Future Internet **2014**, *6*, 242-260; doi:10.3390/fi6020242

OPEN ACCESS

future internet

ISSN 1999-5903

www.mdpi.com/journal/futureinternet

Article

Exploiting Multimedia in Creating and Analysing Multimedia Web Archives

Jonathon S. Hare *, David P. Dupplaw, Paul H. Lewis, Wendy Hall and Kirk Martinez

Web and Internet Science Research Group, University of Southampton, Highfield, Southampton SO17 1PR, UK; E-Mails: dpd@ecs.soton.ac.uk (D.P.D.); phl@ecs.soton.ac.uk (P.H.L.); wh@ecs.soton.ac.uk (W.H.); km@ecs.soton.ac.uk (K.M.)

* Author to whom correspondence should be addressed; E-Mail: jsh2@ecs.soton.ac.uk; Tel.: +44-23-8059-7678; Fax: +44-23-8059-2783.

Received: 25 February 2014; in revised form: 31 March 2014 / Accepted: 16 April 2014 / Published: 24 April 2014

Abstract: The data contained on the web and the social web are inherently multimedia and consist of a mixture of textual, visual and audio modalities. Community memories embodied on the web and social web contain a rich mixture of data from these modalities. In many ways, the web is the greatest resource ever created by human-kind. However, due to the dynamic and distributed nature of the web, its content changes, appears and disappears on a daily basis. Web archiving provides a way of capturing snapshots of (parts of) the web for preservation and future analysis. This paper provides an overview of techniques we have developed within the context of the EU funded ARCOMEM (ARchiving COmmunity MEMories) project to allow multimedia web content to be leveraged during the archival process and for post-archival analysis. Through a set of use cases, we explore several practical applications of multimedia analytics within the realm of web archiving, web archive analysis and multimedia data on the web in general.

Keywords: multimedia; archiving; analysis

1. Introduction

Community memories embodied on the web and within social media are inherently multimedia in nature and contain data and information in many different auditory, visual and textual forms. From

the perspective of web archiving and digital preservation, taking these modalities into account is very important. As we move towards more intelligent archiving and more intelligent post-crawl analysis of the data hidden within digital archives, the ability to leverage the vast array of different data modalities becomes even more crucial.

Adaptive web crawlers, which target specific keywords, are now becoming more commonplace. These crawlers make fast, near-real time, decisions about what (textual) content should be crawled based on the presence or absence of keywords given at the start of the crawl. Multimedia analysis is typically very computationally expensive, especially compared to simple text analysis. This presents many challenges when working with multimedia web archives, especially if we want to integrate multimedia analysis into the decision making that happens inside an adaptive crawl. Within the EU funded ARCOMEM (ARchiving COmmunity MEMories) project [1], we have developed adaptive web and social-web crawlers that produce highly focussed web archives relating to specific entities (people, places, and organisations), topics, opinions and events (ETOEs) [2]. In the ARCOMEM crawlers, ETOEs are primarily detected using state-of-the-art text analysis techniques, although, as we will see later in the paper, ETOEs can also be detected in multimedia content. To deal with the problem of computational complexity in the ARCOMEM crawlers, the real-time decision making required for dynamically controlling the crawl is performed by fast text analysis techniques, which determine whether the outlinks of a document should be followed based on the relevance of the document to a crawl specification (itself consisting of sets of ETOEs). Once a crawl is completed (or potentially, whilst it is still running), a second set of more-complex processing is applied to the crawled documents to select a subset for export, whilst, at the same time, producing vast amounts of metadata and analysis results, which can be exported alongside it. A more complete overview of the ARCOMEM system is given in Section 3.

This paper builds upon our work presented at the 1st International Workshop on Archiving Community Memories [3], where we discussed many of facets of multimedia analysis with respect to web archiving. Within this paper, we discuss some of the needs and opportunities for both analysing the multimedia data on the web and within an archive of community memories. Practical embodiments of the discussed techniques developed within the ARCOMEM project are detailed, with full references to our published works in which these techniques are evaluated. We also discuss how multimedia data might be exploited during the initial creation of the archive, given sufficient computational resources to analyse the data at a high enough rate.

The structure of this paper is as follows: Section 2 gives an overview of some key recent trends, approaches and techniques in multimedia analysis. Section 3 describes the approach to web archiving and automated archive analysis developed in the ARCOMEM project. Section 4 describes a number of use cases for applying multimedia analysis to web and social web data, that can either be used for post-crawl analysis or intelligent document appraisal within a web crawl. Finally, we conclude with some remarks about the future outlook of our research on multimedia analysis with respect to web crawling and web analytics.

2. Current Trends in Multimedia Analysis

Multimedia analysis is a vibrant and challenging research area, and there are many aspects to its use in archiving or cultural heritage. Analysis of multimedia forms a continuum that extends from the extraction and matching of low-level explicit features, such as colour and shape in the case of images, to high-level semantics, such as object identification or more subjective sentiment-based semantics. Much of what has been achieved so far resides at the lower-level end of this continuum, because the explicit features are relatively easy to extract. One key challenge in multimedia analysis is to "bridge the gap" between low-level features and high-level semantics with reliable analysis methodologies [4–6].

Many of the higher-level analysis techniques that exist are specific to particular domains; for example, human face detection and recognition in the case of still images and video. More generalised higher-level descriptions can be sought from multimedia through auto-annotation techniques that rely on heuristics or artificial intelligence to find similarities between the lower-level features extracted from data sets labelled with higher-level ground-truth semantics. Within the ARCOMEM project, most of our work on multimedia analysis focused on the analysis of visual content and the combined analysis of visual content together with text and metadata. For this reason, it makes sense to briefly give an overview of popular and state-of-the-art techniques for visual analysis.

Low-level visual feature extraction and representation is a well-studied problem; effective solutions to it include the definition of robust local features, such as SIFT [7], colour variants of SIFT [8] and scalable methodologies for creating bag-of-words and pyramidal representations using such features [9,11]. In terms of higher-level interpretation, particularly the detection of visual concept and event labels, existing solutions typically rely on SIFT-like features and computationally expensive machine learning algorithms, such as kernel Support Vector Machines (SVMs), where one or more such classifiers are trained independently for each concept or event that we want to detect [8]. The scale of the problem dictates that the training of each SVM is performed independently. However, this also means that the numerical output of the different classifiers is often not directly comparable, making these classifiers suitable for Google-like retrieval of content according to isolated labels, but sub-optimal for selecting the most appropriate concepts for annotating a specific image or video.

Online web repositories that have user-annotated images (such as Flickr [10]) give access to ground-truth data sets providing large amounts of knowledge which, in turn, indicates the overall consensus of a community for image representations of specific concepts. This is ideal for training machine-learning algorithms [12,13], which can be be applied to images with no (or only partial) user provided annotations. By combining semantic taxonomies with feature-based approaches, it has been shown that retrieval results can be either ordered on an axis of diversity [14,15] or ordered such that diversity is maximised within the most relevant results [16–18].

Sentiment analysis of multimedia is a fledgling research area, and current research has investigated two topics in particular. Firstly, there has been work in the area of automatic facial expression recognition [19–21]. Secondly, there has been some work on associating low-level image features with emotions and sentiments [22,23]. The automated detection of facial expression is still a long way from being robustly applicable to unconstrained real-world images, where even the task of reliably detecting a face is not completely solved. The idea of associating low-level visual features with sentiments has

shown some promise, although until recently, it has only been demonstrated on small datasets with relatively simple colour- [24] and texture- [25] based features. Our more recent work on image sentiment analysis has started to explore larger datasets and looks at multimodal analysis, where image features are used in combination with available metadata and text [26].

3. The ARCOMEM Approach

Within the ARCOMEM project, together with our project partners, we have built a system for crawling and analysing samples of web and social-web data at scale. Fundamentally, the software has four goals: Firstly, it provides a way to intelligently harvest data from the web and social web around specific entities, topics and events. Secondly, it provides a scalable and extensible platform for analysing harvested datasets and includes modules for state-of-the-art multimodal multimedia (textual, visual and audio) content analysis. Thirdly, it exposes the results of the analysis in the form of a knowledge base that is interlinked with standard linked-data resources and accessible using standard semantic technologies. Finally, the software provides the ability to export the harvested and analysed dataset in standardised formats for preservation and exchange. In terms of scalability, the ARCOMEM software is designed to work with small (tens/hundreds of gigabytes) to medium (multi-terabyte) web datasets.

3.1. Intelligently Harvesting and Sampling the Web

ARCOMEM has developed web crawlers specifically designed for harvesting multimedia data from both standard web pages (the large scale crawler) as well as social media sources via their APIs (the API crawler). In addition, the web crawler is able to crawl the deep web through a set of modules that know how to extract information from certain kinds of web sites. Both crawlers run in a distributed fashion across a cluster of machines and store the raw content in an Apache HBase column-oriented database [27].

The crawlers can operate in a number of different ways, depending on the requirements of the user, who defines what they want in their dataset. There are three common crawling strategies that can be used individually or together:

- Standard crawling. A standard web crawl starts with a seed list of URLs, and crawls outwards from the seed pages by following the outlinks. Constraints might be added to limit the number of hops the crawler is allowed to make from a seed page or to limit the crawler to specific Internet domains or IP addresses.

- API-directed crawling. In an API-directed crawl, the user provides keywords that describe the domain of the dataset they want to create. These keywords are then fed to the search APIs of common social media sources (e.g., Twitter, Facebook, YouTube, *etc.*), and the returned posts are examined for outlinks that are then used as seeds for a web crawl.

- Intelligent crawling. In an intelligent crawl, the user provides a detailed intelligent crawl specification (ICS) consisting of keywords, topics, events and entities that describe the target domain. A standard crawl and/or API-directed crawl is then started, and as new resources

are harvested, they are scored against the ICS (using pattern matching and machine learning techniques). Scores for the outlinks of each resource are then created (combining the resource score with specific scores computed based on the link), and these scores are fed back to the large scale crawler, which prioritises the next URL to crawl based on the score. URLs with low scores will not be crawled.

3.2. Advanced, Scalable Multimodal Multimedia Content Analysis

Once the data has been harvested, it must be cleaned, processed and analysed in order to allow the end users to explore and query the data. The ARCOMEM system is designed such that most processing occurs as scalable, distributed map-reduce tasks using Hadoop [28] performed over the HBase database. The final output of these processes is stored as Resource Description Framework (RDF) triples in a knowledge base, which is queryable through the SPARQL query language. Interfaces are built on top to allow the end users to explore the data set without the need to know SPARQL.

Useful data on the web is often mixed with lots of irrelevant content, such as adverts. In terms of the archiving of a web page, it is often the case that the archivists want the full content, including adverts, to be preserved to retain the original context surrounding the information on the page. However for automatic analysis and indexing of the archive material, this irrelevant information can get in the way. Within the ARCOMEM system, there are modules (based on BoilerPipe [29] and Readability4J [30] that can attempt to automatically clean the content of a web page to remove the extraneous noisy information before the content analysis modules are invoked.

The content analysis modules, provided by GATE (the General Architecture for Text Engineering) [31,32] and OpenIMAJ (the Open toolkit for Intelligent Multimedia Analysis in Java) [33,34], are primarily based around the detection of entities, topics, opinions and events (ETOEs) in both textual and visual media. Additional information, such as textual terms (words with high term frequency-inverse document frequency values) and near duplicates of images and videos are also recorded. In visual media, specifically, face (and object) detection and recognition techniques are used to detect entities. Visual opinion analysis takes two forms: facial expression analysis and global sentiment/attractiveness/privacy classification [12]. High-level semantic enrichment is used to disambiguate entities and semantically link them to concepts in standard external knowledge bases, such as DBpedia (DBPedia is a machine readable database of knowledge extracted from structured information within the Wikipedia encyclopedia).

Data on the web is in constant flux, with many pages and items often changing or being updated. To deal with this, the ARCOMEM system architecture is specifically designed to allow resources to be crawled multiple times if required. This allows changes over time to be detected and analysed.

3.3. Interoperability, Reusability and Provenance

The web is very dynamic and constantly changing, so it is unlikely that any dataset harvested by a web observatory could ever be re-collected from scratch and end up the same. That being said, it is important that any given dataset has associated provenance that describes the exact strategy that was used to create

it. In ARCOMEM, this is achieved by storing the crawl specification and system configuration in the knowledge base along with information about exactly what was harvested and when.

ARCOMEM allows data to be exported in two forms that can be used together. The raw resource data can be exported as standard ISO 28500 WARC (Web Archive) files [35] and can be used with standard tools for handling WARC files, such as the Wayback Machine (which allows the dataset to be visually reconstructed and explored). The results of the processing and analysis can be exported directly in RDF. Whilst the ARCOMEM system uses its own ontology for describing the analysis, for interoperability, the ontology is provided with mappings to several other standard ontologies.

4. Use Cases for Multimedia Analysis in Archiving Community Memories

There are numerous applications for multimedia analysis within the community memory archiving and analysis scenario, but they fall into four broad categories:

- Guiding the crawl by identifying relevant documents.
- Reducing the size of the archive by removing irrelevant or duplicated content.
- Generating metadata for facilitating searching strategies within the archive.
- Summarising various aspects of the content of the archive (*i.e.*, finding the events, people and places represented by the content of the archive).

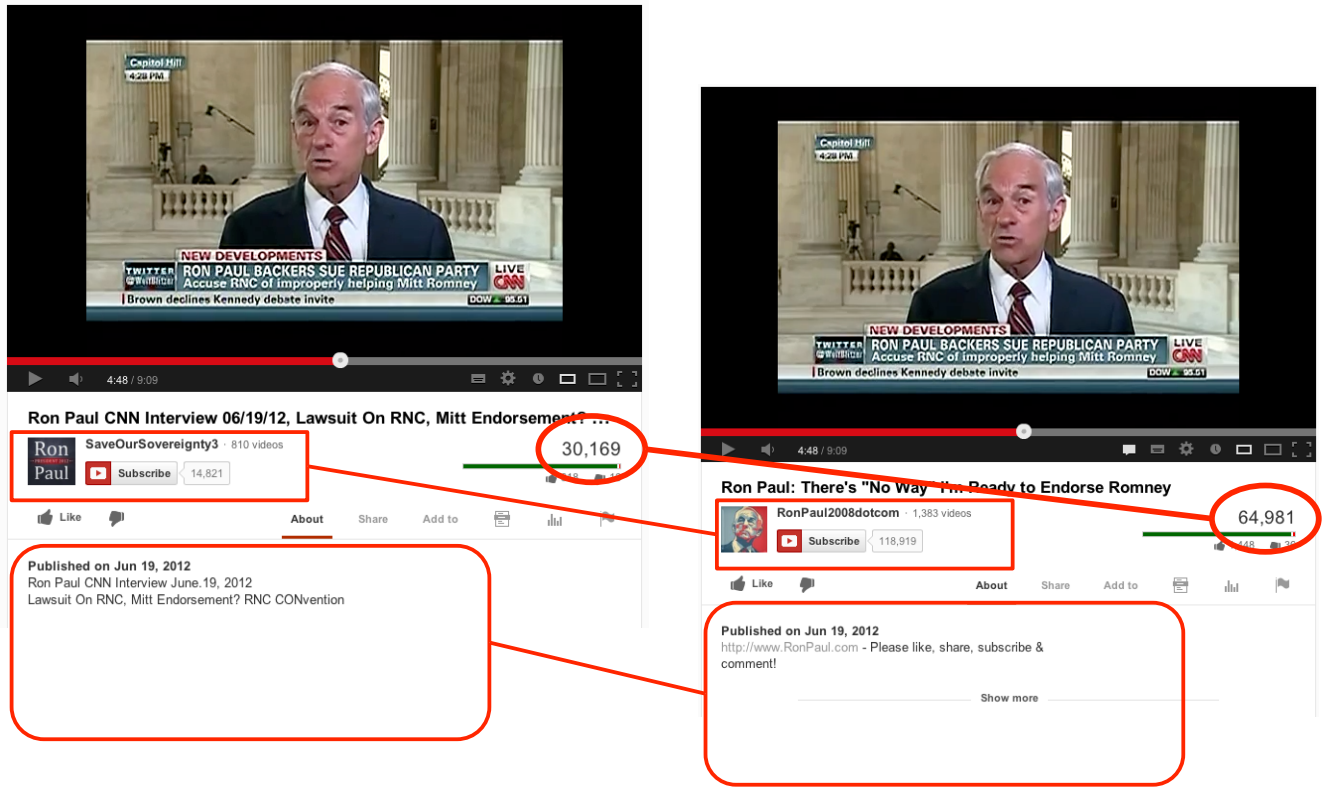
This section of the paper discusses how multimedia analytics has potential applications in each of the categories. These are illustrated through a number of use cases developed within the ARCOMEM project together with a summary of the research and results of the research to support these use cases performed within the project.

4.1. Aggregating Social Commentary

Media is often duplicated on the web. For example, on YouTube, it is common to find multiple versions of the same video, often with minor changes (*i.e.*, different sound tracks or modified visual appearance; it is common for the original broadcaster's logo to be changed in a bootleg copy on YouTube). On Twitter, the same effect can be observed, and it is common for the same/similar media object to be shared and re-posted (at a different URL) many times [36] (see Section 4.2 below). In both of these examples, the communities sharing and providing commentary on the individual instances of a particular media item often do not overlap.

From the perspective of archiving community memories, it is important to find and capture the relationships between these shared objects, as this allows the social context and commentary to be aggregated, building a much fuller and potentially more diverse picture of the community associated with the media object. An example of this is shown in Figure 1, which shows two copies of the same video on YouTube; both of these videos have been watched and commented on by different subsets of the YouTube community, but to draw any balanced conclusions about what people think about what was being said in the video, one must look at both sets of commentary.

Figure 1. Illustration of duplicate videos on YouTube, with the non-overlapping social commentary from different subsets of the user community highlighted. Fully automated video analysis techniques are able to detect these duplicates, and by recognising that the social commentaries are referring to the same content, they can be aggregated.

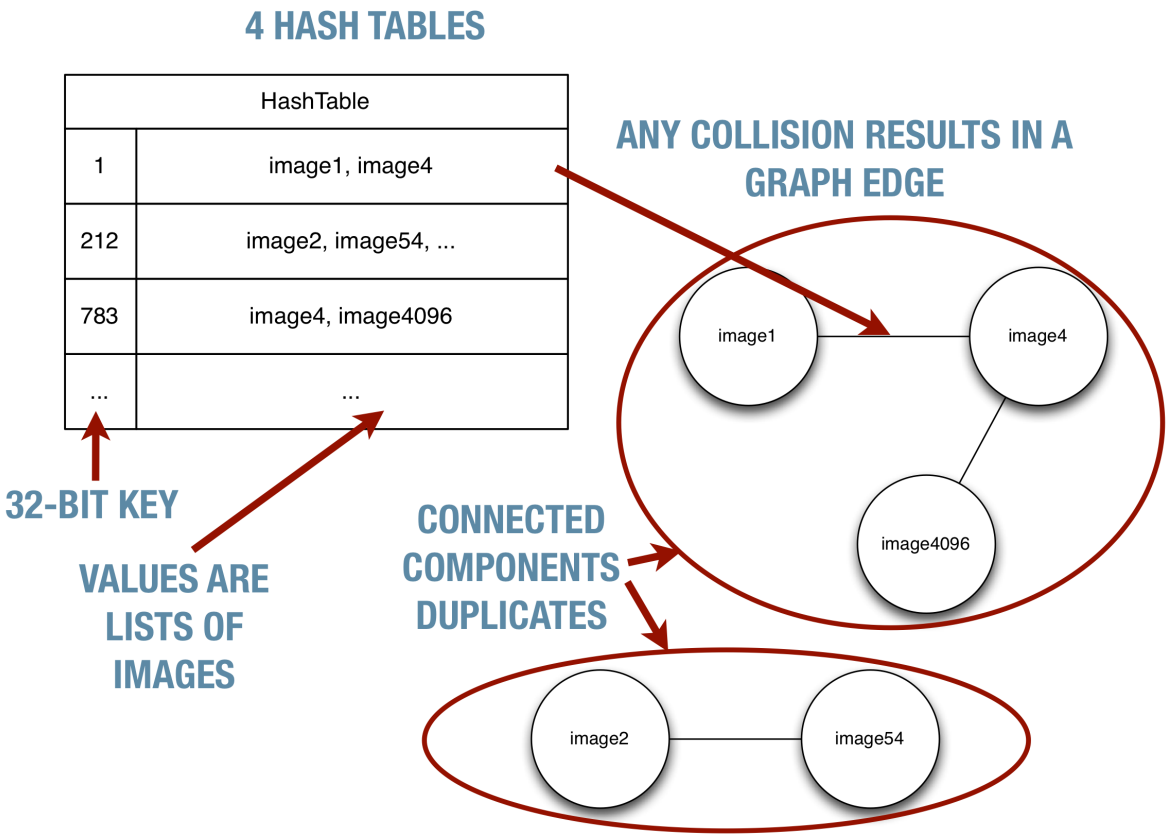


From a practical perspective, it can be very difficult or even impossible to determine whether two media items are duplicated by looking at the metadata alone. Often, retweets use the same URL; however, sharing across social networks (e.g., Twitter to Facebook or YouTube to Facebook) or redirection through a different URL shortener can mean that the URLs of identical media are different. Some networks/sites will resize media as they are shared or uploaded, so it is not even possible to use a checksum-like measure of similarity. However, modern image analysis and indexing techniques allow visual duplicates to be detected efficiently through the use of content-based image feature indexing [36,37]. For example, in [36], we extracted SIFT features [7] to represent the image content. As SIFT features are scale-invariant, they are already able to match across images that have been resized, and they are also robust to a number of other transformations.

Comparing images by directly matching their SIFT features is computationally inefficient and impractical with large image collections. To circumvent this problem and efficiently assess whether features match, locality sensitive hashing [38] (LSH) is used to create sketches (compact binary strings) from the features. The sketches are produced, such that the Hamming distance between sketches approximates the Euclidean distance between the features [39]. Rather than explicitly computing Hamming distances between all features, an efficient, approximate scheme is used to find matching features [40]. In this scheme, 128-bit sketches are partitioned into four 32-bit integers, which are used as keys in four hash tables. The corresponding values are the images in which a particular sub-sketch of a feature appeared. In our approach, to detect all near-duplicate images (or video frames), we build an

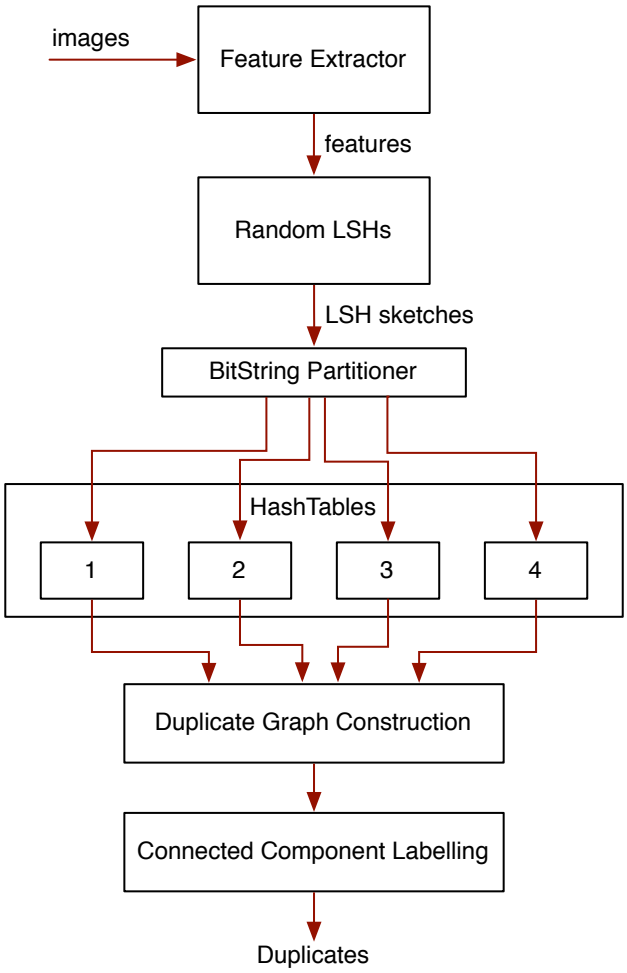
undirected weighted graph, where the vertices are the images and the edge weights represent the number of matching features between the images. As illustrated in Figure 2, the approximate matching scheme, which results in four sets of hash tables, makes the graph building process become trivial [36]. Once the graph is constructed, edges with low weight are pruned, and connected-component analysis is performed to extract all the sets of near-duplicates. It should be noted that the number of features associated with an image varies with the content of the image. This means images with more features are more likely to have more connected edges; however, in practice, this does not seem to be much of a problem. In our experiments, edges with a weight of less than 10 were pruned (e.g., images had to share at least 10 features to be classed as a match). We have begun to explore other more advanced methods for cutting the duplicate graph without the need for a fixed threshold; however, the current technique gives good results on the ARCOMEM test crawls and is also very efficient.

Figure 2. Constructing a duplicate graph from the four hash tables created from sketches of local features.



The entire duplicate detection process is illustrated in Figure 3. Near-duplicate media can then be linked in the archive (also linking the associated commentaries and contexts), or better still, the best (highest resolution, most complete or, in some cases, the smallest) version of the media can be stored alongside the commentary from all of the duplicates. This allows the overall size of the archive to be reduced without losing information, as it is often unnecessary to store multiple copies of content that is identical.

Figure 3. Process for efficiently finding near-duplicates of images (or video frames) by constructing a duplicate graph from the sketches of local features. LSH, locality sensitive hashing.



4.2. Measuring the Temporal Pulse of Social Multimedia

The detection of trends within streams of communication in a social network is often necessary for deciding the important topics or turning points within an archive. Although most analysis is performed in an offline, batch manner, it is possible to perform trend analysis on a real-time stream during the crawling process. Traditional approaches to trend and topic detection are based on the analysis of textual content within the stream; however, it is possible to use the near-duplicate detection techniques described above in Section 4.1 for finding bursts of sharing activity for a single non-text media item, thereby finding trending multimedia. Using the SIFT-LSH near-duplicate detection approach, described above, on a temporally shifting window of images, it is possible to find those images that are being shared within a particular time-period (say the last hour). These can be marked in the archival process as being important to crawl and archive, and the times at which they were deemed important can also be stored as metadata. Our work on finding trending images on Twitter in real-time [36] (a video demo is available at [41]) implements this approach and also has subsequently been adapted to work on archived Twitter data. A screenshot of the visualisation of the trending images is shown in Figure 4.

Figure 4. Visualising trending images on Twitter. The image in the centre was the most popular image at the time of capture; on the right-hand side are some of the instances of this image. The small piles of images on the left correspond to other images that are currently trending (*i.e.*, they appear multiple times in the time-window); these can be clicked on to expand them. The text of Tweets related to the currently selected trend are shown at the bottom.

4.3. Recognising Social Events in Social Media Streams

Combining items from social media streams, such as Flickr photos and Twitter tweets, into meaningful groups can help users to contextualise and effectively consume the torrents of information now made available on the social web. From the point-of-view of a web archive, this grouping of data makes it much more amenable to later browsing and analysis. Performing this grouping can also help separate irrelevant information from relevant information in a guided crawl. The task of performing such a grouping is made challenging because of the scale of the streams and the inherently multimodal nature of the information being contextualised. The problem of grouping social media items into meaningful groups can be seen as an ill-posed and application-specific unsupervised clustering problem. A fundamental question in multimodal contexts is determining which features best signify that two items should belong to the same grouping.

Recently, we have been developing a methodology that approaches social event detection as a multi-modal clustering task. The various challenges of this task are addressed by our approach: the selection of the features used to compare items to one another; the construction of a single sparse affinity matrix; combining the features; the relative importance of features; and clustering techniques that produce meaningful item groups, whilst scaling to cluster very large numbers of items. Our state-of-the-art approach developed in the ARCOMEM project was evaluated using the dataset defined by the 2013 MediaEval (MediaEval is an international multimedia analysis benchmarking initiative [42]) Social Event Detection task [43]. With a near-optimal configuration, we achieved an F_1 score of 0.94, showing that a good compromise between precision and recall of clusters can be achieved. At the time of writing, this is the highest reported score on this dataset. Full details on our approach can be found in [44].

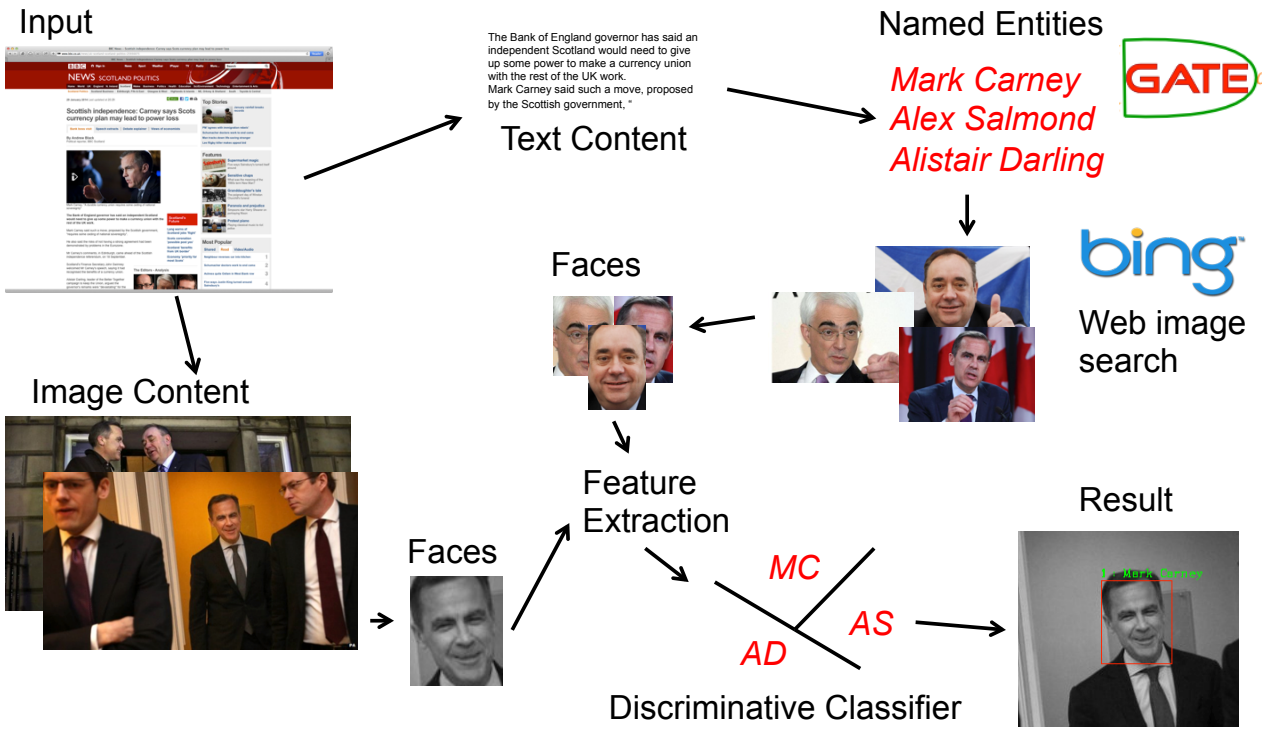
4.4. Detecting Media about Individuals, Organisations and Places

Modern web archiving techniques, such as those being created by the ARCOMEM project, are often centred around the main entities (people, places and organisations) that are often integral to the crawl topic. By ensuring that the archived documents mention relevant entities, the quality and accuracy of the archive can be controlled. Storing this metadata alongside the raw archive also provides enhanced browsing and analytic opportunities when the archive is reused. Analysis of non-text media to determine the entity depiction can further enhance the archive's metadata and can be used to help select relevant multimedia documents by looking for the presence of specific entities in a visual sense.

4.4.1. Recognising People

For people entities, recent advancements in face recognition means that people can be recognised with relatively high accuracy from within a small search space (*i.e.*, a small set of people to choose from). The problem with a general archiving scenario is that the search space is effectively infinite, and current face recognition algorithms tend to deteriorate rapidly as the search space gets larger. One option that we have explored in ARCOMEM is to apply entity recognition to the text to extract mentions of people and, then, to use the specific set of person entities to constrain the face recogniser's search space. In ARCOMEM, another option for identifying relevant people for constraining the face recogniser is to just use the people that are explicitly mentioned in the ICS. For well-known personalities and people whose photos can be found on the Internet, a web-based image search can be used to automatically retrieve example images of those people from which a face recognition algorithm can be trained [45]. An illustration of the overall process used in ARCOMEM is shown in Figure 5.

Figure 5. Illustration of how face recognition is used to detect entities by dynamically training the recogniser based on person entities detected in the textual content of the crawl.



4.4.2. Recognising Organisations

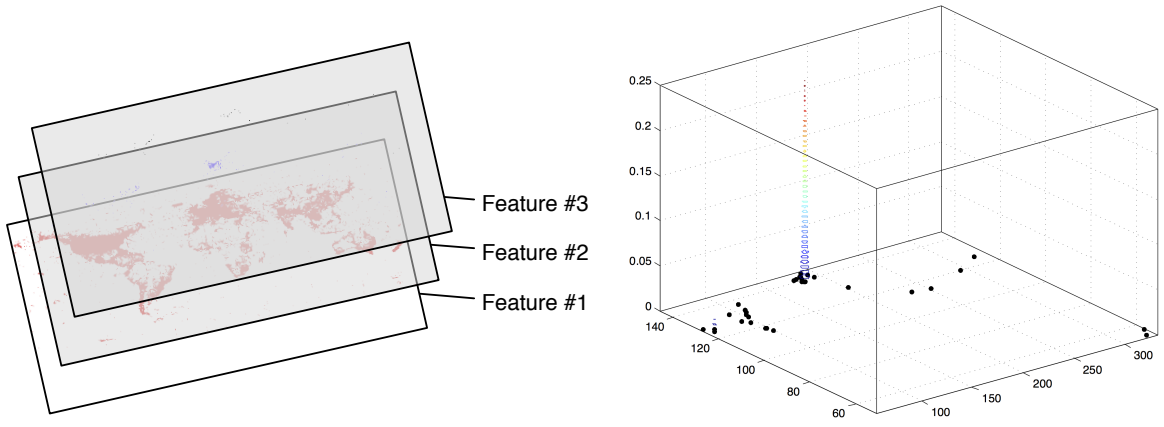
Organisations can be detected by looking for the presence of their corporate logo in images. There are many different ways in which logo recognition can be achieved, but variations on the SIFT keypoint matching technique described above are popular [46], and techniques, like topological matching of SIFT interest points [47], can improve performance. If example logo images cannot be provided at the campaign outset, the technique of finding logo exemplars can be automated using the same technique as for finding people exemplars by utilising web search engines to find images from which classifiers can be trained.

4.4.3. Recognising Places

Finally, place detection is an emerging research area. Given an image that has no location information, as long as the depicted location is visually unique, then it may be resolved by matching against a large corpus of geo-located images [37,48]. If metadata is available, it can be used together with visual information to improve the location estimation.

As part of our participation in the MediaEval 2013 placing task [49], we developed an algorithm that estimates a continuous probability density function (pdf) over the surface of the Earth from a number of features extracted from the query image and, if available, its metadata [48,50]. From the pdf, we find the modes of the distribution and choose the position of the most likely mode as the estimated location. The key ideas of this approach are illustrated in Figure 6.

Figure 6. Our approach to the geographical location of images. (**a**) A number of features are extracted from the content and metadata of a query image; each feature provides a set of latitude-longitude points associated with that feature (learned from the geotagged training images); (**b**) the location of predominant mode of the density distribution of all the features is determined using the standard mean-shift algorithm; this gives an estimate of the location, as well as the probability of it being correct (based on the likelihood of the mode).



(**a**) (**b**)

Experiments with the dataset of 8.6 million Flickr photos from the MediaEval placing task indicate that we are able to achieve state-of-the-art performance with our model and predict the location of almost 25% of the images to within 1 km of there true location. If we exploit additional information about the

users taking the photos (by utilising their Flickr profiles and other photos that they have taken that have know locations), this rises to over 45%. It should be noted that it is predominantly the meta-information that is at work here; the best runs using only visual information at MediaEval achieved about 2.5% accuracy within 1 km of the true location. The reason for this is related to the fact that the majority of images on Flickr do not actually display any visibly recognisable places.

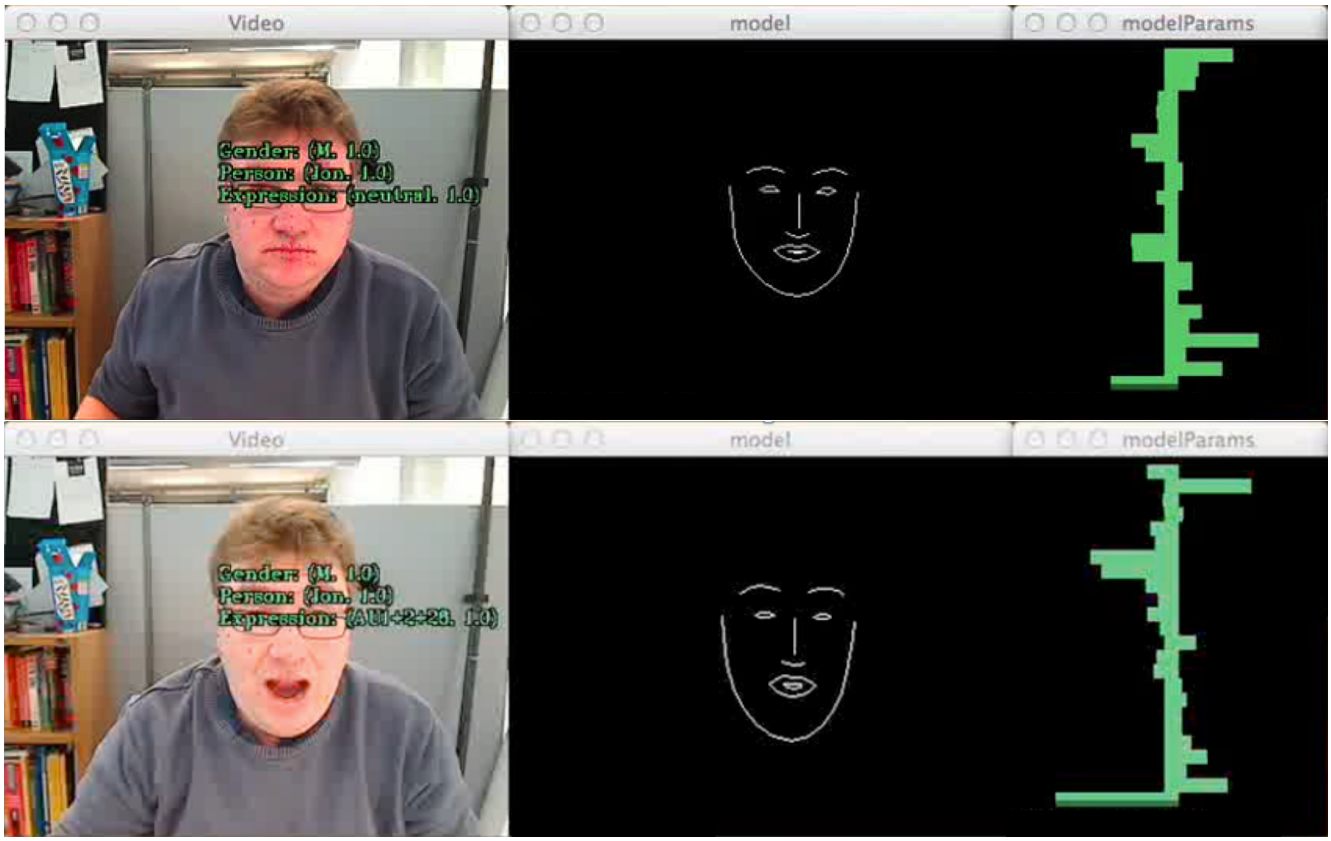
4.5. Measuring Opinion and Sentiment

The analysis of opinions and sentiment within a collection of community memories is often an important requirement of the end users of the archive. Specific examples include sentiment-based search (e.g., "find me positive media about X") and the determination of the key influential media objects (see the previous use cases of determining trends and events).

Images are often used to illustrate the opinions expressed by the text of a particular article. By themselves, images also have the ability to convey and elicit opinions, emotions and sentiments. In order to investigate how images are used in the opinion formation process, we have been developing tools that: (1) allow the reuse of images within an archive to be explored with respect to diverse time and opinion axes; and (2) allow in-depth analysis of specific elements (in particular, the presence and expression of human faces) within an image to be used to quantify opinion and sentiment.

Active shape models [51] (ASMs) and active appearance models [52] (AAMs) are well-known algorithms for fitting a shape to an image using the image's gradients to choose the best position for the vertices of the shape. As these models are parametric and generative (they can be used to reconstruct the shape from a small number of parameters), a large range of poses, expressions and appearances (e.g., skin textures) can be generated. Fitting a model to an image is a constrained optimisation problem in which the parameters of the model are iteratively updated in order to minimise the difference between the generated model and the image. Once a model is fitted to an image, the parameters can then be used as input to an expression classifier that can determine an expression label for the face. This model fits well with the Facial Action Coding System (FACS) [53], which aims to provide a standardised way of describing the expressions of faces. Codes represent muscular actions in the face (such as "inner eyebrow raising" or "lip corner puller"), and when combined, they represent emotions (for example, activation of the lip corner puller AU6 (*'Action Unit number 6'*) and the cheek raiser AU12 actions imply happiness). These muscular movements map to combinations of parameters in the face model, so a classifier can be trained on the model to recognise these actions. Of course, this relies on accurate face model fitting. Unfortunately, it is difficult to build a model that will accurately fit all faces and poses, which is essential for the accurate measurement of the shape parameters needed for expression classification. Another problem is that accurate detection of a face is required to initialise the fitting of a face model; whilst face detection techniques are quite mature, they can still have major problems working on real-world images, where the faces are not exactly frontal to the camera or there are shadows or contrast issues. Our research in ARCOMEM has shown that there is some promise in using these kinds of approaches, but they are not yet mature enough to be applied to most real-world imagery. A screenshot of our experimental system working under laboratory conditions is shown in Figure 7.

Figure 7. Screenshots of our experimental expression recognition system. The left window shows the input from a video camera, with live predictions of the name and gender of the subject(s), together with predictions of facial action units. The middle image shows the reconstructed shape model, and the third window shows the shape parameters of the model as a bar chart.



In less constrained visual media, we cannot rely on there being faces in the images, and sentiment may be carried by other visual traits. Indeed, images may intrinsically have sentiment associated with them through design (for example, a poster for a horror film) or through association with a specific subject matter, which may be context sensitive (say a photo of wind generators in the context of climate change). For these situations, there are no specific algorithms that we can use for extracting the sentiment, and whilst predictions of the opinions and sentiment of visual content can be made by considering the visual content alone, a richer approach is to consider both the image and the context in which it appears. State-of-the-art research on the sentiment analysis of images [25,26,54] has already begun to explore how the analysis of textual content and the analysis of visual content can complement each other. In particular, in one of our experiments, the results of sentiment classification of images based on the sentiment scores of Flickr tags was combined with a sentiment classifier that was based purely on image features. The performance of the multimodal fusing of the classifiers outperformed either classifier alone [26]. Recently, work has also been exploring aspects of attractiveness [55] and privacy [12], both of which could aid sentiment classification. In the case of privacy classification, sentiment could be inferred by seeing if a very private photo is being shown in a public place.

5. Conclusions and Outlook

This paper has discussed some of the opportunities created by combining modern multimedia analysis techniques with tools for crawling and performing analysis on archives of community memories. The paper has also described how some of these techniques are being made available within the tools developed by the ARCOMEM project.

Looking ahead to the future, there are a number of research areas in which the use of multimedia analysis could be expanded with respect to the archiving of community memories. Two specific areas for further exploration are outlined below:

1. Image-entity-guided crawling. At the moment, within the ARCOMEM tools, the visual entity tools are applied to the archive after it has been created. However, visual entities could potentially be used to directly influence the crawl process; for example, if a crawl was specified to look at topics surrounding the Olympic games, then the crawl specification could contain images of the Olympic rings logo, and the visual analytics tools could be used to detect the presence of this logo in images. Pages with embedded images with the logo and the outlinks of these pages could then be given higher priority by the crawler.

2. Image-entity co-reference resolution in multilingual corpora. Image content is inevitably reused across different documents; often, the image will have been scaled or cropped as it is used in different documents. Our tools for detecting this kind of reuse are now quite robust and are capable of providing coreference resolution of the images and the entities they depict. This has many practical uses; for example, it could be used to link documents in different languages as being related, even though we may not have natural language processing tools for the languages in question. In turn, this coreference information could be used to help guide the crawler to new relevant content.

Acknowledgments

This work was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 270239 (ARCOMEM). The authors would also like to thank the project partners within the ARCOMEM consortium, without whom this work would not have been possible.

Author Contributions

All authors worked extensively together to conceive the work the work presented in this manuscript and were involved its preparation. J.S.H. and D.P.D. were jointly responsible for the implementation of the various techniques and their integration with the ARCOMEM platform. J.S.H. and P.H.L. were responsible for the evaluation of the techniques. W.H., K.M. and P.H.L. supervised the project.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. ARCOMEM: Archiving Community Memories. Available online: http://www.arcomem.eu/ (accessed on 21 April 2014).
2. Tahmasebi, N.; Demartini, G.; Dupplaw, D.; Hare, J.; Ioannou, E.; Jaimes, A.; Lewis, P.; Maynard, D.; Peters, W.; Risse, T.; *et al.* Models and Architecture Definition/Contribution to Models and Architecture Definition. ARCOMEM Deliverable D3.1/D4.1. Available online: http://www.arcomem.eu/wp-content/uploads/2012/05/D3_1.pdf (accessed on 21 April 2014).
3. Hare, J.S.; Dupplaw, D.; Hall, W.; Lewis, P.; Martinez, K. The Role of Multimedia in Archiving Community Memories. In Proceedings of the 1st International Workshop on Archiving Community Memories, Lisbon, PT, USA, 6 September 2013.
4. Enser, P.G.B.; Sandom, C.J.; Hare, J.S.; Lewis, P.H. Facing the reality of semantic image retrieval. *J. Doc.* **2007**, *63*, 465–481.
5. Hare, J.S.; Sinclair, P.A.S.; Lewis, P.H.; Martinez, K.; Enser, P.G.; Sandom, C.J. Bridging the Semantic Gap in Multimedia Information Retrieval: Top-down and Bottom-up Approaches. In Proceedings of the 3rd European Semantic Web Conference, Budva, Montenegro, 12 June 2006.
6. Smeulders, A.W.M.; Worring, M.; Santini, S.; Gupta, A.; Jain, R. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intel.* **2000**, *22*, 1349–1380.
7. Lowe, D. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110.
8. Van de Sande, K.E.A.; Gevers, T.; Snoek, C.G.M. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intel.* **2010**, *32*, 1582–1596.
9. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; Volume 2, pp. 2169–2178.
10. Flickr Photo Sharing. Available online: http://www.flickr.com/ (accessed on 21 April 2014).
11. Hare, J.; Samangooei, S.; Lewis, P. Efficient Clustering and Quantisation of SIFT Features: Exploiting Characteristics of the SIFT Descriptor and Interest Region Detectors under Image Inversion. In Proceedings of the 1st ACM International Conference on Multimedia Retrieval, Trento, Italy, 17–20 April 2011.
12. Zerr, S.; Siersdorfer, S.; Hare, J.; Demidova, E. Privacy-Aware Image Classification and Search. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, OR, USA, 12–16 August 2012; pp. 35–44.
13. Huiskes, M.J.; Lew, M.S. The MIR Flickr Retrieval Evaluation. In Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval, Vancouver, BC, Canada, 26–31 October 2008.
14. Hare, J.; Lewis, P.H. Explicit diversification of image search. In Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, Dallas, TX, USA, 16–19 April 2013; pp. 295–296.

15. Zontone, P.; Boato, G.; Natale, F.G.B.D.; Rosa, A.D.; Barni, M.; Piva, A.; Hare, J.; Dupplaw, D.; Lewis, P. Image Diversity Analysis: Context, Opinion and Bias. In Proceedings of the First International Workshop on Living Web: Making Web Diversity a True Asset, Collocated with the 8th International Semantic Web Conference, Washington, DC, USA, 25–29 October 2009.
16. Agrawal, R.; Gollapudi, S.; Halverson, A.; Ieong, S. Diversifying Search Results. In Proceedings of the Second ACM International Conference on Web Search and Data Mining, Barcelona, Spain, 9–12 February 2009; pp. 5–14.
17. Ionescu, B.; Menéndez, M.; Müller, H.; Popescu, A. Retrieving Diverse Social Images at MediaEval 2013: Objectives, Dataset and Evaluation. In Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, 18–19 October 2013.
18. Jain, N.; Hare, J.; Samangooei, S.; Preston, J.; Davies, J.; Dupplaw, D.; Lewis, P.H. Experiments in Diversifying Flickr Result Sets. In Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, 18–19 October 2013.
19. Fasel, B.; Luettin, J. Automatic facial expression analysis: A survey. *Pattern Recognit.* **2003**, *36*, 259–275.
20. Tian, Y.l.; Kanade, T.; Cohn, J.F. Facial Expression Analysis. In *Handbook of Face Recognition*; Springer: New York, NY, USA, 2005; pp. 247–275.
21. Pantic, M.; Sebe, N.; Cohn, J.F.; Huang, T. Affective Multimodal Human-Computer Interaction. In Proceedings of the 13th annual ACM international conference on Multimedia, Singapore, 6–11 November 2005; pp. 669–676.
22. Wang, W.; He, Q. A Survey on Emotional Semantic Image Retrieval. In Proceedings of the International Conference on Image Processing, San Diego, CA, USA, 12–15 October 2008; pp. 117–120.
23. Zontone, P.; Boato, G.; Hare, J.; Lewis, P.; Siersdorfer, S.; Minack, E. Image and Collateral Text in Support of Auto-annotation and Sentiment Analysis. In Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, Uppsala, Sweden, 16 July 2010; pp. 88–92.
24. Wang, W.; Yu, Y.; Jiang, S. Image Retrieval by Emotional Semantics: A Study of Emotional Space and Feature Extraction. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Taipei, Taiwan, 8–11 October 2006; Volume 4, pp. 3534–3539.
25. Yanulevskaya, V.; van Gemert, J.C.; Roth, K.; Herbold, A.K.; Sebe, N.; Geusebroek, J.M. Emotional Valence Categorization Using Holistic Image Features. In Proceedings of the IEEE International Conference on Image Processing, San Diego, CA, USA, 12–15 October 2008; pp. 101–104.
26. Siersdorfer, S.; Hare, J.; Minack, E.; Deng, F. Analyzing and Predicting Sentiment of Images on the Social Web. In Proceedings of the International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 715–718.
27. Apache HBase Home. Available online: http://hbase.apache.org (accessed on 21 April 2014).
28. Apache Hadoop Home. Available online: http://hadoop.apache.org (accessed on 21 April 2014).

29. Kohlschütter, C.; Fankhauser, P.; Nejdl, W. Boilerplate Detection Using Shallow Text Features. In Proceedings of the Third ACM International Conference on Web Search and Data Mining, New York, NY, USA, 3–6 February 2010; pp. 441–450.
30. Hare, J.; Matthews, M.; Dupplaw, D.; Samangooei, S. Readability4J—Automated Webpage Information Extraction Engine. http://www.openimaj.org/openimaj-web/readability4j/ (accessed on 21 April 2014).
31. Cunningham, H.; Maynard, D.; Bontcheva, K.; Tablan, V.; Aswani, N.; Roberts, I.; Gorrell, G.; Funk, A.; Roberts, A.; Damljanovic, D.; *et al.*. *Text Processing with GATE, Version 6*; Gateway Press: Louisville, KY, USA, 2011.
32. GATE. General Architecture for Text Engineering. Available online: http://www.gate.ac.uk (accessed on 21 April 2014).
33. Hare, J.S.; Samangooei, S.; Dupplaw, D.P. OpenIMAJ and ImageTerrier: Java Libraries and Tools for Scalable Multimedia Analysis and Indexing of Images. In Proceedings of the 19th ACM international conference on Multimedia, Scottsdale, AZ, USA, 28 November–1 December 2011; pp. 691–694.
34. OpenIMAJ. Open Intelligent Multimedia Analysis in Java. Available online: http://www.openimaj.org (accessed on 21 April 2014).
35. International Organization for Standardization. *Information and Documentation—WARC File Format*; ISO 28500:2009; ISO: Geneva, Switzerland, 2009.
36. Hare, J.S.; Samangooei, S.; Dupplaw, D.P.; Lewis, P.H. Twitter's Visual Pulse. In Proceedings of the 3rd ACM International Conference on Multimedia Retrieval, Dallas, TX, USA, 16–19 April 2013; pp. 297–298.
37. Hare, J.; Samangooei, S.; Dupplaw, D.; Lewis, P. ImageTerrier: An Extensible Platform for Scalable High-Performance Image Retrieval. In Proceedings of the ACM International Conference on Multimedia Retrieval, Hong Kong, China, 5–8 June 2012.
38. Gionis, A.; Indyk, P.; Motwani, R. Similarity Search in High Dimensions via Hashing. In Proceedings of the 25th International Conference on Very Large Data Bases, Edinburgh, Scotland, UK, 7–10 September 1999; pp. 518–529.
39. Dong, W.; Charikar, M.; Li, K. Asymmetric Distance Estimation with Sketches for Similarity Search in High-Dimensional Spaces. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, Singapore, 20–24 July 2008; pp. 123–130.
40. Dong, W.; Wang, Z.; Charikar, M.; Li, K. High-Confidence Near-Duplicate Image Detection. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieva, Hong Kong, China, 5–8 June 2012; pp. 1:1–1:8.
41. Dupplaw, D.P.; Hare, J.S.; Samangooei, S. Twitter's Visual Pulse Demo. Available online: https://www.youtube.com/watch?v=CBk5nDd6CLU (accessed on 21 April 2014).
42. MediaEval. MediaEval Benchmarking Initiative for Multimedia Evaluation. Available online: http://www.multimediaeval.org (accessed on 21 April 2014).

43. Reuter, T.; Papadopoulos, S.; Mezaris, V.; Cimiano, P.; de Vries, C.; Geva, S. Social Event Detection at MediaEval 2013: Challenges, Datasets, and Evaluation. In Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, 18–19 October 2013.
44. Samangooei, S.; Hare, J.; Dupplaw, D.; Niranjan, M.; Gibbins, N.; Lewis, P.; Davies, J.; Jai, N.; Preston, J. Social Event Detection Via Sparse Multi-Modal Feature Seating Contest and Incremental Density Based Clustering. In Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, 18–19 October 2013.
45. Parkhi, O.; Vedaldi, A.; Zisserman, A. On-the-Fly Specific Person Retrieval. In Proceedings of the 13th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Dublin, Ireland, 23–25 May 2012; pp. 1–4.
46. Psyllos, A.; Anagnostopoulos, C.N.; Kayafas, E. M-SIFT: A New Method for Vehicle Logo Recognition. In Proceedings of the IEEE International Conference on Vehicular Electronics and Safety, Istanbul, Turkey, 24–27 July 2012; pp. 261–266.
47. Kalantidis, Y.; Pueyo, L.G.; Trevisiol, M.; van Zwol, R.; Avrithis, Y. Scalable Triangulation-Based Logo Recognition. In Proceedings of the 1st ACM International Conference on Multimedia Retrieval, Vancouver, CB, Canada, 26–31 October 2008; pp. 20:1–20:7.
48. Davies, J.; Hare, J.; Samangooei, S.; Preston, J.; Jain, N.; Dupplaw, D.; Lewis, P.H. Identifying the Geographic Location of an Image with a Multimodal Probability Density Function. In Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, 18–19 October 2013.
49. Hauff, C.; Thomee, B.; Trevisiol, M. Working Notes for the Placing Task at MediaEval 2013. In Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, 18–19 October 2013.
50. Hare, J.S.; Davies, J.; Samangooei, S.; Lewis, P.H. Placing Photos with a Multimodal Probability Density Function. In Proceedings of the International Conference on Multimedia Retrieval, Glasgow, Scotland, UK, 1–4 April 2014.
51. Cootes, T.F.; Taylor, C.J.; Cooper, D.H.; Graham, J. Active shape models—Their training and application. *Comput. Vis. Image Underst.* **1995**, *61*, 38–59.
52. Cootes, T.; Edwards, G.; Taylor, C. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intel.* **2001**, *23*, 681–685.
53. Ekman, P.; Friesen, W. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*; Consulting Psychologists Press: Palo Alto, CA, USA, 1978.
54. Schmidt, S.; Stock, W.G. Collective indexing of emotions in images. A study in emotional information retrieval. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 863–876.
55. San Pedro, J.; Siersdorfer, S. Ranking and Classifying Attractiveness of Photos in Folksonomies. In Proceedings of the 18th International World Wide Web Conference, Madrid, Spain, 20–24 April 2009; pp. 771–780.