QMUL @ MediaEval 2012: Social Event Detection in Collaborative Photo Collections

Markus Brenner, Ebroul Izquierdo
School of Electronic Engineering and Computer Science
Queen Mary University of London, UK
{markus.brenner, ebroul.izquierdo}@eecs.qmul.ac.uk

ABSTRACT
We present an approach to detect social events and retrieve associated photos in collaboratively annotated photo collections as part of the MediaEval 2012 Benchmark. We combine data of various modalities from annotated photos as well as from external data sources within a framework that has a classification model at its core. Experiments based on the MediaEval Social Event Detection Dataset demonstrate the effectiveness of our approach.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms
Design, Experimentation, Performance

Keywords
Benchmark, Photo Collections, Event Detection, Classification

1. INTRODUCTION
The Internet enables people to host and share their photos online, e.g. through websites like Flickr. Collaborative annotations and tags are commonplace on such services. The information people assign varies greatly but often seems to include some sort of references to *what* happened *where* and *who* was involved. In other words, such references describe observed experiences or occurrences that are simply referred to as *events* [7]. In order to enable users to explore, retrieve and associate such events in their photo collections or on online services, effective approaches are needed to both detect events and retrieve corresponding photos. The MediaEval Social Event Detection (SED) Benchmark [4] provides a platform to compare different such approaches.

1.1 Background
There is much research in the area of event detection in web resources in general. The subdomain we focus on is photo websites, where users can collaboratively annotate photos. Recent research like [1] put emphasis on detecting events from Flickr photos by primarily exploiting user-supplied tags. [6] and [5] extend this to place semantics, the latter incorporating the visual similarity among photos as well. Our aim, however, is to also use information from external sources to find photos corresponding to the same events. [2] is an example that goes further in our direction by exploiting Wikipedia classes.

1.2 Objective
In this paper, we outline a framework (we give details in [8]), where we utilize external sources to detect social events and retrieve associated photos in collaborative photo collections. We test our approach against one of the three challenges laid out by the MediaEval 2012 SED Benchmark: The goal of Challenge II relates to soccer events taking place in two given cities.

The remainder of this paper is structured as follows: In the next few sections, we explain what external data we utilize how, and then we outline the design of our classifier-based framework. We finish by presenting our benchmark results and conclusions.

This work is partially supported by EU project CUbRIK.

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

2. GATHERING EXTERNAL DATA
2.1 Expanding the Topic
Social events often revolve around a topic like festivals or sport events. Imagine a set of collaboratively tagged photos on a photo website. Users assign keywords that might relate to the topic of the scene depicted in the photo, but they do not adhere to a controlled vocabulary. To account for this fact, we expand the textual representation of a given topic (e.g. *concert* by *festival*, *gig*, etc.). We do this through a combination of WordNet and DBpedia based on some of our own initial evidence (a few commonly associated terms with the topic).

2.2 Handling Geographic Locations
The venue location of a social event is an important component and indicator where an event happened. For that reason, we gather location-centric information like suburb, region and the geographic coordinates for each venue. Thus, we can later match geo-tagged photos against venues. We implement the automatic lookup through the Google Geocoding API service.

Whereas the above expands the query venue, we also wish to identify and understand any textual annotations in photos that refer to geographic locations (e.g. *Norway*). We use this information later in the retrieval process to isolate photos that do not likely correspond to the venue of a queried event. For practical reasons we limit ourselves to countries and larger cities extracted from the public GeoNames database.

2.3 Specific Extension: Soccer Matches
Our strategy for detecting soccer matches or events is to first find all soccer clubs and associated stadiums for the given cities in the challenge query. We automatically retrieve this information from DBpedia by means of the SPARQL interface. For each soccer club, we also gather its club- and nickname. Similarly, we request alternative names for the stadiums. Note that we discard stadiums (and thus clubs) with a capacity of less than 15000 people.

3. PREPROCESSING
3.1 Matching Geographic Locations
As geo-tagged photos become more and more popular, we can identify photos as belonging and *not* belonging to a venue (and ultimately an event when also considering the date and time). For each venue, we compile two sets of photos: photos that lie within and photos that lie outside a venue's relaxed bounds.

3.2 Translating Terms and Stop-words
Photos are shared and accessed across geographical and cultural boundaries. To factor this in, we translate stop-words (that we

introduce next) and the topic-related terms compiled beforehand into other languages. We limit ourselves to the languages prevailing in the countries in which the query venues are located. We retrieve the translations via the Google Translate API.

3.3 Composing Textual Features

We compose text features of each photo's title, description and keywords. During the training stage (detailed later in Section 4), we also include the list of expanded topic terms as well as any available event or venue information gathered in Section 2.

Then, we apply a Roman preprocessor that converts text into lower case, strips punctuation as well as whitespaces and removes accents. It also eliminates common stop-words like *and*, *cannot*, *you* etc. Moreover, we discard all words that are less than three characters in length. We also ignore numbers and terms commonly associated with photography (e.g. *Nikon*). Finally, photos with less than two words overall are filtered out.

In the next step, we split the words into tokens. The text assigned to photos by users on online services such as Flickr is often not *clean*: Words have spelling errors and different suffixes and prefixes. Furthermore, traditional natural language processing steps, e.g. word-stemming, are often tailored to the English language. To accommodate other languages, we do not apply a word-based tokenizer but a language-agnostic character-based tokenizer. We also take all preprocessed words in their full and non-tokenized form into account.

We then use a vectorizer to convert the tokens into a matrix of occurrences. To make up for photos with a large amount of textual annotations, we also consider the total number of tokens. This approach is commonly referred to as Term Frequencies (TF).

4. EVENT DETECTION AND RETRIEVAL

The MediaEval Benchmark defines an event as a distinct combination of date and location. In the simplest case, one could start from a list of all suitable date-venue combinations. In our framework though, we first narrow down the list of candidate events based upon a temporal clustering process that discards clusters with few photos. If we retrieve multiple photos (as explained next) that match a venue's location but do not fall into any of that venue's known or already detected events, we consider them as part of another *new* event.

For actual retrieval, we employ a Linear Support Vector Classifier [3]. For each event, we train a separate classifier. Basically, we perform binary classification: photos which are either related or not related to an event (including its location). However, we also introduce a third class that reflects events of the same topic to potentially improve performance. We can utilize the non-relating class to include features of other topics as well. Note that we aggregate all textual terms into single samples, as it performs better than considering multiple samples (with the same class label).

5. EXPANDING FEATURE SPACE

We use an iterative two-step process to expand the feature space of the training data based on query information alone. We accomplish this by first training an initial classifier on the few query terms available, and then compiling a new list of textual terms based upon the predicted outcome over all applicable photos. Finally, we use these gained terms to refine our initial query terms.

6. LIMITING SEARCH AND PRUNING

In general, the date and time a photo was captured are effective cues to bound the search space. Therefore, for each event's prediction step, we consider only those photos that lie within the event's temporal search window. Unless the latter is specified by the query, we estimate it through a temporal clustering scheme. At this stage, we also exclude all photos that refer to a location other than that of the event's associated venue. Lastly, we introduce a pruning-step (following the retrieval stage) based on visual features extracted from photos. In particular, we extract MPEG-7 color and texture features to train (using photos relating to the same venue) an additional classifier that we utilize in a similar fashion as before.

7. EXPERIMENTS AND RESULTS

We perform experiments on the MediaEval 2012 SED Dataset that consists of 167.332 Flickr photos with accompanying metadata. First, though, we use the MediaEval 2011 SED Dataset (with available ground truth) to estimate suitable parameter values for the overall framework, including the classifiers.

For Challenge II, we identify multiple soccer stadiums (and clubs) for each given city. We find several thousand geo-tagged photos not associated with any venue, thus substantially reducing the search space while providing a large amount of training samples for each event's non-relating class.

In the following table, we present our test results (as evaluated by the organizers of the MediaEval Benchmark). We notice that discarding unlikely candidate events most notably improves performance. However, while visual pruning only leads to slight gains in precision, both including features of additional topics and feature expansion do not show much effect in the benchmark. Additional tests (based upon the 2011 SED Dataset) reveal that both show only notable gains when training features are sparse. That is the case when there are only few geo-tagged photos.

Table 1: Results depending on configuration

	P	R	F-score	NMI
Default configuration	79.0	67.1	72.6	0.65
With features of another topic	79.1	67.0	72.5	0.65
With feature expansion	79.0	66.9	72.5	0.65
With basic event detection	56.0	69.6	62.0	0.53
With visual pruning	83.2	61.9	71.0	0.63

8. CONCLUSION

We present an approach to both detect social events and retrieve associated photos in tagged photo collections such as Flickr. We combine external information with data extracted from photos to train a classifier. The listed benchmark results validate our approach. In the future, we wish to additionally detect events from the photos' annotations and from social networks like Twitter.

9. REFERENCES

[1] Chen, L. and Roy, A. 2009. Event detection from Flickr data through wavelet-based spatial analysis. *ACM CIKM*.

[2] Firan, C.S. et al. 2010. Bringing order to your photos: Event-driven classification of Flickr images based on social knowledge. *ACM CIKM*.

[3] Keerthi, S.S. et al. 2008. A sequential dual method for large scale multi-class linear SVMs. *ACM KDD*.

[4] Papadopoulos, S. et al. 2012. Social Event Detection at MediaEval 2012: Challenges, Dataset and Evaluation. *MediaEval Workshop*.

[5] Papadopoulos, S. et al. 2010. Cluster-based landmark and event detection on tagged photo collections. *IEEE Multimedia*.

[6] Rattenbury, T. et al. 2007. Towards automatic extraction of event and place semantics from Flickr tags. *ACM SIGIR*.

[7] Troncy, R. et al. 2010. Linking events with media. *I-Semantics*.

[8] Brenner, M. and Izquierdo, E. 2012. Social Event Detection and Retrieval in Collaborative Photo Collections. *ACM ICMR*