Multimed Tools Appl (2009) 42:73–96
DOI 10.1007/s11042-008-0249-5

Attributing semantics to personal photographs

Rodrigo F. Carvalho · Sam Chapman · Fabio Ciravegna

Published online: 14 November 2008

Abstract A major bottleneck for the efficient management of personal photographic collections is the large gap between low-level image features and high-level semantic contents of images. This paper proposes and evaluates two methodologies for making appropriate (re)use of natural language photographic annotations for extracting references to people, location and objects and propagating any location references encountered to previously unannotated images. The evaluation identifies the strengths of each approach and shows extraction and propagation results with promising accuracy.

Keywords Photographs · Semantic capture · Information extraction · Clustering · Image · Annotation

1 Introduction

In recent years digital cameras have become an essential gadget in the household. With the increasing adoption of mobile photography, inexpensive network transmissions, cheap data storage and a decline in physical printing there is an inevitable expanding number of photographs in both public and private digital collections.

In the domain of personal photography, and more specifically within family centred communities, there is a growing need to exploit the existence of semantic metadata for facilitating retrieval, organisation and reuse of images. However the

R. F. Carvalho (✉) · S. Chapman · F. Ciravegna
Department of Computer Science, Natural Language Processing Group,
The University of Sheffield, 211 Portobello, S1 4DP, Sheffield, UK
e-mail: rodrigo@dcs.shef.ac.uk

S. Chapman
e-mail: sam@dcs.shef.ac.uk

F. Ciravegna
e-mail: fabio@dcs.shef.ac.uk

issue of obtaining such semantic metadata for photographs remains unsolved. The development of an approach that tackles this problem presents one main challenge: how to design and implement efficient metadata capture and reuse strategies. Previous studies by Frohlich et al. [8], and Miller and Edward [16], highlight that most users in the personal photograph domain are unlikely to make extensive manual contributions to generating semantic metadata relating to images; this attitude could be exasperated by the ever increasing size of personal photograph collections.

Automated solutions aiming to index images for retrieval are incomplete as they frequently fail to address user needs: retrieval from user relevant concepts relating to individual images. Such approaches are problematic due to two main issues, (1) sparsity of manually provided metadata, (2) the *semantic gap* between automatable approaches and actual users concepts. For example *C*ontent *B*ased *I*mage *R*etrevial techniques, CBIR [25], index visual artifacts and patterns within images (not user relevant concepts themselves) which aid retrieval from examples but not conceptual retrieval. Other techniques try to use the user to provide additional semantic metadata for images they own or view. Online photo sharing services are examples of such systems that encourage image reuse and sharing by utilising additional user input, i.e. comments, tags, temporal and categorical groupings and organisation.

In this paper we examine the sparsity of semantic metadata taking a large corpus of personal photographs gathered from a leading online photo sharing service considered to have a high degree of metadata.[1]

In Section 4 we present an information extraction methodology for capturing conceptual semantics from image captions and descriptions. In order to overcome linguistic problems specific to this domain, our approach combines a machine learning framework (T-Rex) [13] with a rule based extractor (Saxon) [9] obtaining promising results.

We also propose and evaluate a strategy for (re)using such semantic metadata for generating more metadata for other images by using the concept of semantic propagation.

Section 5 examines the use of temporal data in conjunction with visual features for determining the usefulness of either when used separately or in conjunction with one another for performing semantic propagation.

2 Existing approaches

Many approaches aim to address the problem of facilitating image sharing and reuse. Current techniques focus upon one of four basic approaches, each of which is now detailed briefly.

2.1 Image analysis

Image analysis techniques attempt to extract meaning from the pixel content of an image automatically. Veltkamp [25] surveys the state of the art techniques such as

[1] Flickr http://www.flickr.com.

face recognition, edge detection, image segmentation, region classification etc. Such techniques, however, are largely problematic in real world scenarios for two reasons:

1. **Semantic gap**—extracted regions are visual artifacts within pixels and not semantic concepts which users require, for example, *an objects boundary edge* and not semantic entities like *My brothers car*, *dad* or *the Eiffel tower*.
2. **Accuracy**—state of the art has an unacceptable precision and recall to be considered useful in that objects and classifications can be frequently misapplied. Barla et al. [4] indicate a miss-classification of 20.7% in even rudimentary binary classifications such as cityscape vs non-cityscape.

2.2 Improved structured knowledge representations

Representing knowledge in a standard format is of huge importance as it facilitates its reuse and retrieval. In recent years a number of exchange formats have been developed focusing specifically upon exchanging information regarding digital images. Exif[2] includes detailed camera settings set at the time of digital image capture. Some of this information is of use for retrieval but again suffers from the issue of the *semantic gap* where it fails to embed semantic meaning needed by users. Newer standards such as *MPEG-7* [15, 21] provide a mechanism to encode extended information including regional semantic metadata within an image; unfortunately, although a format exists for its representation, there is as yet no agreed method to obtain the needed metadata.

2.3 Metadata propagation

To cope with minimal semantic metadata it is possible to expand existing metadata, increasing its richness by using external resources like wordnet [2]; such approaches, however, do not increase coverage over sparse metadata, they merely enrich where it already exists. Also, such approaches decrease precision. Alternatively, some research has been undertaken to propagate metadata associated to images that have not been annotated [14]. Hare [10] provides a method to propagate metadata by just using image features and existing metadata.[3] Such approaches however suffer from the issues of semantic gap as well as precision and recall, being largely unacceptable to users. Other more promising approaches try to propagate purely from the social graph of connected information [5] upon the assumption that connected information is likely to be linked semantically.

2.4 User (or community) input

2.4.1 Manual annotation

Enlisting user support in image classification has had a recent resurgence in popularity following the success of the ESP game [1] and the development of online

[2] *EX*changeable *I*mage file *F*ormat, was created by the Japan Electronic Industries Development Association (JEIDA). Version 2.1 (the first public release) was released June, 1998, and later updated to version 2.2 in April 2002.

[3] See 2.4.2 for wider usage of such features.

photo sharing websites such as Flickr, KodakGallery and many more. Such interfaces empower users to perform individual or collective annotation/archival of digital photographs. One issue with such approaches is that only a small proportion of the population put in reasonable efforts regarding annotation. Given such systems it is imperative to make maximal usage of any photographic annotations. Many attempts have been made to extract maximal meaning from photographic annotations such as [19, 23] but most approaches have focused upon a complete natural language parse which is costly to scale to a large scale solution and problematic especially in shorter often ungrammatical photographic descriptions.

2.4.2 Image retrieval based on relevance feedback

In the image retrieval research field, many approaches have been developed to date that make use of relevance feedback for addressing the constraints presented by the limited existence of semantic metadata within image databases. These techniques expand image search results by allowing users to perform keyword searches on image databases and then use the feedback from the user for expanding results retrieved based on visual features of the images indicated as relevant to the query. The work in [3] involves the use of relevance feedback to warp the feature space that contains all images in a VRML (Virtual Reality Modeling Language) database in order to use positive feedback to pull vector points (i.e. images) closer to the query vector and negative feedback to push irrelevant vector points away from the query vector. In [26] images are represented as part of a semantic network that is composed of the images themselves, keywords and weighted links between the two. After running textual queries, relevance feedback is used to reinforce the weights of the connections between images and keywords and has been shown to produce confident results achieving over 95% accuracy after 8 relevance feedback iterations. For a more comprehensive survey on the uses of relevance feedback using CBIR techniques please refer to [27].

While such approaches have been demonstrated to perform well for generic retrieval tasks, they have crucial drawbacks in the domain of personal photography where clear difficulties exist in trying to tie concepts of interest to users to the visual contents of images alone. The concept of *location* can serve as an example of this

Fig. 1 The first image depicts India while the second depicts Turkmenistan. Comparing these images by using their visual similarities alone, however, would result in an incorrect classification regarding their location

where visual similarity is rarely enough for discriminating between two images (see Fig. 1).

3 The photographic domain

This paper focuses on attributing semantics to photographs according to the user needs and preferences. A study performed by Naaman et al. [17] determined the usefulness of various metadata in aiding users to locate their own photos. The cues rated by users as contextually important for recalling images were found in order of importance to be:

1) indoors opposed to outdoors pictures,
2) the identity of people within a photo,
3) the location,
4) the event depicted,
5) the number of people,
6) the season,
7) the year,
8) the time of the day,
9) the weather conditions,
10) the date,
11) the mood of the photograph

Further research undertaken by industrial bodies confirmed these results.[4] It is relevant to notice how some features from the above list could be obtained from sources other than the image annotations themselves. For example some of the above features could be extracted from Exif metadata, while basic image analysis could obtain features like e.g. indoor vs outdoor environments [4] having 93% accuracy. However, there is still a lack of relevant features that cannot be obtained reliably by employing the above mentioned methods, therefore more advanced analysis methods must be envisaged.

In this paper we focus on three main features:

- **Location**: a textual location that the image might depict. This includes not only geographical location names but also far less exacting locations such as *home*, *my road*, *my garden* as well as synonyms for place names such as *the big apple*.
- **Person**: people's names or general references to people such as *dad*, *mum*, *brother*.
- **Object**: conceptual objects depicted in an image. This concept was only identified when a term of obvious importance did not fall into any of the previous categories, such as *football* in the description *Dave and his football*.

With the existence of online photo management and sharing services such as Flickr for almost half a decade, users of this technology have grown accustomed to organising their photo collections by using textual metadata such as single (or multiple) words known as *"tags"* as well as textual descriptions of an arbitrary length. The existence of such metadata about images has opened opportunities for the development of novel techniques for the extraction of information about images by using approaches that take advantage of features from different media such as text (Natural Language Processing—NLP) and visual descriptors (*MPEG-7*).

[4] Internal communications with Kodak Limited.

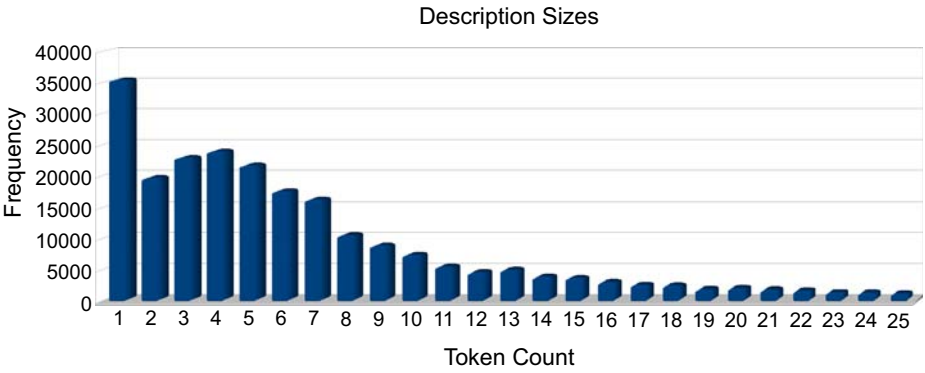

Fig. 2 Descriptions' token count

It is believed that an optimum solution for attributing semantic metadata to images in the domain of personal photography that makes use of image descriptions would have to address the following requirements:

- It must be computationally cheap (*light-weight*) in order to be scalable.
- The extractions produced must be highly precise while maintaining recall.
- It must make maximum (re)use of any metadata provided (e.g. descriptions, tags, etc.).

3.1 Corpus collection

To support the development of a solution for attributing semantic metadata to images in the domain of personal photography, 2,325 online Flickr users were contacted over a period of 4 months for permission to use their public images to build a corpus of image metadata. During this period there were a total of 414 responses, of which 391 replied positively.

The corpus gathered includes over 2.3 million tokens[5] distributed among over 240K image descriptions. This is largely characterised by short disconnected snippets of text (see Fig. 2) describing users photographs.

An overview of the corpus reveals two of the main obstacles to using image descriptions for attributing semantic metadata to images in this domain:

- You can see from Fig. 2 that over 64% of image descriptions in the corpus gathered from Flickr have less than 10 tokens even before performing any stop-word removal. Performing Information Extraction (IE) from such short snippets of text using typical Natural Language approaches can be problematic due to their limited linguistic content.
- What Fig. 3 confirms is that the existence of textual metadata in the form of descriptions tends to cover only a portion of the corpus as a whole and,

[5]A token is a categorized block of text. In the context of this work, a token is a word in a text separated by white spaces from other tokens.

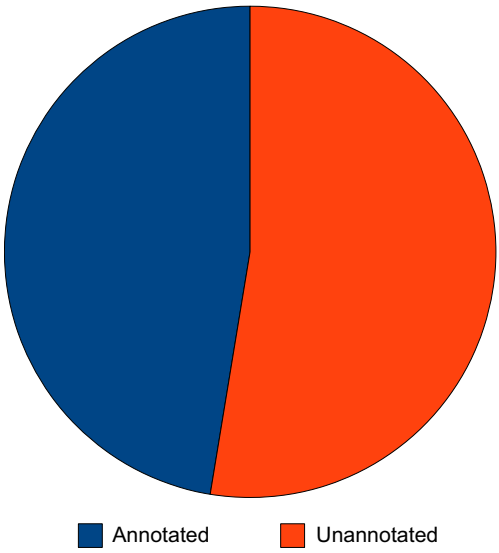

Fig. 3 Proportion of annotated vs. unannotated photographs in the main corpus

while the actual ratio of annotated versus unannotated images may vary across the collections of different users, it is a major obstacle in attributing semantic metadata to photographs for later retrieval.

What we propose in this paper is the use of an approach for extracting information from image descriptions alone, excluding tags or image titles, that takes advantage of the flexibility of machine learning data models as well as of the precision of rule based extractors. Given a very limited initial training dataset as well as a limited number of rules, we aim to combine these two approaches not only for performing more confident extractions from image descriptions, but also to control levels of precision and recall by maintaining a balance over which technique is more influencing in the extractions. Furthermore, taking into consideration the scarcity of image descriptions, we propose and evaluate a strategy for propagating *location* semantic captures obtained from image descriptions to unannotated images within a user's collection that uses an unsupervised clustering technique and exploit visual and temporal similarities between images.

The rest of this paper is organised as follows: Section 4 will introduce the machine learning framework, the rule based extractor and a hybrid approach for extracting semantics from image descriptions and Section 5 introduces the algorithms used in devising a propagation strategy as well as the proposed strategy itself. What follows from that is a separate evaluation of the hybrid approach for semantic capture and semantic propagation before a final discussion and conclusions.

4 Semantic capture

As can be seen from Fig. 2, performing semantic capture from image descriptions is a challenging task due to the limited linguistic content of text in this domain.

4.1 Machine learning

It is widely known that given a small set of training data, machine learning systems are capable of creating a generic model and apply it to previously unseen data. More specifically, in the field of NLP, textual features of tokens (e.g. part of speech, orthography, the tokens themselves, etc.) together with the features of other neighbouring tokens are used in the creation of this model, what makes this an extremely flexible technique for extracting information from text.

4.1.1 T-Rex

One such system that achieves competitive results when applied to several corpora is the *T*rainable *R*elation *EX*traction framework (T-Rex[6])[13]. T-Rex is a highly configurable support vector machine based IE framework that uses a canonical graph-based data model. Its strength comes from decoupling its data representation from the machine learning algorithms allowing configurable extensions.

4.1.2 Training data

For building a training dataset for the machine learning framework a total of 1,660 English image descriptions (24,215 tokens) belonging to 54 distinct users were randomly collected from the main corpus. This smaller corpus was then manually annotated by a group of 7 researchers according to the concepts introduced in Section 3 generating a total of 2,522 annotations. More specifically, 566 annotations were assigned to the concept of *Person*, 747 to *Location* and 1,209 to *Object*. This dataset was then subdivided into 2 sets: the annotated data used by T-Rex as training data and Saxon as a basis from which to build extraction rules (40%), the remaining data was used for testing. Further image descriptions were also collected from the main corpus at a later stage for evaluating the approach.

4.2 Rules

On the opposite end of the spectrum there are rule based extractors that apply manually written Hearst pattern [12] style rules to textual data. Precise extractions can then be performed according to the granularity of rules.

4.2.1 Saxon

It is a rule based tool for annotating documents and is built upon the Runes framework [9].[7] It relies on the document being represented as a graph, with nodes representing document elements (tokens, sentences, etc.) and edges representing relationships between elements (belongsTo sentenceXYZ, follows tokenXYZ, etc.). Saxon rules are defined as regular expressions detailing how to move between elements of the graph. A rule has three main parts: a starting point, a regular expression (describing how to move between sections of the graph) and an update

[6] http://www.sourceforge.net/projects/t-rex

[7] Saxon—http://nlp.shef.ac.uk/wig/tools/saxon/index.html
Runes—http://runes.sourceforge.net/

rule (detailing how the graph should be updated if the rule matches). Further to these, a rule can also make use of external gazetteer lists for reinforcing its precision by detecting better matches within a concept. The full flexibility of Saxon lies, however, in the ability to specify unrestricted Java code as the right hand side of a rule. The output of a rule can be either other annotations or unrestricted actions specified within the rule.

Below is an example of one such rule for capturing *person* instances:

```
1. Rule:PersonWithTitleUsingLookup
2. Lookup{major_type{=title}}
3. ((?lookup_has_last_token)(token_next{token_pos{=NNP}})+)
4. =>
5. [Person]
```

The rule above is aided by the existence of a gazetteer list that contains examples of personal titles such as Mr., Mrs., Ms., Dr., etc. So its starting point (line 2) is a token in a sentence that indicates a personal title. Once the starting point is matched, the regular expression part (line 3) defines what to do. In this case it navigates forward in the sentence collecting all the proper nouns it detects in a sequence and returns this as a potential person name. The last line is the update rule and in this case it simply instructs Saxon that the graph nodes matched should be annotated with the token *Person*.

4.2.2 Rule development

The development of rules was an iterative process and involved the manual encoding of patterns from annotated data for capturing specific concepts. The process took place in 3 stages: one for each concept defined in Section 3. At the end of the process, 15 generic rules were developed. Four rules for *'Person'* aided by the use of gazetteer lists for detecting common first names, references to family relatives (e.g. mom, dad, brother, etc.) as well as person titles (e.g. Mr, Dr, etc.). Six rules for *'Location'*, 5 of which were reinforced by gazetteers for detecting common locations (e.g. countries, cities, etc.) as well as tokens indicative of references to a location (e.g. museum, street, etc.). Five rules for *'Object'* were extracted, 3 were reinforced by organisation gazetteers to detect instances that refer to branded objects (e.g. McDonald's sandwich, Lincoln engine, etc.).

4.3 Hybrid IE

In order to successfully extract information from image descriptions, it is arguable that either technique implemented by T-Rex or Saxon could be singularly applied to the task. However, because of the constraints imposed by the domain and the requirements introduced at the beginning of Section 3, each system carries with it disadvantages.

Despite implementing a flexible approach for IE, T-Rex depends heavily upon the size and coverage of the initial training dataset which is costly to develop. Also, when configured for performing highly accurate extractions, computational cost can be impractical for use in scalable applications. Saxon, on the other hand, while being less computationally expensive, requires time consuming development of rules for capturing every desirable case within the text, which makes it less flexible for

performing IE. What we propose here is that the combining of the two techniques implemented by T-Rex and Saxon not only lessens their disadvantages, but also gives way to improved precision and recall while maintaining the approach as scalable as possible.

One of the first issues to be addressed by the combination of the two techniques is an architectural one. Machine learning approaches, as mentioned previously, utilise tokens' textual features from training data to build a generic data model that can be applied in previously unseen cases. In order for this data model to be highly accurate, multiple features must be recorded about as many neighbouring tokens as possible implying complexity and increased computational cost for an extraction task.

The domain of image descriptions as discussed previously is unique. This means that for typically short texts, the size of the context a token can be placed in almost always shrinks down to 1 or 2 neighbouring tokens. The creation of a machine learning data model that reflects this reduces overall computational costs. On the other hand, in reducing the size of the contextual information gathered for the creation of an appropriate data model, the accuracy of extractions performed by T-Rex are also decreased. It now takes less constraints to be satisfied for finding a token fitting the model created. A potential solution would be to produce a greatly expanded training dataset but this would be a prohibitive option since it would not only be costly but also difficult to obtain a dataset that is comprehensive enough. The most suitable solution for improving the accuracy of extractions could therefore lie in the use of a rule based extractor.

Unlike in singularly developed rule based extractors, in developing a hybrid approach to IE, Saxon rules can be built in a generic way, thus speeding up development (i.e. less rules), as well as improving recall. While this would have a massive effect on precision in exclusively rule based systems, in a hybrid approach extractions can be compared according to different resources, thus giving rise to improved precision.

The essence of the approach therefore lies in extracting information from an annotation by using a combination of the extraction suggestions from each system. So in order to better combine these extractions, a *weighted voting strategy* was devised that gave rise to an opportunity for taking advantage of both systems' strengths while attenuating the effects of their weaknesses. This *voting* method can be subdivided into three distinct phases (Fig. 4):

- **Extraction**: each system puts forward potential extractions found in an image description.
- **Voting**: based on their separate findings, Saxon and T-Rex *"vote"* on each extraction according to a set of weights attributed to each system.
- **Ranking**: the number of votes cast on the tokens of each extraction are used to give it a *"confidence"* ranking according to heuristically specified ranges (i.e. between 0.8 and 1 - *high*, between 0.5 and 0.79 - *medium* or between 0 and 0.49 - *low*).

In the *extraction* phase, an image description is passed to each system separately and both generate a list of potential extractions from the original text together with their corresponding classifications (i.e. person, location or object). Once the potential extractions are identified, systems vote on the set of extractions based on their own findings and pre-defined weights. An obvious example of this would be in

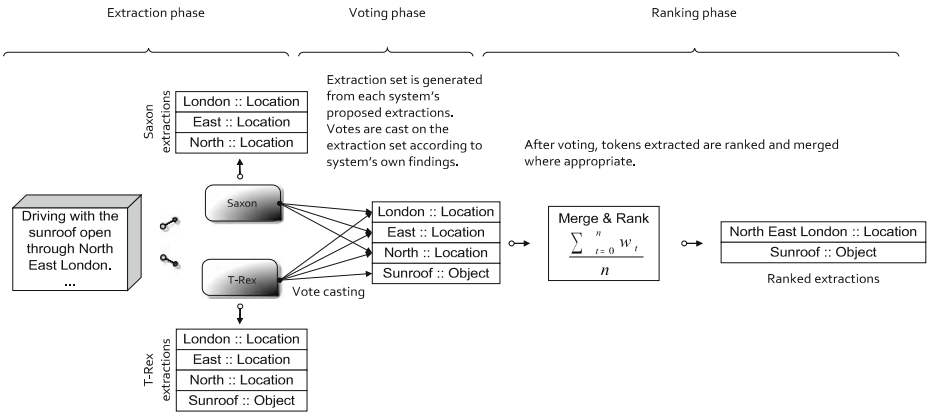

Fig. 4 Voting strategy

the description *"Driving with the sunroof open in North East London."* whereby both T-Rex and Saxon vote for all the tokens within *"North East London"* as referring to a location and only T-Rex votes for the token *"sunroof"* as referring to an object.

The set of votes V_t each token t receives can then be represented as $V_t = \{w_0, ..., w_r\}$ where w_r is the weight of the vote received from resource r (i.e. Saxon or T-Rex). The accumulated weight w_t for each token is obtained from the sum of the vote weights w_r that make up V_t for token t (see Eq. 1). The confidence ranking r_e for each merged extraction E that is composed of n tokens where $E = \{t_0, ..., t_n\}$ can be obtained from the sum of each tokens' accumulated weight w_t divided by the number of tokens n that compose the extraction (see Eq. 2).

$$w_t = \sum_{w_r \in V_t} w_r \ . \tag{1}$$

$$r_e = \frac{\sum_{t=0}^{n} w_t}{n} \ . \tag{2}$$

As exemplified, the votes are cast at the level of tokens. This allows extractions to be ranked according to what T-Rex and Saxon find regarding each single token that may be part of a larger entity. Once the accumulated weights for tokens are obtained, neighbouring tokens are then merged according to a combination of their weight, their extraction type and the confidence ranking expected from each extraction (i.e. *high*, *medium* or *low*). So in the example above, the tokens *North*, *East* and *London* are merged since their overall confidence ranking is very high (i.e. 1) and they were classified with the same type. However, not all extraction combination scenarios are complementary.

One of the strengths of this strategy is its ability to resolve overlapping extractions according to the three levels of confidence mentioned previously. A typical example is *"Autumn in Arlington cemetery"* whereby T-Rex extracts the token *Arlington* as a location and Saxon extracts *Arlington cemetery*. Both extractions are conceptually correct although one is more complete than the other. After voting the token *Arlington* would arise as being a **high confidence** extraction, whereas the token *cemetery* would be classified as **medium confidence**. Depending on the confidence

ranking expected, the final result could either be an extraction ranked with **medium confidence** that incorporates both tokens *Arlington cemetery* or an extraction ranked with **high confidence** that only includes the token *Arlington*. This is one of the advantages of using a *weighted voting strategy* in that it enables not only decisions on which extractions are the strongest, but also consider the ones that are not so strong as opposed to simply discarding them. One feature that arises from the existence of such rankings is that they allow the final extractions to be geared towards either one of high precision or high recall.

More problematic conflicts such as the disagreement regarding an extraction classification cannot be resolved by simply applying the three levels of confidence introduced above. This is where the full flexibility of a weighted voting strategy lies, in that the assigning of weights to votes can not only be used for ranking extractions but, when tweaked to reflect a higher confidence in the more precise technique at hand, can be used for resolving extraction type disputes across systems. An example found during the testing of the approach that would fit into this situation comes from descriptions such as *"Auray in Brittany; North-West France"*, where *Auray* is classified by T-Rex as a person and a location by Saxon. It is clear in this instance Saxon has classified the extractions correctly and this can be mainly attributed to the tokens being a correct match to an existing rule for extracting locations that is reinforced by a gazetteer list, thus yielding more precise extractions. Therefore, in order to resolve conflict as exemplified above, the same voting strategy is used, but with the weights reflecting a higher confidence in Saxon as being the more precise technique in such circumstances and providing a means to resolve problems previously presented to either an exact match combination or an overlapping extraction.

Despite the usefulness of this semantic capture approach for extracting information from image descriptions, it still relies heavily on comprehensive human input for each and every image in a collection. Contextual information about the image can help remediate this where the existence of semantic metadata about a small proportion of images in a user's collection can be used in conjunction with appropriate similarity metrics for propagating textual metadata to unannotated images.

5 Semantic propagation

Semantic propagation aims to take advantage of semantic metadata that is known about photographs in a collection for attributing semantics to photographs that have no such metadata assigned to them. Within this work in specific, making use of this type of approach attempts to augment the impact of user input for certain key images within a collection while reducing the laborious and intensive process of human annotation. That is, we assume that descriptions or annotations have been created for only a few images in a user's collection and aim to propagate these annotations to other images by performing similarity comparisons.

The susceptibility of images to an external context in this domain makes this type of approach especially attractive. It is often the case that photographs belong to a collection or album, and the likelihood of any one photograph depicting a certain concept c, such as location, will be heavily influenced by the content of other photographs in the same collection. For instance, unannotated images that belong

Fig. 5 All images above belong to the same collection, however only the first image in this series contains textual metadata (i.e. *Everest*)

to an album where other images have been annotated with *Everest* have a much stronger chance of also depicting the same location than images in other albums.

Figure 5 exemplifies a situation where the existence of an external context in photographic collections can be explored to remediate the lack of semantic metadata for unannotated photographs. By taking advantage of contextual information that connects a seed image to other images in the same collection the information that is know about a seed image could be propagated to other photographs.

So in the example above, the location information extracted from the description of the first image (i.e. *Everest*) could be applied to the other 2 photographs in the series by determining that they look visually similar.

Naturally, visual similarity is not the only determining factor in whether two photographs depict similar concepts, and could in fact be very misleading when considered separately from other features such as the timestamp of a photograph. What can be seen from Fig. 1 is a typical example where the visual similarity between two images fails to play a significant factor in distinguishing two photographs in the same collection.

An important fact in this domain is that photographs capture moments in *time* and *space* and a typical characteristic of personal photographic collections is the existence of a timeline that determines not only a logical sequence of events, but also their duration. Concepts such as location can be shown to be highly correlated with time. Figures 6 and 7 plot the usage of location tags against time in the collections

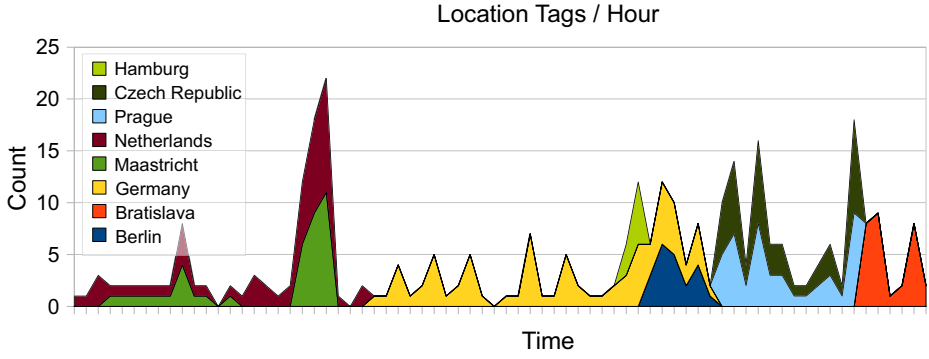

Fig. 6 Location tags usage per hour for user A during periods of photographic activity

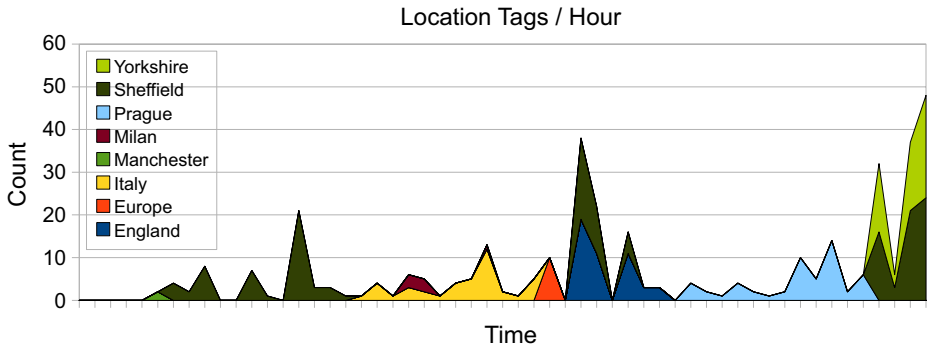

Fig. 7 Location tags usage per hour for user B during periods of photographic activity

of 2 arbitrary contributors to the main corpus during a period of 6 months. Each graph's temporal granularity is hours, and clearly demonstrate not only the stable relationship between time and location but also the potential to use temporal cues as a means for propagating location semantic captures to other images in each user's collection.

In order to address the issue of generating semantic metadata for photographs in the family domain we propose a strategy that explores this relationship between the concept of *location* and time. The basic assumption behind the approach being that users are likely to generate some semantically relevant metadata about key images within a collection.

5.1 Propagating semantic captures

In order to overcome the drawbacks of relying on extensive user input or imprecise computer vision methods for determining the semantic contents of images, we propose a strategy for propagating the semantic concepts of key images within a user's collection to other previously unannotated images. The approach takes advantage of unsupervised clustering techniques such as K-Means [7, 11] for contextualizing photographs into semantically proximate partitions. The aims of this work are (1) to determine the usefulness of a propagation approach in the domain of personal photography and (2) to evaluate the use of temporal and visual data in semantically linking photographs.

5.1.1 Propagation approach

The approach described here takes advantage of the visual content of images as well as the temporal nature that predominates in photographic collections with the aim of extending the IE approach presented previously and making it more robust when faced with unannotated photographs. In order to maximise the semantic proximity between photographs, the approach makes use of X-Means [20], a variation of the K-Means clustering algorithm that searches for an optimum number k of clusters.

To evaluate the usefulness of temporal and visual features for propagating location semantics 3 clustering scenarios were considered: (1) temporal only, (2) visual

only, (3) visual and temporal. These scenarios were aimed at determining the impact of either feature separately and in conjunction with each other on the precision of the propagation of semantic concepts to unannotated images.

The essence of the strategy proposed here lies in (re)using any semantic metadata associated with images. To achieve this, the strategy is divided into three distinct phases:

1. Capture of semantic metadata for images.
2. Clustering of images according to either visual or temporal features or both.
3. Propagation of existing metadata to unannotated images.

During the first phase, the semantic capture methodology explained in Section 4.3 is applied to all images containing descriptions in a user's collection for extracting location names. For the purpose of further enriching the extractions, place name gazetteers were used to detect location tags associated with images. Once the data capture has finished, only the most confident extractions from the image descriptions are kept in conjunction with any tags found to refer to locations according to the gazetteer lists. The use of gazetteer lists is also applied to control the level of granularity for location tags which is currently maintained at the country level.

The second phase involves creating clusters of images according to either their visual or temporal features or both. The timestamp of photographs, obtainable from Exif headers, contains the date and time when the photograph was taken and is used as the temporal feature for photographs. With the growing adoption of *MPEG-7* as the multimedia content description standard, two standards compliant visual descriptors were used for computing photographs' visual features:

- *Scalable Colour Descriptor* (SCD): is a Colour Histogram in HSV Colour Space, which is encoded by a Haar transform to reduce the large size of its representation. It measures the colour distribution over an entire image.
- *Edge Histogram Descriptor* (EHD): represents the spatial distribution of four directional edges (0°, 45°, 90°, 135°) as well one non-directional. Images are divided into 16 non-overlapping blocks and an extraction scheme is applied to extract the five types of edges and their relative populations.

Previous works have shown the usefulness of both SCD and EHD for image matching [18, 24], but more specifically, it has been found that the combination of both is beneficial in certain domain problems [22] and especially in performing image-to-image matching of similar semantic meaning. For more information about *MPEG-7* compliant descriptors please see [15].

During the second phase, the creation of such clusters follows from the idea that semantic propagation should only happen when images demonstrate a high enough level of relatedness. The clustering of photographs therefore helps put images into the context of other semantically proximate images. While it may be argued that visual similarities for event clustering are not entirely appropriate, content similarity can predominate for photographs with sufficient temporal proximity [6]. There would be a higher likelihood of semantic metadata being propagated within such clusters.

Parting from the premise above that visual and temporal similarities (separately or combined) can be used for partitioning an image collection into clusters of highly

inter-related photographs, the following clustering experiments were devised for evaluation:

- **Visual Clustering:** both SCD and EHD descriptors were used for obtaining the visual features of images and X-Means was then used to partition the vector space of a user's collection into k clusters according to the Euclidean distance between vector points.
- **Temporal Clustering:** due to the linear nature of temporal data and findings in previous studies [6], it was found that a threshold based clustering technique would suffice for partitioning a user's collection according to the temporal proximity between adjacent photographs. The timestamp of photographs was used to determine the temporal proximity between photographs according to a threshold of 24 h selected *a priori*. The selection of this particular threshold is based on the level of granularity of location extractions which is currently maintained at the country level.
- **Temporal + Visual:** in order to preserve the linear effect of a timeline on a user's photographic collection, the approach taken to combine both content and time involved an iterative clustering strategy, whereby the collection for a user would be first clustered temporally as indicated in the previous point and then within each temporal cluster several visual clusters would be created to maximise the semantic proximity between images.

Once the clusters were obtained for each of these strategies, a *hard* propagation methodology was adopted that comprised of attributing location references from annotated photographs within a cluster to other photographs within the same cluster uniformly. It is understood that applying a probability distribution over clustered images in order to propagate semantics would potentially yield better results, but in order to answer the question of whether a temporal clustering strategy would outperform other approaches based on visual features it was found that this propagation methodology would suffice.

The next section presents an evaluation of both the semantic capture and semantic propagation methodologies presented in this paper. What follows from that is the conclusion and final remarks on the work presented here.

6 Experimental results

This section is divided into 2 parts: The first evaluates the semantic capture approach presented in Section 4 while the second evaluates the semantic propagation strategy proposed in Section 5.

Our evaluation goals range from finding exceptional circumstances beyond the grasp of the approaches as well as:

- comparing traditional solutions (i.e. rules and machine learning) against the proposed approach to the problem of extracting entities from short image descriptions.
- comparing the effectiveness between visual and temporal features in clustering for the task of *location* semantics propagation. Do visual features help at all in the propagation process or would the use of temporal features alone suffice?

6.1 Semantic capture

So the evaluation of the task involved the detection of all occurrences of locations, people and objects in an image description. The definition of how we decide whether extractions made are correct or not is crucial for the computation of evaluation scores. For the evaluation of the hybrid approach detailed earlier three different possibilities were considered:

- **exact rule**: a prediction is only correct, if it is exactly equal to an answer.
- **contain rule**: a prediction is correct, if it contains an answer, plus possibly a few extra neighbouring tokens.
- **overlap rule**: a prediction is correct, if it contains a part of a correct instance, plus possibly some extra neighbouring tokens.

An evaluation set of 100 previously unseen image descriptions that spanned the collections of 3 different users was randomly selected from the main corpus. This set was then manually annotated before being processed both by T-Rex and Saxon individually and as part of a hybrid system. The following results were obtained for extractions that were ranked in the *high* and *medium* confidence ranges.

	T-Rex		Saxon		Hybrid	
Concept	Precision	Recall	Precision	Recall	Precision	Recall
Person	67%	63%	70%	76%	86%	82%
Location	80%	62%	91%	77%	92%	79%
Object	75%	61%	75%	60%	73%	63%

As it can be seen from the results above, the hybrid approach outperforms T-Rex and Saxon when run individually for extracting instances of *'Person'* and *'Location'* from image descriptions, while for instances of *'Object'* there is no noticeable overall improvement. Each system's extractions were then shaped by their strengths and weaknesses and in most cases combined with great success using the hybrid approach. For instance, T-Rex was able to contextually detect the uses of unknown words to refer to locations depicted in the photograph such as *La Louvre* in *"Floaton on a fountain by La Louvre"* where Saxon failed. On the other hand, the usefulness of gazetteers and the precision of rules allowed Saxon to detect tokens such as *Harry Potter* in *"Of Harry Potter fame"* while T-Rex failed to do so.

Classification types can be corrected, in *"Low cloud on Mont Victoire"* T-Rex misclassified *Mont Victoire* as a person and Saxon correctly resolved the entity to a location. Other examples such as *"Big French sandwich"* and *"Worst seat in the best court"* demonstrated the flexibility of Saxon rules in complementing T-Rex's extractions of *sandwich* and *seat* with *Big French sandwich* and *Worst seat* which undoubtedly represent better conceptual extractions.

Although the approach performed well in combining both techniques there where some cases of misclassification. In most cases these occurred due to overgeneralisation both of Saxon rules and the T-Rex data model. Instances such as *life* in *"I have never seen anything like this in my life"* and *whole new meaning* in *"A whole new meaning for drive through"* were wrongly extracted as objects by either Saxon or T-Rex or a combination of both at times. Further to this, occasional entities such as *lines* in *"The people were lines up like crazy to get into this place"* cluttered the extraction set without adding any semantic value to it.

The issue of useful instances being overlooked by both systems can be partly attributed to part of speech misclassifications in descriptions such as *"Artwork! Sculptures in the sea at Crosby"*, *"Lifeboat on car ferry to France"*. In all descriptions, the references to objects of relevance within the photo (i.e. artwork, sculptures and lifeboat) are contextually difficult to be classified as objects (i.e. nouns instead of proper nouns), since their linguistic context also lends itself to other interpretations.

Finally, cases where there is not enough linguistic content for performing extractions using only machine learning and rules or a hybrid approach are exemplified by descriptions such as *"Mull"* and *"French Riviera"*. Unless such noun phrases were already part of a pre-compiled gazetteer, the lack of a sentence structure surrounding such examples makes it very difficult to tackle the IE problem from a purely NLP perspective.

6.2 Semantic propagation

The evaluation of semantic propagation proposed in Section 5.1.1 involved propagating location captures both from image descriptions as well as image tags to unannotated photographs. A small subset of 50 contributors to the main corpus was randomly selected and the evaluation was performed on a selection of each user's tagged/captioned images that reached a total of 19K images for the evaluation dataset. By only selecting images that had associated textual metadata the evaluation could be carried out automatically.

Each user's collection was processed using the semantic capture approach presented previously and then clustered using either the temporal or visual features separately and in conjunction with each other. For each cluster, one image would be randomly selected as the seed of semantic metadata and any location metadata associated with it would be propagated uniformly to all members of the same cluster. After running the experiments for each clustering strategy, the results shown in Fig. 8 were obtained.

It must be noted here that during the evaluation any one element clusters generated were discarded. The precision of propagations was calculated by obtaining

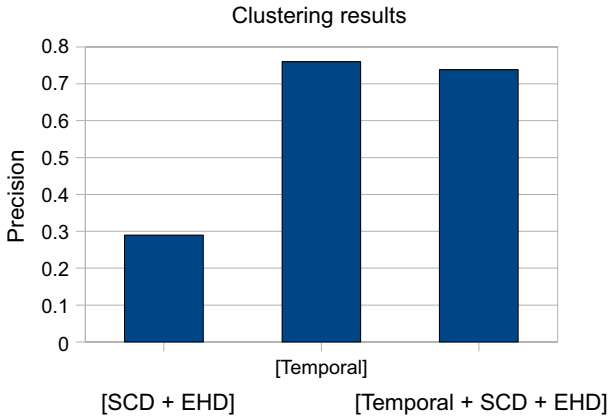

Fig. 8 Evaluation results for semantic propagation approach

the number of correctly propagated concepts (i.e. location) and dividing it by the total number of photographs in a cluster.

$$precision = \frac{number\ of\ correct\ propagations}{number\ of\ images\ in\ a\ cluster} \tag{3}$$

In order to obtain the number of correct propagations the metadata propagated from the seed image was compared to the existing metadata for the other photographs in the same cluster (i.e. tags or extractions from descriptions).

To the best of our knowledge, there exists no gold standard for the domain of personal photography, and due to the difficulty in obtaining an extensive set of manually annotated photographs, the average size of clusters will be presented alongside the precision for each feature set in order to contrast the precision measures obtained with how the model fits the dataset.

From Fig. 8 it can be seen that clustering photographs visually and performing a uniform propagation of tags across clusters achieved very low precision with the mean precision across the collections of 50 users being 29.3% and a σ (standard deviation) of 22% where 84% of values lie within 1 σ of the mean. It was found that the results obtained in this scenario were caused by visually similar images albeit semantically different being clustered together thus emulating the example from Fig. 1. In this case, all 19K photographs were partitioned into a total of 5,171 clusters that held an average of approximately 4.5 images per cluster. Figure 9 demonstrates that, although excluded from the precision calculation, just under 35% of the clusters were found to contain one element.

When clustering items temporally only, the results were clearly better where the approach achieved a mean precision over the collections of all 50 users of 76.4% with a σ of 13.7% where 67.4% of values lie within 1 σ of the mean. The drawback of using temporal features alone was that visually similar photographs that were captured at distinct times (e.g. visits to London at different times of the year) were not clustered together, thus affecting the reach of the semantic propagation strategy. The 19K photographs were partitioned into 3,756 clusters with approximately 5.5 images per cluster and Fig. 9 shows that similarly to the visual feature set 35% of clusters contained one element.

On the other hand, despite the attempt to maximise the semantic proximity of photographs by clustering images according to the temporal + visual strategy, less encouraging results were obtained. Following this strategy lead to a mean precision of 73.8% over the collections of all 50 users with a σ of 15.4% where 68.5% of values

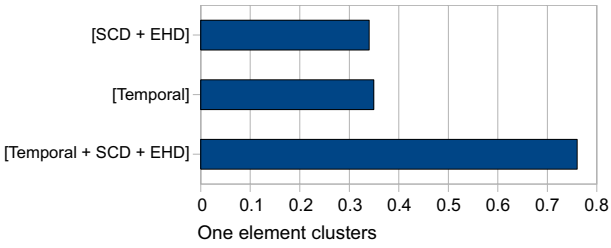

Fig. 9 One element clusters for each feature set

lie within 1 σ of the mean. It has been observed, however, that creating visual clusters within each temporal cluster has lead to a massive overfitting of the data model over the evaluation dataset, which resulted in an increase of the overall number of clusters to 11,919 and a consequent decrease in the average cluster size to approximately 1.5 images per cluster. This caused many photographs to be isolated from others in the same temporal cluster, aggravating the reach of the semantic propagation strategy even further. Figure 9 confirms this and demonstrates the large increase in one element clusters that renders this feature set unusable for the propagation of location semantics.

What can be seen from the results above is that although the visual similarity between images can be used as a good indicator of semantic meaning, understandably, the same assumption cannot be generalized to all cases in the domain of personal photography. However by using the temporal proximity between two photographs, semantic concepts such as *location* can be propagated to other images with promising accuracy. One issue that must be highlighted here however is that even by using the timestamp of photographs for creating contextually focused clusters, there seems to be a slight overfitting of the model over the dataset. Although better temporal clustering techniques could be used for this task, the results could have also been affected by the nature of the dataset collected from Flickr where users do not tend to upload their entire photographic collection, but rather a small sample of it for the purposes of sharing.

The results also suggest that while the approach is certainly valid for improving the quality and quantity of metadata for personal photographs, it still relies on user input, albeit to a much reduced degree. Contextualising photographs by using temporal clustering could help focus users' efforts in annotating their collections where it matters by propagating the semantics of any input across the entire collection.

6.3 Conclusion

In this paper, we have detailed a hybrid approach for extracting information from image descriptions that takes advantage of the combined results produced by systems that implement widely used techniques for IE. More specifically we considered the combination of T-Rex, a machine learning framework, and Saxon, a rule based extractor, for addressing issues of computational cost as well as precision and recall when extracting information from such short snippets of text.

Furthermore, in order to overcome the scarcity of semantic metadata in this domain, we have proposed and evaluated a strategy for propagating location semantics from annotated photographs to unannotated ones in the same collection. The strategy takes advantage of the temporal proximity between adjacent images for determining whether they depict the same location.

6.4 Future work

As seen in the evaluation results, the use of a hybrid approach for extracting information from image descriptions is promising, however levels of precision and recall could be improved by using external knowledge for reinforcing the extractions. For instance, cases such as in the description *"Highland near Ben Nevis"* could be

placed in the context of the user (e.g. does s/he know anyone called "Ben Nevis"?), the image itself (e.g. GPS positioning) or other image descriptions within the same collection (e.g. "Ben Nevis" was previously classified as a location/person). Another possible refinement to the approach, that has been previously applied with success in the past for the task of image annotations [1], is that of involving the user in the process for reinforcing system decisions, such as confirming the outcome of a conflict resolution.

Furthermore, some extraction examples, such as in the description *"Vicky and dad at local bus stop"* where *local bus stop* is extracted as an object, suggest that certain concepts may need further refinement. This would allow in this case for the object instance found to be also assigned geographic properties, given the contextual information about the image. Also, the concepts used here are an incomplete list of those useful within an image description. One important area for future work is extraction of further concepts used by people to describe their images (e.g. time, events, mood, etc). The same applies for the propagation strategy and in order to propagate other concepts to unannotated images a future study will have to take a closer look at how each such concept behaves according to time or other features (e.g. visual or others). For instance it is expected that for propagating event information (e.g. birthday, barbecue, party, etc.) a similar approach to that for propagating location will be used, but it is expected that visual features will play a bigger role in this case where it is important to detect the presence of the same people and the same visual background. For propagating instances of objects or people however it is expected that time will play a smaller role than visual features.

Acknowledgements This work was sponsored by Kodak Limited. We would also like to thank the 391 online photo sharing users who donated their photographs and respective metadata.

References

1. Ahn L, Dabbish L (2004) Labeling images with a computer game. In: CHI '04. ACM, New York, pp 319–326
2. Alp Aslandogan Y, Thier C, Yu CT, Zou J, Rishe N (1997) Using semantic contents and wordnet in image retrieval. In: SIGIR '97: proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 286–295
3. Bang HY, Zhang C, Chen T (2004) Semantic propagation from relevance feedbacks. In: 2004 IEEE international conference on multimedia and expo. ICME '04, 27–30 June. vol 1. IEEE, Piscataway, pp 81–84
4. Barla A, Odone F, Verri A (2003) Old fashioned state-of-the-art image classification. In: Proc. of ICIAP 2003, Mantova, 17–19 September 2003, pp 566–571
5. Budura A, Michel S, Cudre-Mauroux P, Aberer K (2008) To tag or not to tag ? harvesting adjacent metadata in large-scale tagging systems. In: The 31st annual international ACM SIGIR conference, 20–24 July 2008, Singapore
6. Cooper M, Foote J, Girgensohn A, Wilcox L (2003) Temporal event clustering for digital photo collections. In MULTIMEDIA '03: proceedings of the eleventh ACM international conference on multimedia. ACM, New York, pp 364–373
7. Duda RO, Hart PE (1973) Pattern classification and scene analysis. Wiley, New York, pp 98–105
8. Frohlich D, Kuchinsky A, Pering C, Don A, Ariss S (2002) Requirements for photoware. In: CSCW '02: proceedings of the 2002 ACM conference on Computer supported cooperative work. ACM, New York, pp 166–175

9. Greenwood M, Iria J, Ciravegna F (2008) Saxon: an extensible multimedia annotator. In LREC' 08: proceedings of the 6th international conference on language resources and evaluation, Morocco, May 2008
10. Hare JS, Lewis PH (2005) Saliency-based models of image content and their application to auto-annotation by semantic propagation. In: Multimedia and the semantic web/European semantic web conference 2005, Heraklion, 29 May–1 June 2005
11. Hartigan JA, Wong MA (1979) A K-means clustering algorithm. Appl Stat 28:100–108
12. Hearst M (1992) Automatic acquisition of hyponyms from large text corpora. In: Proc. of COLING 1992, 23–28 August 1992, Nantes, pp 539–545
13. Iria J, Ireson N, Ciravegna F (2006) An experimental study on boundary classification algorithms for information extraction using svm. In: Proc. of EACL 2006, Montreal, 22–27 April 2006
14. Keyvanpour M, Asbaghi S, Fathy M (2007) A new scheme of automatic semantic propagation in the image data base using a hierarchical structure of semantics. In: DEXA '07: proceedings of the 18th international conference on database and expert systems applications. IEEE Computer Society, Washington, DC, pp 59–63
15. Manjunath B (2001) Color and texture descriptors. IEEE Trans Circuits Syst Video Technol 11:703–715
16. Miller AD, Edwards KW (2007) Give and take: a study of consumer photo-sharing culture and practice. In: CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, New York, pp 347–356
17. Naaman M, Harada S, Wang Q, Garcia-Molina H, Paepcke A (2004) Context data in geo-referenced digital photo collections. In: Proc. of ACM MM, October 2004
18. Park DK, Jeon YS, Won CS (2000) Efficient use of local edge histogram descriptor. In: MULTIMEDIA '00: Proceedings of the 2000 ACM workshops on Multimedia. ACM, New York, pp 51–54
19. Pastra K, Saggion H, Wilks Y (2002) Extracting relational facts for indexing and retrieval of crime-scene photographs. In Macintosh A, Ellis R, Coenen F (eds) Applications and innovations in intelligent systems X, British Computer Society Conference Series. Springer, Heidelberg, pp 121–134
20. Pelleg D, Moore A (2000) *X*-means: Extending *K*-means with efficient estimation of the number of clusters. In: Proc. 17th International Conf. on Machine Learning. Morgan Kaufmann, San Francisco, pp 727–734
21. Salembier P, Sikora T (2002) Introduction to MPEG-7: multimedia content description interface. Wiley, New York
22. Spyrou E, LeBorgne H, Mailis T, Cooke E, Avrithis Y, O'Connor N (2005) Fusing MPEG-7 visual descriptors for image classiffication. In: Duch W, Kacprzyk J, Oja E, Zadrozny S (eds) In: Artificial neural networks, part II: formal models and their applications, vol 3697. Springer, Heidelberg, pp 847–852
23. Srihari R (1995) Automatic indexing and content-based retrieval of captioned images. Computer 28(9):49–56
24. Stauder J, Sirot J, Le Borgne H, Cooke E, O'Connor NE (2004) Relating visual and semantic image descriptors. In: EWIMT 2004—European workshop on the integration of knowledge, semantics and digital media technology
25. Veltkamp R, Tanase M (2000) Content-based image retrieval systems: a survey. Technical report UU-CS-2000-34, Dept. of Computing Science, Utrecht University
26. Zhang D, Tsotras VJ (2001) Improving min/max aggregation over spatial objects. In: ACM-GIS, pp 88–93
27. Zhang H, Chen Z, Li M, Su Z (2003) Relevance feedback and learning in content-based image search. World Wide Web 6(2):131–155

Rodrigo F. Carvalho is a Ph.D. student at the University of Sheffield, UK. His research interests lie in the application of contextual and social information for enhancing image related metadata with the intent of improving future retrieval and sharing of photographic resources. He has worked previously on European and commercial research projects targeted towards the extraction of information for use in emergency response scenarios and for the management of personal photographic memories.

Sam Chapman is a senior Research Associate at the University of Sheffield, UK. His research investigates cutting edge semantic technology to facilitate knowledge processes across large organisations with a focus upon search, acquisition and integration of knowledge from various media. He works on a number of european, national and commercial research projects concerning the needs of aerospace, historical research, archaeology support, personal photographic memories and emergency response amongst others. He is also the Director of Technology for Knowledge Now Ltd where he commercialises knowledge acquisition and query technologies to aid a wide variety of industries.

Fabio Ciravegna is Professor of Language and Knowledge Technologies at the University of Sheffield. He is Director of the European Integrated Project IST X-Media (www.x-media-project.org), and principal investigator in several European and National projects. He coordinates industrial projects funded by Rolls-Royce plc, Kodak Eastman and Lycos Europe. He is member of the editorial board of the International Journal on "Web Semantics" and of the "International Journal of Human Computer Studies". Fabio is general chair of the 6th European Semantic Web Conference (2009) (http://www.eswc2009.org/). He is director of K-Now Ltd, a spin-off company supporting dynamic distributed communities in large organizations. He holds a Ph.D. from the University of East Anglia and a doctorship from the University of Torino, Italy.