford Parser [21]). Each event contains on average 7,280 tweets which are further grouped into 135.6 sub-clusters, on average.

We extend this collection[11] by extracting images from websites referenced within tweets. Table 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.

7.2 Building a Test Set

In order to determine the selection & ranking effectiveness of the various systems, described in Sections 4.4 and 5.3, we must create

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

	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 2: Images collected from tweets and URLs in tweets

a test set where relevant images are determined for each event. We achieve this by asking users in a crowdsourced experiment.

Crowd-sourcing (i.e. outsourcing a task to a network of online workers) experiments have grown in popularity in recent years [18, 13] and have been adopted to carry out tasks which are often difficult for computers but easy for humans e.g. image classification [29, 11]. Recently, Nowak *et al.* [29] showed that by using a majority voting scheme for an image annotation task, the quality of Turker judgements were in-line with those made by experts. The ImageNet collection was also built using a crowd-sourced experiment where internet images were mapped to WordNet nodes [11].

For this work, we use the paid crowdsourcing service CrowdFlower[12] (CF) for evaluation purposes. CrowdFlower provides a number of advantages over platforms such as Amazon Mechanical Turk (MTurk)[13] 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 [15, 19].

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 each selected image in the top 5 ranks of each approach, we ask workers a number of questions to create this ground truth. If a worker accepts our experiment, or Human Integellence 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 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:

[12] http://www.crowdflower.com
[13] https://www.mturk.com/

High quality photograph (a) | Average quality photograph (b) | Low quality photograph (c) | Computer generated image (d)

Figure 3: Image categories for Q2 in our crowdsourced experiment

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

They are then asked to complete a number of test questions in order to judge their understanding of the task. One of the major problems with crowdsourcing is that workers often spam or try to complete tasks with as little effort as possible in order to maximize their profits [20]. This can lead to poor quality submissions. Test questions (i.e. those with "known answers") are used to identify spamming workers. Specifically, we ask users 8 test questions, in which they must answer *all* correctly, before they are allowed to continue to the HIT. By doing so, we ensure 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 must have atleast a "Level 1 Contributor[14]" 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 workers accepted our HIT, of which 114 (57.6%) passed the 8 test questions making it to the evaluation stage, with 84 (42.4%) failing. 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 an annotator agreement score of 0.75. Worker agreement, computed by CrowdFlower, "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 an annotator agreement of 0.87.

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

7.3 Metrics

To evaluate our image selection and ranking methods, we use the following traditional information metrics, 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 4.2, we also use intent-aware metrics proposed in text based diversification. 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, queries are often ambiguous or multi-faceted [8]. For example, *Java* is an ambiguous query since it has different interpretations e.g. programming language, the island, and the coffee. Further, *java programming language* is a multi-faceted query since it has several aspects, e.g. books, tutorials *etc*. Ambiguous queries are an issue for search engines and were originally not addressed by retrieval algorithms. In these approaches, intent-aware metrics (or diversification metrics) are used to compute the coverage, novelty and relevance of rankings; in our work, we use diversification metrics to discount those systems which promote duplicate images in their top positions. Therefore, by "grouping" duplicate images into clusters (using the technique described in Section 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 [7] 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 subtopics [7].

7.4 Evaluating Summary Presentations

In order to compare the effectiveness of our event summary presentations, discussed in Section 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 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 [13] 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[15]). 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.

8. RESULTS

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

8.1 Image Selection & Ranking

Table 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 result 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.

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 RQ2. In particular, S-WEB improves significantly (+55% for P@5) 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

[15]"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."

Table 3: Comparison of image selection & ranking approaches. Bold 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.

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****	0.8333	**0.621***	0.570	0.543
S-COM	**0.500***	0.500**	0.813	0.614	**0.616***	**0.578***

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 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: In Table 5 we compare 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", "Arts & Entertainment" and "Sports", we can infer that there exist more relevant images online documenting these types of events and therefore they are most suitable for visual event summarisation purposes, thus addressing RQ4. On the contrary, "Law & Politics" events are more likely to happen "behind closed doors" and therefore more difficult to capture for amateur photographers.

Table 4: Comparison of image quality in the top 5 ranks for each system. Bold values indicate the highest value for each criteria.

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

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

8.2 Event Summary Presentation

For the survey results gathered from users judging the 3 event summary interfaces, as detailed in Section 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 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 RQ3. 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.

9. CONCLUSION

In this work we proposed the automatic visual summarisation of events detected on noisy social media streams. In order to achieve our goal, we proposed image selection & ranking methods which aimed to automatically collect and order the most relevant images related to events. In order to do so, we overcame issues such as: (i) irrelevant images, (ii) data sparsity, (iii) dealing with duplicate image content, (iv) and dealing with spam. In a crowdsourced evaluation we demonstrated the advantages of describing events using images in comparison to 2 text based baselines.

The main results of our work showed that images help the user to understand the content, people and locations present in events and that they can increase user interest and engagement. Further, we demonstrated that by combining filtered images from multiple sources (i.e Twitter and related URLs) and selecting the most pop-

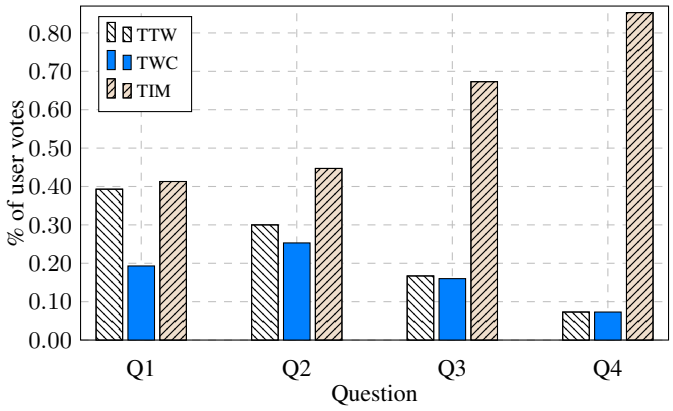


Figure 4: Survey results for event summary presentation evaluation

ular images from the largest semantic clusters, we were able to automatically select *relevant* and *diverse* images for summarisation purposes. Finally, the findings of 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.

In future work, we plan to exploit new evidences for image ranking/selection purposes (e.g. filenames, explicit/implicit user feedback) as well as explore more elaborate summarisation interfaces, such as using more than one image. In addition to this we plan to consider which kinds of photograph attract users as well as compare against images collected from image sharing websites (e.g. Flickr, Instagram *etc*).

10. REFERENCES

[1] C. C. Aggarwal and K. Subbian. Event detection in social streams. *In SIAM DM, 2012.*

[2] J. Allan, V. Lavrenko, and H. Jin. First story detection in TDT is hard. *In ACM CIKM, 2000.*

[3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. *In CEAS, volume 6, 2010.*

[4] D. Chakrabarti and K. Punera. Event summarization using tweets. *In AAAI ICWSM*, 2011.

[5] C. H. Chang, M. Kayed, M. Girgis, and K. Shaalan. A survey of web information extraction systems. *In IEEE KDE, 2006.*

[6] O. Chum, J. Philbin, and A. Zisserman. Near duplicate image detection: min-hash and tf-idf weighting. *In BMVC, 2008.*

[7] C. Clarke, N. Craswell, and I. Soboroff. Preliminary report on the trec 2009 web track. *In TREC notebook, 2009.*

[8] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. *In ACM SIGIR, 2008.*

[9] L. K. D. A. Shamma and E. Churchill. Statler: Summarizing media through short-message services. *In ACM CSCW, 2010.*

[10] M. Del Fabro and L. Böszörmenyi. Summarization and presentation of real-life events using community-contributed content. *In ACM MM, 2012.*

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *In IEEE CVPR, 2009.*

[12] Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum. An empirical study on learning to rank of tweets. *In COLING, 2010.*

[13] C. Eickhoff and A. P. Vries. Increasing cheat robustness of crowdsourcing tasks. *In Inf. Retr., volume 16, 2013.*

[14] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open information extraction from the web. *In Commun. ACM, volume 51, 2008.*

[15] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating named entities in twitter data with crowdsourcing. *In NAACL HLT, 2010.*

[16] J. J. Foo and R. Sinha. Pruning sift for scalable near-duplicate image matching. *In ACM ADC, 2007.*

[17] J. J. Foo, J. Zobel, R. Sinha, and S. M. M. Tahaghoghi. Detection of near-duplicate images for web search. *In ACM CIVR, 2007.*

[18] M. Hirth, T. Hoßfeld, and P. Tran-Gia. Cheat-detection mechanisms for crowdsourcing. *In Technical Report 474, University of Würzburg, 2010.*

[19] J. Hong and C. F. Baker. How good is the crowd at "real" wsd? *In ACL LAW V, 2011.*

[20] A. P. D. V. J. Vuurens and C. Eickhoff. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. *In ACM SIGIR Workshop on Crowdsourcing for Information Retrieval, 2011.*

[21] D. Klein and C. D. Manning. Accurate unlexicalized parsing. *In Annual Meeting on ACL, 2003.*

[22] B. Y.-L. Kuo, T. Hentrich, B. M. . Good, and M. D. Wilkinson. Tag clouds for summarizing web search results. *In ACM WWW, 2007.*

[23] R. Long, H. Wang, Y. Chen, O. Jin, and Y. Yu. Towards effective event detection, tracking and summarization on microblog data. *In WAIM, 2011.*

[24] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *In IJCV, volume 60, 2004.*

[25] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. *In ACM SIGCHI, 2011.*

[26] A. J. McMinn, Y. Moshfeghi, and J. M. Jose. Building a large-scale corpus for evaluating event detection on twitter. *In ACM CIKM, 2013.*

[27] P. J. McParlane and J. Jose. Exploiting twitter and wikipedia for the annotation of event images. *In ACM SIGIR, 2014.*

[28] J. Nichols, J. Mahmud, and C. Drews. Summarizing sporting events using twitter. *In ACM IUI, 2012.*

[29] S. Nowak and S. Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. *In ACM MIR, 2010.*

[30] S. Ravikumar, R. Balakrishnan, and S. Kambhampati. Ranking tweets considering trust and relevance. *In ACM IIWeb, 2012.*

[31] R. Rivest. The md5 message-digest algorithm. 1992.

[32] M. Sahuguet and B. Huet. Socially motivated multimedia topic timeline summarization. *In ACM SAM, 2013.*

[33] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. *In ACM WWW, 2010.*

[34] P. Salembier and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc., 2002.

[35] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. *In ACM SIGSPATIAL, 2009.*

[36] B. P. Sharifi, D. I. Inouye, and J. K. Kalita. Summarization of twitter microblogs. *In BCS Computer Journal, 2013.*

[37] Z. Tang, Y. Dai, and X. Zhang. Perceptual hashing for color images using invariant moments. *In Appl. Math, volume 6, 2012.*

[38] F. Wang and M. yen Kan. Npic: Hierarchical synthetic image classification using image search and generic features. *In ACM CIVR, 2006.*

[39] J. Weng and B.-S. Lee. Event detection in twitter. *In AAAI ICWSM, 2011.*

[40] J. Wood, J. Dykes, A. Slingsby, and K. Clarke. Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. *In IEEE TVCG, 2007.*

[41] A. Zubiaga, D. Spina, E. Amigó, and J. Gonzalo. Towards real-time summarization of scheduled events from twitter streams. *In ACM HT, 2012.*