MediaEval 2013: Social Event Detection, Retrieval and Classification 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 a framework to detect social events, retrieve associated photos and classify the photos according to event types in collaborative photo collections as part of the MediaEval 2013 benchmarks. We incorporate various contextual cues using both a constraint-based clustering model and a classification model. 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 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 sorts of references to *what* happened *where* and *who* was involved. In other words, such references describe observed experiences that are planned and attended by people, which we simply refer to as *events* [1]. In order to enable users to exploit events in their photo collections or on online services, effective approaches are needed to detect events and retrieve corresponding photos, and additionally, to understand event types. The MediaEval Social Event Detection (SED) Benchmark [2] provides a platform to compare different such approaches.

2. BACKGROUND AND RELATED WORK
There is much research in the area of event detection in web resources in general. The subdomain we focus on is photo websites, wherein users can share and collaboratively annotate photos. Recent research [3] put emphasis on detecting events from Flickr photos by primarily exploiting user-supplied tags. Other works [4], [5] extend this to place semantics, the latter incorporating the visual similarity among photos as well. Our framework relates to event clustering approaches, particularly in personal photo collections [6]. However, we also embody the context of social events to improve detection and retrieval performance. We believe that further understanding and research are needed on how to best exploit and process the information collaborative photo collections hold in the context of social events.

This work is partially supported by EU project CUbRIK.

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

3. OBJECTIVE AND APPROACH
In this paper, we outline a framework that builds upon and extends our previous works [7] and [8], and where we detect social events and retrieve associated photos in collaborative photo collections. Moreover, we classify photos according to event types such as *music concerts* or *sport games*. We test our approach against both challenges laid out by the MediaEval 2013 SED Benchmark: the goal of Challenge I relates to detecting social events and retrieving associated photos, and the goal of Challenge II relates to classifying photos according to event types.

3.1 Preprocessing: Propagating Locations
The most useful information to us with respect to social events are: involved people (based on the username of the person who uploaded the photos); date and time (the photos are captured); and the geographic location (venue) an event takes place. Our reasoning for this is the assumed constraint that photos sharing the same involved people, date and time as well as geographical location shall belong together to the same event. Likewise, photos that differ in at least one constraint shall *not* belong together. Thus, we extract, propagate and incorporate as much information from these three domains as possible. While date and time as well as usernames (involved people) are usually available, the geographic location is often unavailable (for example, only newer smartphones embed location coordinates). As detailed in our previous paper [8], we take advantage of this constraint in a preprocessing step to propagate geographic locations across a photo collection based on some photos that include geographic coordinates or textual references such as *Barcelona*.

3.2 Feature Extraction
To aid event detection, retrieval and classification as explained in the forthcoming two sections, we extract and compose textual features of each photo's title, description and keywords. First, we apply a Roman preprocessor that converts text into lower case, strips punctuation as well as whitespaces and removes accents. In the next step, we split the words into tokens. To accommodate other languages as well as misspelled or varied terms, we apply a language-agnostic character-based tokenizer rather than a word-based tokenizer. 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*. Instead of decomposing the resulting feature matrix, we simply limit the amount of features to 9600, which results in almost comparable performance at much lower required complexity.

In addition to textual features, we also extract and incorporate visual GIST features (a feature vector with 960 elements) for each photo. To fuse textual and visual features, we normalize both features and concatenate them into a combined feature vector. We

also incorporate a weighting ratio that allows us to emphasize one or the other feature.

3.3 Event Detection and Retrieval

We define an event as a distinct combination of a spatial window (5km clusters) and a temporal window (8h clusters). We start with a list of all suitable spatio-temporal window combinations (the results of Section 3.1). If we retrieve more than two photos, as we explain next, we consider the combination as a detected event and the retrieved photos as part of that event.

For actual retrieval, we first include all photos whose date, time and available geographic coordinates fall into an event's spatio-temporal window (we denote these candidate photos as X_C). Thereafter, we employ a Linear Support Vector Classifier (using the features whose extraction we explain in Section 3.2) for all remaining photos that only fall into an event's temporal window, but whose spatial window we are not aware of. For each event, we train a separate model and perform binary classification: photos which are either related or not related to an event. We use X_C for the *related* class, and a small, random subset of photos (that do not fall within the same spatio-temporal window) for the *not-related* class.

In this last step of our event-driven retrieval framework, we include photos that are likely relevant to a retrieval query but may have been mistakenly discarded by the classification step. In particular, these might be photos that are linked to users who have multiple photos relevant to a retrieval query. The assumption is that if a user attends a social event and takes photos, then it is likely that most of his photos taken over the time that he attends the event are of the event.

3.4 Classification of Event Type

In this section, we extend our framework to classify the event type that a photo or multiple photos belong to. We perform the same initial constraint-based spatio-temporal clustering as in Section 3.3. This allows us to compile a larger training set by including all photos of an event in case the training ground truth is only given for some photos of an event.

Using this extended overall training set, we then train a multi-class Linear Support Vector Classifier (as in Section 3.3, based on features that we extract in Section 3.2). In the simplest case, we can thereafter predict the event type of any given test photo. However, instead of treating any test photos separately, it is also possible to consider multiple photos (that belong to the same event) together. To do so, we simply assign the most often predicted event type within an event to all its associated photos.

4. EXPERIMENTS AND RESULTS

We perform experiments on the MediaEval 2013 SED Dataset that consists of a total of 437370 Flickr photos (Challenge I) and 57165 Instagram photos (Challenge II) with accompanying metadata. We use the provided training sets to estimate suitable parameter values and train our event classification model required for Challenge II.

In the following two tables, we present our results (as evaluated by the organizers of the MediaEval Benchmark) with respect to the testing sets. The results of Challenge I show us that it is important to consider temporal clusters (as *newly* detected events) that are not *clearly* associated with any geographic location (or spatial cluster). In our case, this improves the F1-score from 0.59 to 0.76. We also see that an additional classification-based expansion of an event's candidate set does not necessarily always improve detection and retrieval results, or does so only in conjunction with other steps. For example, if we consider and select only one spatial cluster per matching temporal window for each involved person (username), we can further improve results by a small margin.

For Challenge II, the results detail that we can better classify photos as non-events (F1-score of ~0.94) rather than as a specific event type. Of eight possible event types that we trained our model on, we can best classify the types *concert* (~0.52), *protest* (~0.37) and *theater-dance* (~0.31). We see the worst performance with *fashion* (~0.07) and *other* (~0.05). On average, we achieve an event classification F1-score of 0.50 in our best performing configuration. Surprisingly, neither training set expansion nor event-wide joint classification notably improves results.

Although we use the same feature extraction configuration for both challenges, the addition of visual features (compared to using only textual features) has a much larger positive impact for Challenge II than for Challenge I.

Table 1: Results of Challenge I depending on configuration

	F1	NMI
Run 1: Run 5 - visual features	0.78	0.94
Run 2: Basic	0.59	0.64
Run 3: Run 2 + include temporal clusters	0.76	0.94
Run 4: Run 3 + expansion	0.74	0.93
Run 5: Run 4 + include rest + max. user	0.78	0.94

Table 2: Results of Challenge II depending on configuration

	F1 Non-Event	F1 Event
Run 1: Without visual features	0.93	0.37
Run 2-5: Default	0.95	0.50

5. CONCLUSION

We present a framework to detect social events, retrieve associated photos and classify the photos according to event types in tagged photo collections such as Flickr. We combine various contextual information using a constraint-based clustering and classification model. The listed benchmark results validate our approach. In the future, we wish to improve event detection by incorporating information from social networks.

REFERENCES

[1] R. Troncy, B. Malocha, and A. T. . Fialho, "Linking events with media," in *I-SEMANTICS*, 2010, pp. 1–4.

[2] T. Reuter, S. Papadopoulos, V. Mezaris, P. Cimiano, C. de Vries, S. Geva, and C. De Vries, "Social Event Detection at MediaEval 2013: Challenges, Datasets, and Evaluation," in *MediaEval 2013 Workshop*, 2013, pp. 2–3.

[3] L. Chen and A. Roy, "Event detection from flickr data through wavelet-based spatial analysis," in *CIKM*, 2009, pp. 523–532.

[4] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from Flickr tags," in *SIGIR*, 2007, pp. 103–110.

[5] S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, and A. Vakali, "Cluster-based landmark and event detection on tagged photo collections," *MultiMedia*, no. 99, pp. 1–1, 2010.

[6] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox, "Temporal event clustering for digital photo collections," *TOMCCAP*, pp. 269–288, 2005.

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

[8] M. Brenner and E. Izquierdo, "Event-driven Retrieval in Collaborative Photo Collections," in *WIAMIS*, 2013.