University of Glasgow

Glasgow2014

ICMR

Proceedings of the

International Workshop on Social Events in Web Multimedia '14

In conjunction with the ACM Conference on Multimedia Retrieval '14

April 1st, 2014

Glasgow, UK

Workshop Chairs

Vasileios Mezaris (CERTH, Greece)

Raphael Troncy (Eurecom, France)

Georgios Petkos (CERTH. Greece)

Philipp Cimiano (University of Bielefeld, Germany)

MEDIA MIXER LINKEDTV social sensor

acm FXPAL sicsa* Springer SIGMM SIGIR AXES

Organization

Chairs

Vasileios Mezaris (Information Technologies Institute / CERTH)
Raphael Troncy (Eurecom)
Georgios Petkos (Information Technologies Institute / CERTH)
Philipp Cimiano (University of Bielefeld)

Program Committee

Giulia Boato (University of Trento)
Xavier Giro-i-Nieto (Universitat Politecnica de Catalunya)
Nikolaos Gkalelis (Information Technologies Institute / CERTH)
Lynda Hardman (CWI)
Michiel Hildebrand (University of Amsterdam)
Lyndon Kennedy (Yahoo! Research)
Paul Lewis (University of Southampton)
Erik Mannens (iMinds – University of Ghent)
Diana Maynard (University of Sheffield)
Francesco De Natale (University of Trento)
Symeon Papadopoulos (Information Technologies Institute / CERTH)
Timo Reuter (Universitat Bielefeld)
Lars Schmidt-Thieme (University of Hildesheim)
Stefan Siersdorfer (L3S Research Center)
Alean Smeaton (Dublin City University)
Thomas Steiner (Google)
Ruben Verborgh (iMinds – University of Ghent)
Maia Zaharieva (University of Vienna)
Qianni Zhang (Queen Mary, University of London)

Keynote talk: Mining Events from Multimedia Streams

Speakers: Jonathon Hare, Sina Samangooei

The aggregation of items from social media streams, such as Flickr photos and Twitter tweets, into meaningful groups can help users contextualize and effectively consume the torrents of information on the social web. This task is challenging due to the scale of the streams and the inherently multimodal nature of the information being contextualized. In this talk we'll describe some of our recent work on trend and event detection in multimedia data streams. We focus on scalable streaming algorithms that can be applied to multimedia data streams from the web and the social web. The talk will cover two particular aspects of our work: mining Twitter for trending images by detecting near duplicates; and detecting social events in multimedia data with streaming clustering algorithms. We will describe in detail our techniques, and explore open questions and areas of potential future work, in both these tasks.

Jonathon Hare is a Lecturer in the Web and Internet Science group at the University of Southampton. His research interests lie in the area of multimedia information mining, analysis and retrieval, with a particular focus on large-scale multimodal approaches. He has published nearly 60 papers in peer-reviewed conferences and journals.

Sina Samangooei is a Research Fellow in the Web and Internet Science Research group at the University of Southampton. His research interests include streaming data, multimedia retrieval and large-scale machine learning.

Keynote talk: Semantic Encodings for Recognizing and Recounting Video Events

Speaker: Cees Snoek

What defines an event in video? Answers from the recent literature indicate success can be obtained with a color Fisher vector, a histogram of motion and trajectories, or, even better, a potpourri of multimedia descriptors and representations. In this talk I will highlight our progress on encoding video, and events, by semantic detector predictions, which can not only recognize but also explain events. First I will present our study on the characteristics of a universal semantic encoding for arbitrary-event recognition in web video. Then I will introduce an algorithm that learns from examples what concepts in a semantic encoding are most informative per event. Finally, I will end by showing event recounting capabilities of the semantic encodings, which open up the possibility to automatically describe and explain why a particular video was found.

Cees G. M. Snoek is currently an associate professor at the University of Amsterdam. He was previously at Carnegie Mellon University and the University of California at Berkeley. His research interest is video and image search. Dr. Snoek is the principal investigator of the MediaMill Semantic Video Search Engine, which is a consistent top performer in the yearly NIST TRECVID evaluations. He is member of the editorial boards for IEEE Multimedia and IEEE Transactions on Multimedia, and general co-chair of ACM Multimedia 2016. Cees is recipient of an NWO Veni award, an NWO Vidi award, and the Netherlands Prize for ICT Research 2012. Several of his Ph.D. students have won best paper awards, including the IEEE Transactions on Multimedia Prize Paper Award and the SIGMM Best Ph.D. Thesis Award.

Multimodal Detection, Retrieval and Classification of Social Events in Web Photo Collections

Markus Brenner and Ebroul Izquierdo
School of EECS, Queen Mary University of London, UK
{m.brenner,e.izquierdo}@qmul.ac.uk

ABSTRACT

We present a framework to detect or cluster social events in web photo collections, retrieve associated photos and classify these photos according to event types. Compared to traditional approaches that often consider only textual or visual features without the notion of social events, our approach jointly utilizes both features while also incorporating other event-related contextual cues like date and time, location and usernames. Experiments based on the MediaEval Social Event Detection Dataset demonstrate the effectiveness of our combined constraint-based clustering and classification model.

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, Theory, Experimentation, Performance

Keywords

Context, Detection, Retrieval, Social Events, Collaborative Photo Collections

1. INTRODUCTION

The transition from traditional, film-based photography to digital photography has led to a situation where more consumers take many more photos than ever before. Similarly, the ways we can store and share photos have also changed. Nowadays, the Internet enables users to host, access and share their photos online; for example, through websites like Flickr and Facebook. Collaborative annotations and tags, as well as public comments, are commonplace on such services. The information users assign varies greatly but often seems to include references to *what* happened *where* and *who* was involved. Such references often describe observed experiences that are planned and attended by people, which we simply refer to as *events* [16].

It is desirable to exploit such events for many reasons. For example, studies have proven that users find it easier to search through photo collections if the photos are grouped into events (that can be more easily browsed) rather than only by their dates of capture [7]. It is also possible to link events in photo collections to public social media like online news feeds. Thus, aiding users in exploring photo collections and facilitating the *mesh-up* or linkage of web media reinforce why effective approaches are needed in the first place to detect or cluster events and retrieve their associated photos, and additionally, to understand the kind or type of those events.

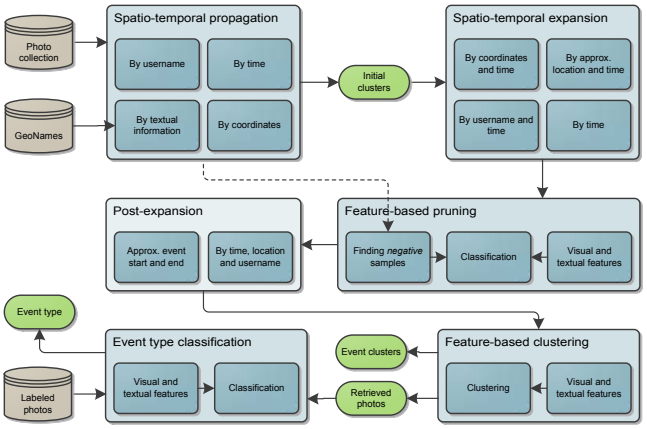

Figure 1: Overview of our proposed event-based detection, retrieval and classification framework.

2. BACKGROUND AND RELATED WORK

There is a wealth of research in the area of general event detection in web resources. Works like [1,17] study event detection in social media, particularly in social online networks such as Facebook or Twitter. They share some aspects with our work but do not focus on the photo domain, as we do. More specifically, we target photo websites, where users can collaboratively annotate photos. Since we wish to retrieve photos relating to social events (thus also require the detection of such events), our task is different from generic event clustering approaches used in personal photo collections [6] that do not embody the context of social events. Research

ICMR 2014 SEWM Workshop, Glasgow, Scotland
.

that target collaborative photo collections such as [4] often focus on exploiting user-supplied tags without considering many of the other modalities available, especially those related to events. [9] targets photos and events but leverages only the association patterns between generic activities and their geographical locations.

Works [2, 12, 13] focus on events and combine more varied semantic information, such as the spatial-temporal domain in relation to users (photographers) uploading photos. Of these works, however, only [12] and [2] consider visual similarities among photos. While [12] classifies events, [2] does not classify events and instead emphasizes external semantic data. Our prior work [3] also focuses on social events in photo collections, but it is limited to photo retrieval and does not involve the detection and classification of social events.

For the benefit of event detection and event-driven photo retrieval (especially when linked to social events), further research is needed on how to best exploit and process the information collaborative web photo collections hold.

3. OBJECTIVE AND APPROACH

We present a framework (overview in Figure 1) to detect and cluster social events and retrieve associated photos in photo collections. In particular, we target collaborative web photo collections (such as Flickr) which contain photos with rich but uncontrolled annotations and that are linked to their users. Moreover, we show how to classify photos and events according to event types like *music concerts* or *sports games*.

The foremost domains defining a social event are date and time, location, involved people and their observable activities [16]. Note, however, that we primarily target social events that are public and attended by many people because these events are likely to be better represented across popular social media websites and channels. We do not consider private events such as a single person's vacation. Also, note that our proposed work is generally different from traditional photo retrieval frameworks, which are often purely based on image content, as we incorporate the notion of social events. In the usage-scenario that we envision, all social events within a photo collection are *automatically* detected and their associated photos *automatically* retrieved – both without requiring any user knowledge or interaction. Additionally, these photos or events are *automatically* classified according to event type.

The remainder of this paper is structured as follows: In the next two sections, we describe an initial spatio-temporal clustering procedure to increase the amount of location-aware photos, and we also describe our methods to extract textual and visual features from photos. Thereafter, we explain how we detect and cluster social events and retrieve the photos that are associated with those events. We devote the subsequent section to event type classification. Lastly, we detail our experiments and evaluate their results.

3.1 Spatio-Temporal Propagation

Although collaboratively annotated photos may provide several information domains, the most useful to us with respect to social events are: involved people (based on the username of the person who uploads the photos); date and time of photo capture; and the geographical location (venue) an event takes place. Our reasoning for this is the assumed

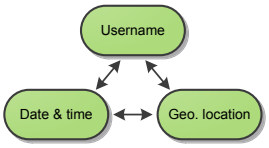

Figure 2: Due to our assumed constraint, photos sharing the same username, date and time and geographical location shall belong to the same social event. Likewise, photos that differ in at least one domain shall not belong together.

constraint that photos sharing the same involved people, date and time and geographical location shall belong to the same event. Likewise, photos that differ in at least one domain shall not belong together. Thus, we extract, propagate and incorporate as much information from these three domains as possible.

Analyzing photos to determine which people are depicted and thus involved in a social event is difficult, especially when people are not known beforehand. However, it is a valid assumption that the users who upload and share photos are the people involved. Since photo services like Flickr use unique identifiers or usernames for their users when they upload and share photos, we are able to associate each photo with a user.

Almost all photos are shot with digital cameras or *smartphones* nowadays. As such, the date and time of capture is automatically embedded into photos. However, this is often not yet the case with location, as many cameras still lack the capability of determining geographical location. Camera-equipped *smartphones* usually offer this capability but cannot provide and embed such location information at all times; for instance, GPS satellite signals within or below buildings are often too weak to fix the location. To determine the location of as many photos as possible, we propagate location information from photos that embed it to photos that do not. We take advantage of the constraint that a person cannot be at multiple locations at the same time. We relax the constraint by linking it to a particular temporal duration. For each user u, we compile two sets of photos: set X_L with photos that provide their location, and set $X_{\bar{L}}$ with photos that do not. Then, we find for each photo $x \in X_{\bar{L}}$ a set of photos $X_{\check{L}}$ out of X_L whose time difference is below a threshold d_x, and assign the location most often associated with photos within $X_{\check{L}}$ to $X_{\bar{L}}$.

Additionally, we analyze each photo's textual annotation t (e.g. title, keywords, etc.) for references to geographical locations (e.g. *London*). We first compile a list of geographical locations (where each list item represents the name of a location and its geographic coordinates). Theoretically, we would need to consider all possible geographical locations and places worldwide for optimum coverage. However, we limit ourselves to larger cities that we extract from the publicly available GeoNames dataset. We train a Linear Support Vector Classifier (that scales well with a high number of classes) on the compiled list, and thereafter predict the location of each photo based on t. Then, we compute and average text edit distances (based on Jaro-Winkler [5]) among prediction and consecutive word token combinations within t to discard predictions that fall below a certain probability threshold. Employing a classifier that directly emits prob-

abilities is another option. For all approximate locations $\tilde{L}$ (the coordinates of the city centers) that we determine, we apply the same username-based location propagation procedure as mentioned before. This time, however, we perform some additional filtering by only considering those approximate locations $\tilde{L}$ associated with each user if they are also associated with at least n_a photos. We signify the gained set of photos with approximated locations as $X_{\tilde{L}}$. Lastly, we compile a combined set S that includes all spatio-temporal clusters of $X_{\bar{L}}$ and $X_{\tilde{L}}$. We also compile a combined set T that includes all temporal clusters of $X_{\bar{L}}$ and $X_{\tilde{L}}$.

3.2 Textual and Visual Feature Extraction

To aid event detection, retrieval and classification as explained in the forthcoming two sections, we extract and compose textual features from each photo's title, description and keywords. First, we apply a Roman preprocessor that converts text into lower case and strips punctuation, whitespaces and accents. Next, we split the words into tokens. To accommodate multiple languages as well as misspelled and varied terms, we apply a language-agnostic character-based tokenizer (limited to windows of three to six characters within word boundaries) rather than a typical word-based tokenizer. We then use a vectorizer to convert the tokens into a matrix of occurrences. To account for photos with a large amount of textual annotations, we also consider the total number of tokens by ignoring tokens that appear *often* throughout documents. This approach is commonly referred to as *Term Frequencies*. Instead of decomposing the resulting feature matrix to one with a lower and fixed number of dimensions, we limit the amount of features to a fixed value that corresponds to the n_f most frequent terms. According to initials tests, this results in almost comparable performance at much lower required complexity.

In addition to textual features, we also capture and incorporate the scene or *gist* of photos in terms of visual attention (e.g. color and texture). For every photo, we extract GIST signatures [11] of the patches of a 4×4 grid *spanned* over the photo and compose a final feature vector with 960 elements. 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

Let X be the entire set of photos within a dataset and let an event be a distinct combination of a spatial and temporal window or cluster (e.g. $5km$ and $8h$ as in our later experiments). We start with the list of spatio-temporal clusters S that is the result of Section 3.1. We consider each spatio-temporal cluster as a detected event if we can associate at least n_d photos with it. In each case, we consider these associated photos (we denote them as X_C) as *belonging* to an event. Together, they form our initial retrieval result that we expand in the next steps.

3.3.1 Expansion and Feature-based Pruning

We propose a supervised approach to retrieve any remaining photos that only fall into an event's temporal window (we signify the set of these photos as X_E), but whose spatial window we are not aware of. Our intention is to classify the photos of X_E as belonging to an event (we define this resulting set as X_P) or not belonging to an event.

In particular, we train a binary classification model based upon the features whose extraction we explain in Section 3.2: one class represents photos that belong to an event and its training data is represented by X_C, and another class represents photos that do not belong to an event. For that latter class, we compile a small, random subset of photos of X that does not intersect with X_E (in other words, photos that do not fall within the same date and time or location *boundaries*) of a given event. We utilize a Support Vector Classifier (with the penalty parameter denoted as C) as our method of classification. Initial tests show superior performance of this method over other common classification methods.

3.3.2 Post-Expansion

In this step, we include photos that likely belong to a detected event but may have been *mistakenly* discarded by the feature-based pruning step in the prior Section 3.3.1. In particular, these might be photos that are linked to users who have multiple photos belonging to an event. 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 this event are *of* this event.

We first approximate the temporal beginning and end of each event. We do this by averaging the capture date and time d of n_e photos for both ends of the temporal window that is spanned by photos of the set union $X_U = X_C \cup X_P$. Next, we determine all users U_p with at least n_p represented photos within X_U. Based on our assumption in Section 3.1 (and shown in Figure 2), we finally include all photos of U_p that fall within the approximated event beginning and end times (we allow for some additional temporal leeway d_p), and, if given, whose exact or approximated location is within a threshold l_p.

3.3.3 Clustering

We propose an optional clustering step to further improve detection and retrieval performance for datasets that consist mostly of photos belonging to events rather than of photos not belonging to any events. Recall that our detection and retrieval approach starts with a set of spatio-temporal clusters S. To detect more events and retrieve more photos, we additionally incorporate the set of temporal clusters T. In other words, we now also consider those clusters that are *defined* by photos that do not encode any geographical information. Note, however, that we only consider such temporal clusters of S whose associated photos (we denote this set of associated photos as X_G) do not yet *belong* to any detected events; in other words, photos that are not yet retrieved. Likewise, we compile a set of photos that we have already retrieved (we denote this set as X_R).

Using K-Means [8], we then cluster the union set $X_G \cup X_R$ based upon the features that we extract in Section 3.2. Since the photos in X_R are associated with detected events, we use the event labels of these photos to assign event labels to the photos of X_G according to closest cluster. Next, we omit all assigned event labels with an overall frequency less than n_c, and we further propose to optionally omit all but the most frequent event label for each user. Our assumption is that a user is likely to attend only one social event over a certain period (that is implicitly defined by the temporal cluster). We denote the set of photos whose assigned event label we do not omit or discard as X'_G. We propose two

Figure 3: Exemplary photos from the MediaEval SED Dataset.

variants for determining which photos of X'_G to include as part of the final retrieval result. In the first variant, we consider only those photos of X'_G that correspond to the k most frequently assigned event labels. In the second variant, we do not use k and simply consider all photos of X'_G. Note, however, that in both variants we levy an additional constraint where the portion of photos that are associated with both X'_{Ge} and X_{Re}, where e signifies a given event, must be below a threshold β.

3.4 Event Type Classification

In this section, we extend our framework to classify the event type that one or multiple photos may belong to. We propose a supervised classification model that is similar to Section 3.3, and that is similarly based upon the features that we extract in Section 3.2. First, we train a multi-class model using a set of photos X_{tr}, where each photo is labeled with one of the event types we wish to predict. We can then utilize this trained model to predict the event type of any *new* and unlabeled photo x_{te}. We extend this basic approach in two ways.

One option is to expand the set of labeled photos X_{tr} in the training step. We perform the same spatio-temporal clustering as in Section 3.3. For every training photo $x_{tr} \in X_{tr}$ that falls within any of these spatio-temporal clusters, we consider all photos of any matching spatio-temporal cluster as additional training photos, and assign the same event type label of x_{tr} to these additional photos.

Another option is to not treat each testing photo x_{te} separately in the prediction step, but to consider whether a testing photo x_{te} belongs to an event cluster. We can then consider how a testing photo x_{te} relates to other photos belonging to the same event cluster. Since photos belonging to an event cluster can be predicted and thus labeled differently, we perform a simple arbitration by defining the *overall* type of an event cluster based upon its most frequently associated event type label. As such, we assign the same event type labels to all testing photos belonging to the same event clusters.

4. EXPERIMENTS

4.1 Dataset

We perform experiments on the MediaEval Social Event Detection (SED) Training Dataset [14] released in 2013 (exemplary photos in Figure 3). The dataset consists of two sets: set D_1 specifies 306150 Flickr photos for the task of event detection and retrieval, and set D_2 specifies 57165 Instagram photos for the task of classifying photos according to event types. Accompanying metadata (unique Flickr ID, capture date and time, username, title, description, keywords and, in about 46% and 27% of the cases, geographic coordinates) are provided in a separate XML file. Ground truth that associates photos (specified by their Flickr IDs) with event clusters (14882 classes) or types (sporting events, protest marches, BBQs, debates, expositions, festivals, concerts and *others*) is separately provided.

4.2 Experimental Setup

We evaluate the event detection and retrieval performance on D_1 in its entirety due to the unsupervised nature of the detection and retrieval task. To evaluate the event type classification performance, however, we perform a five-fold stratified shuffle split and average its results. For each fold, we use one third of D_2 as a testing set and the remaining photos of D_2 for training a classification model as explained in Section 3.4. We base all experiments on the following default model configuration and parameter values: $n_f = 9600$, $n_d = 3$, $n_e = 3$, $n_a = 3$, $n_p = 2$, $n_c = 3$, $d_p = 30min$, $l_p = 5km$, $\beta = 0.33$ and $C = 0.1$.

4.3 Results and Evaluation

We report and evaluate the performance of our framework using four common scoring measures: Precision (P), Recall (R), Normalized Mutual Information (NMI) and F1. All four scores are in the range $[0, 1]$, where higher values indicate better results.

4.3.1 Spatio-Temporal Propagation

As mentioned in Section 4.1, the 2013 MediaEval SED Training Dataset that we use for our experiments provides geographic coordinates for some but not all photos. For the event detection and retrieval set D_1, geographic coordinates are provided for 140472 photos out of a total of 306150 photos, or 45.88%. Using our approach from Section 3.1, we are able to propagate location information and thus increase the number of photos whose exact or approximate location we are aware of by the following percentages: 0.99% based on only the usernames; 32.86% when analyzing the photos' textual information and inferring their approximate locations (on a city-level); and 33.84% when combining both to a total of 79.73%. As a result, we are able to compile a large number of spatio-temporal clusters S that we exploit in the subsequent event detection and retrieval steps.

4.3.2 Event Detection and Retrieval

Corresponding to Section 3.3, the left plot of Figure 4 shows our baseline performance for event detection and photo retrieval without subsequent clustering. We achieve a P of 0.55, a R of 0.49 and a NMI of 0.69. When expanding the spatio-temporal clusters in terms of their spatial and temporal windows (and thus including additional candidate photos), and then pruning these candidate photos based upon their textual and visual features, we are able to increase R by as much as 0.06, and NMI by up to 0.08.

After event detection and initial photo retrieval (but prior to subsequent clustering), we propose to re-include *relevant* photos mistakenly pruned in the prior spatio-temporal expansion step. The left plot of Figure 4 illustrates that this step impacts performance only marginally with gains of around 0.02. The proposed post-expansion procedure seems

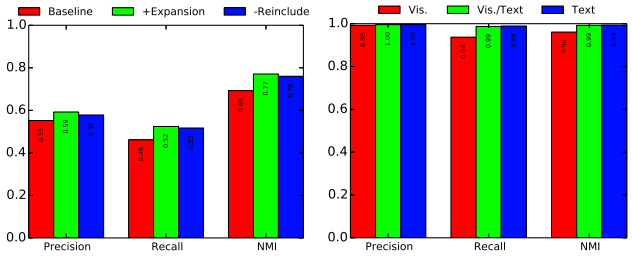

Figure 4: Event detection and photo retrieval. Left: *without* subsequent clustering, both spatio-temporal candidate expansion and re-inclusion of samples in a post-expansion step improve performance. Right: *with* subsequent clustering, textual features moderately outperform visual features.

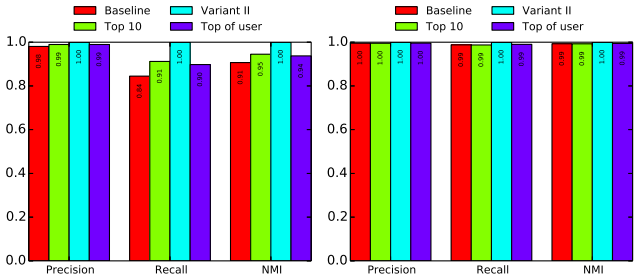

Figure 5: Event detection and photo retrieval *with* subsequent clustering: performing spatio-temporal candidate expansion helps achieve *near-optimum* performance regardless of variation (right); *Variant II* performs best, otherwise (left).

to be most effective when using smaller values for d_p such as $30min$, in which case P decreases less while still re-including some mistakenly pruned photos.

The two plots in Figure 5 show that subsequent clustering can significantly increase performance when, as in our case, the dataset covers mostly only events (as opposed to also including a significant amount of photos that do not relate to events). When performing spatio-temporal expansion and when using both textual and visual features throughout our framework, we achieve *near-optimum* results as the right plot shows. Our best results are thus in line with [10] and [15], which are among the top-performing approaches in the 2013 MediaEval SED Benchmark. However, our performance lessens when not performing spatio-temporal expansion (left plot). In that case, our first clustering variant (we perform experiments using the top five and top ten most frequent clusters per event) performs less effectively than our second clustering variant. Moreover, we notice that considering only photos according to the most frequent cluster label further improves performance (e.g. R up to 0.06 compared to our baseline configuration).

The right plot of Figure 4 illustrates that textual features moderately outperform visual features (R and NMI increase by 0.05) when performing clustering. We also experiment with the number of clusters when performing K-Means clustering. We notice better performance when the number of chosen clusters is in the range of 20 to 50 (we use a default

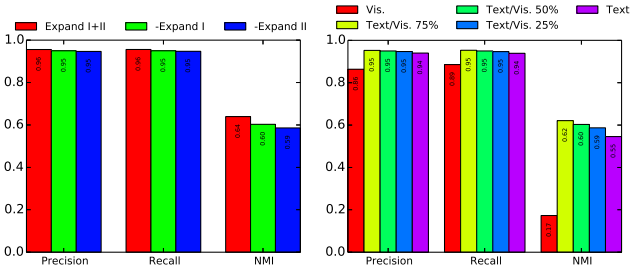

Figure 6: Event type classification: two optional variations slightly further improve performance (left). Textual features clearly outperform visual features, but a combination of both features leads to best performance (right).

of 30 clusters) than when using either a much lower or much higher number of clusters.

4.3.3 *Event Type Classification*

Figure 6 shows the event classification performance of our framework and, in particular, the classification approach detailed in Section 3.4. The figure's left plot shows that we achieve a P and R of 0.96, and a NMI of 0.64, in our baseline configuration. It also shows that our two additional variants can further increase the NMI score by as much as 0.04. Although we use the same feature extraction method and configuration as in the event detection and retrieval step detailed in Section 3.3, visual features show a larger positive impact in this event classification step. As the right plot of Figure 6 illustrates, we achieve a P and R of over 0.95, and a NMI of up to 0.62, when using both textual and visual features. This translates to a gain of 0.07 (NMI) compared to only using textual features. When utilizing only visual features, NMI performance drops significantly to 0.17 while P and R only drop to 0.86 and 0.89, respectively.

A closer evaluation also reveals that our approach classifies photos as *non-events* notably better than as specific events. Of nine possible event type classes (seven explicit types, *other* and *non-event*) on which we train our model, our approach best classifies the event types *concert* (F1-score of 0.52), *protest* (0.37) and *theater-dance* (0.31). *Fashion* and *other* perform the worst with a F1-score of under 0.10.

5. SUMMARY

We present a framework to detect or cluster social events in web photo collections, retrieve associated photos and classify these photos according to event types. We combine various event-related contextual cues such as date and time, location, and usernames with both textual and visual information using a constraint-based clustering and classification model. We report and evaluate results that validate the effectiveness of our approach. For future research, we intend to also incorporate information from social networks.

Acknowledgments

This work is partially supported by EU project CUbRIK under grant agreement FP7-287704. We would also like to acknowledge the MediaEval Multimedia Benchmark for providing the utilized datasets.

6. REFERENCES

[1] E. Benson, A. Haghighi, and R. Barzilay. Event Discovery in Social Media Feeds. In *NAACL HLT*, 2011.

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

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

[4] L. Chen and A. Roy. Event Detection from Flickr Data through Wavelet-based Spatial Analysis. In *CIKM*, 2009.

[5] W. Cohen, P. Ravikumar, and S. Fienberg. A Comparison of String Metrics for Matching Names and Records. In *KDD Work. Data Clean. Object Consol.*, 2003.

[6] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal Event Clustering for Digital Photo Collections. *TOMCCAP*, 1(3), 2005.

[7] A. Graham and H. Garcia-Molina. Time as Essence for Photo Browsing through Personal Digital Libraries. In *Digit. Libr.*, 2002.

[8] J. Hartigan and M. Wong. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.*, 1979.

[9] D. Joshi and J. Luo. Inferring Generic Activities and Events from Image Content and Bags of Geo-tags. In *CIVR*, 2008.

[10] T. Nguyen, M. Dao, and R. Mattivi. Event Clustering and Classification from Social Media: Watershed-based and Kernel Methods. In *MediaEval Work.*, 2013.

[11] A. Oliva and A. Torralba. Building the Gist of a Scene: The Role of Global Image Features in Recognition. *Brain Res.*, 155, 2006.

[12] S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, and A. Vakali. Cluster-Based Landmark and Event Detection for Tagged Photo Collections. *MultiMedia*, 18(1), 2011.

[13] T. Rattenbury, N. Good, and M. Naaman. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. In *SIGIR*, 2007.

[14] T. Reuter, S. Papadopoulos, V. Mezaris, P. Cimiano, C. de Vries, S. Geva, and C. D. Vries. Social Event Detection at MediaEval 2013: Challenges, Datasets and Evaluation. In *MediaEval Work.*, Oct. 2013.

[15] S. Samangooei, J. Hare, and D. Dupplaw. Social Event Detection via Sparse Multimodal Feature Selection and Incremental Density-based Clustering. In *MediaEval Work.*, 2013.

[16] R. Troncy, B. Malocha, and A. T. Fialho. Linking Events with Media. In *I-SEMANTICS*, 2010.

[17] K. Watanabe, M. Ochi, M. Okabe, and R. Onai. Jasmine: A Real-time Local Event Detection System based on Geolocation Information Propagated to Microblogs. In *CIKM*, 2011.

Photo Clustering of Social Events by Extending PhotoTOC to a Rich Context

Daniel Manchon-Vizuete
Pixable
New York, USA
dmanchon@gmail.com

Irene Gris-Sarabia
Universitat Politecnica de Catalunya
Terrassa, Catalonia
i.gris@hotmail.com

Xavier Giro-i-Nieto
Universitat Politecnica de Catalunya
Barcelona, Catalonia
xavier.giro@upc.edu

ABSTRACT

The popularisation of the storage of photos on the cloud has opened new opportunities and challenges for the organisation and extension of photo collections. This paper presents a light computational solution for the clustering of web photos based on social events. The proposal combines a first over-segmentation of the photo collections of each user based on temporal cues, as previously proposed in PhotoTOC. On a second stage, the resulting mini-clusters are merged based on contextual metadata such as geolocation, keywords and user IDs. Results indicate that, although temporal cues are very relevant for event clustering, robust solutions should also consider all these additional features.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Systems**]: Information Storage and Retrieval

General Terms

Design,Experimentation,Performance

Keywords

Clustering, Photo Collections, Event Detection

1. MOTIVATION

The International Telecommunications Union (ITU) announced that in 2014, the amount of active cellular phones would for the first time exceed the world population. Most of these devices are equipped with a photo camera, which is regularly used by the owners to capture, among others, relevant events of their lives. Many of these images are transmitted and stored on third-party services on the cloud, in many cases, through the same cellular network or wireless connections. There exist two main motivations for transferring these data to the cloud: firstly, sharing content with other users and, secondly, saving these memories on a storage facility which is considered safer, cheaper and more usable than the offline photo collections on users' personal computers.

Storing personal photos of relevant memories on the cloud offers new opportunities in terms of enhancing these digital records. Assuming that a user will only choose to capture and store photos from relevant events in his life, it is also probable that he will be interested in expanding the collection with photos coming from other users. Social events correspond to periods in the life of every user where there is exist a high probability that other users have captured complementary content that are willing to share. Additional photos may offer better image quality, new points of view, missing moments or completely novel information for the user. All these services could be offered by the cloud providers in addition to the basic storage, both for private events such as family and friends reunions, or for a public audience such as sports games or music concerts.

In addition to increasing and enhancing the visual content from a social event, photo collections on the cloud can also benefit from sharing contextual data related to the event. One of the main challenges that personal photo collections present is their retrieval, given that usually only a small portion of them has associated semantic metadata. Nevertheless, a photo with missing annotations may import annotations from other photos associated to the same event that had been generated by other users. The tedious process of manual annotation may become more appealing if it only requires a review of suggested tags from other photos associated to the same even [13], or even active and fun if a gamification scheme is adopted [10]. Also automatic annotation can benefit contextual data [22], for example by considering the expansion of missing metadata from other photos associated to the same social event. In any of these cases, it is necessary to identify these social event and the photos that depict it. This paper proposes a solution to this problem, by clustering a large collection of photos in a previously unknown amount of events.

The described services based on social event detection suggest a computational solution to be run on a centralised and shared service on the cloud, in contrast to other scenarios where the personal data of the user is processed on the client side. Any computation on the cloud typically implies an economical cost on the server which motivates extremely efficient solutions, even at the cost of some accuracy. For this reason, it is of high priority that any solution involves only light computations, discarding this way any pixel-related operation which would require the decoding and processing of

ICMR 2014 SEWM Workshop, Glasgow, Scotland
.

the images. In addition, the proposed algorithm is based in a sequential processing of data which on the temporal sorting, which easily allows the introduction of new photos in the collection. Computational costs are also limited to a sliding window, which provides a scalable solution capable of dealing with large amounts of data coming from large amounts of users.

The work presented in this paper was assessed in the benchmark prepared by the MediaEval 2013 Social Event Task [19]. Eleven solutions from different research groups participated in this campaign on a common dataset and metrics for social event detection. The work presented in this paper achieved the second best result in terms of precision and third best result in terms of F1-Measure in the task of photo clustering.

These paper is structured as follows. Section 2 reviews some of the previous works in the field of event clustering and, more specifically, in its application to social media on the web. Section 3 describes the photo clustering technique proposed in this paper, firstly with a description of the PhotoTOC algorithm and later with its adaptation to the contextual metadata available on web photos. Later, Section 4 reports on the experiments run to assess the proposed solutions based on a public dataset and backed by a scientific benchmark. Finally, Section 5 provides the insights learned and points at future research directions.

2. RELATED WORK

The detection of events in personal photo collections has received the attention of several previous works inside and outside the MediaEval benchmark.

A first wave of works was published in parallel with the popularisation of personal digital collections, basically addressing the problem of an offline creation of photo albums based on events. In these first works, the contextual information was very limited because users did not generate much textual information and most cameras did not include geolocation sensors. Loui and Savakis [11] proposed a system to define events and sub-events based firstly on date/time and, secondly, a content-based approach using color histograms. The system included a quality-screening software to discard those photos presenting underexposure, low contrast, camera defocus or movement.

The contribution from Cooper et al [4] also combined time stamps and visual content but, in this case, though, low frequency DCT textures were used to assess the visual similarity. In their work they highlighted that temporal clustering should not be limited to compare adjacent sets of pictures, but expanded to a controlled and local neighbourhood. The *PhotoTOC (Photo Table of Contents)* system by Platt et al [16] focused on collections from single users and generated an initial set of event boundaries based on time stamps. Whenever the algorithm generated a cluster with more than 23 elements, the cluster was considered too large and was split according to color features. This splitting was addressed to the final application of PhotoTOC, which was actually generating a visual table of contents for a photo collection. Using visual features for this over-segmentation aimed at providing color diversity in the generated thumbnails. Our work has adopted this time-based clustering solution due to its simplicity and effectivity, but has expanded it to a multi-user framework with rich metadata available. For this reason, this approach is described in detail in Section 3.2.

The introduction and popularisation of GPS sensors in photo cameras enriched the problem of event detection with a new feature: geolocation [12] [2]. Cao et al [3] added these metadata to the time stamps and used it to annotate photo collections. The process benefited from a hierachical clustering of the photos based first on events and secondly in scenes, where scenes were to be understood as semantic labels. This work already remarked the challenges that poses working with photo collections where, in general, only a part of the photos will have geolocation data available.

Recent works have focused on the particularities of photos shared on the web, mainly through social networks. A first effort focus on social media was published by Becket et al [1], where they proposed a method for learning multi-feature similarity metrics based on the rich context metadata associated to this type of content. In their work they argued that clustering techniques based on learned thresholds are more appropriate that those solutions which require a prior knowledge on the amount of clusters (eg. K-Means or EM), or other based on graph partitioning. In particular, they suggested a single-pass incremental clustering that would compare each non-classified photo with a set of existing clusters. If the similarity to one of these clusters satisfied a certain threshold, the photo will be assigned to the cluster; if not, a new cluster was created. The similarity is defined as the average of similarities between the non-classified photo with a centroid computed in each existing cluster. This way, the features of a non-classified photo do not need to be computed with each classified photo, but only with the centroid of the clusters that contain them. We have also adopted a threshold-based approach based on cluster centroids, but applied in two passes: a first one that considers each user isolated, and a second one that exploits the rich context metadata.

Petkos et al [15] proposed a solution based in spectral clustering that would introduce a known clustering from the same domain (supervisory signal) that would determine the importance of each feature. The introduction of this example clustering guides the output in a semantic way, for instance, providing more relevance to geolocation features if the landmark determines the event nature, or to textual tags if the event has a strong semantics not related to a specific location (eg. Christmas).

Reuter and Cimiano [17] proposed a system where, given a new photo, a reduced set of candidate events were retrieved. Each pair of new photo and retrieved event was represented by a feature vector of multimodal similarities. This feature vector was assessed with a classifier trained to identify correct pairs or whether the new photo should be associated to a new event.

The problem of photo clustering from social media specifically addressed in this paper has been extensively studied in the framework of the MediaEval benchmark for Social Event Detection [19]. This scientific forum allowed the comparison of different techniques in a common dataset and evaluation metrics. During the 2013 edition, Samangooeu et al [20] obtained the best performance in terms of F1-Score by applying a DBSCAN clustering [6] on an affinity matrix built after a fusion of the different features associated to the image. Their experiments indicated that textual information such as title, description and tags should not be fused; and that visual features did not provide any gain despite of the required computational effort. Another relevant contribu-

tion from Dao et al [5] defined a 2D a user-time image which was over-segmented by applying the watershed algorithm. As a second step, the resulting clusters were considered for merging considering different types of contextual metadata.

Compared to the presented approaches, our work gives special relevance to the temporal features, leaving the rest of modalities in a second term. We have prioritised a one-pass exploration of the data that would focus on a local temporal neighbourhood. This way, our solution is light weighted in terms of computational effort, having in mind its application on existing services of photo storage on the cloud.

3. EVENT CLUSTERING

In this paper, we present an extension of the PhotoTOC system [16] in the context of social events represented by rich contextual metadata. The architecture of the proposed solution is depicted in Figure 1. In this example, the photo collections of two users are represented on a temporal axis based on the time stamps associated to each image. During a first stage, each photo collection is split in mini-clusters based on their timestamps, according to a previous work [16]. The resulting sets of photos are sequentially compared to assess their possible merges based on rich contextual metadata, such as keywords, user information and geolocation data. The final result is a clustering of photos from different users to represent social events.

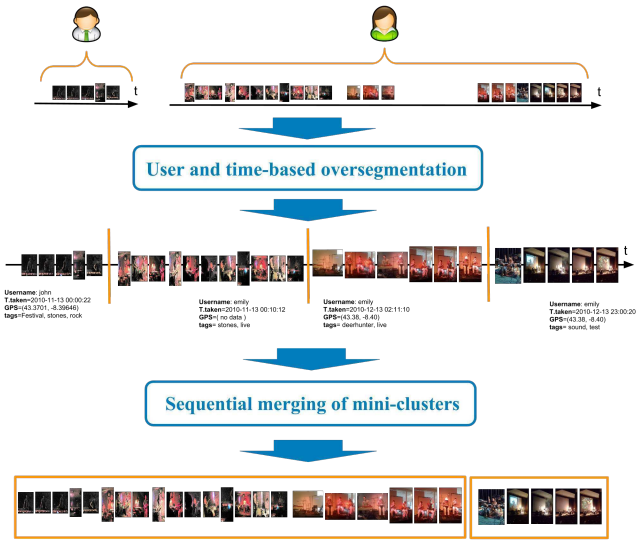

Figure 1: System architecture.

3.1 Context-based metadata

The presented system considers four types of contextual metadata which are commonly associated to photos on the web:

Time stamp: If available, this metadata field reflects when the photo was taken.

Geolocation coordinates: Latitude and longitude of the camera when the photo was taken.

Tags: One or more keywords associated to the image that were added by the user. These type of textual metadata typically present less non-relevant terms for classification, such as articles, conjunctions, connectors, prepositionsâĂę

User ID: A unique identifier of the individual who uploaded the video to the cloud.

In our work, time features are chosen as pivotal in the system as they provide a sorting criteria that allows a sequential processing of the dataset. This decision facilitates the addition of new photos in the collection, which can be easily inserted in the timeline and compared with the existing events. Using time as a pivotal feature is also supported by other authors [7] [11] [16] [14].

3.2 User and time-based over-segmentation

The first step in the proposed solution considers the photos of each user separately and clusters them in small sets that aim at providing a high recall of the actual event boundaries.

This stage corresponds to the *PhotoTOC* solution [16] already introduced in Section 2. According to that algorithm, photos from each user are initially sorted according to their creation time stamp and are sequentially clustered by estimating the location of event boundaries. A new event boundary is created whenever the time gap (g_i) between two consecutive photos is much larger than the average time differences of a temporal window around it. The extension of the temporal window is determined by parameter d, which corresponds to the amount of previous and posterior time gaps which are considered in the averaging.

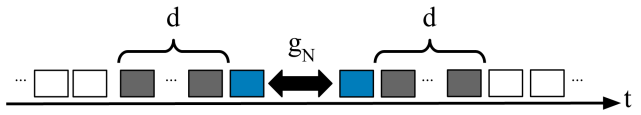

Figure 2: A new event boundary is created when time difference g_N exceeds the average time differences in the neighbourhood defined by d.

In particular, a new event is created whenever the criterion shown in Equation 1 is satisfied. This way, a new event boundary is created when a time gap is significantly larger than the averaged time gaps in its neighbourhood.

$$log(g_N) \geq K + \frac{1}{2d+1} \sum_{i=-d}^{d} log(g_{N+i}) \qquad (1)$$

As a result, an over-segmentation of mini-clusters is obtained. Each mini-cluster is characterised by combining the metadata of the photos they contain. This combinations are used in the posterior stages to assess the similarity between pairs of these mini-clusters.

3.3 Sequential merging of mini-clusters

The collection of time-sorted clusters is sequentially analysed in increasing time value, as depicted in Figure 3. Each cluster is compared with the posterior M clusters, a time window set to avoid excessive computational time. Two clusters are merged whenever a distance measure is below

a learned threshold. Thresholds are learned during a previous training stage by selecting those values which optimise a measure of quality for the whole system. This stage does not process the mini-clusters of each user separately, as in Section 3.2.

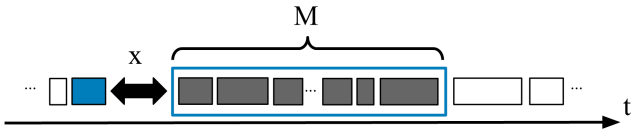

Figure 3: Each mini-cluster is compared to the following M mini-clusters, and merged if their relative distance x is below a certain threshold .

The distance x between two mini-clusters is assessed with a weighted and linear combination of normalised distances from the different features available, as presented in Equation 2. Each similarity $\bar{s_i}$ corresponds to a different contextual metadata, such as geolocation, keywords or user identifications.

$$x = \sum_i w_i \bar{s_i} \tag{2}$$

3.3.1 Distances metrics

Each mini-cluster is characterised in terms of time stamps, geolocation, user ID and textual tags. The different types of contextual metadata for mini-clusters are computed and compared as follows:

Time: L1 distance on the averaged time stamps of every photo in each mini-cluster, as in [15].

Geolocation coordinates: Harversine distance on the averaged latitudes and longitudes of every photo in each mini-cluster. This distance provides the great-circle distances between two points on a sphere.

Tags: All the tags are aggregated to represent each mini-cluster. The similarity between two mini-clusters is assessed with the Jaccard Coefficient, which compares the sum of shared terms between two mini-clusters to the sum of terms that are present in either of the two mini-clusters but which are not shared [9]. In case that no tags are available for any of the two mini-clusters to be compared, this modality is ignored when assessing the distance.

User ID: Mini-clusters are created, by definition, associated to a unique user ID. In this case the distance is binary-valued, 1 when the user ID from the two mini-clusters is the same, 0 otherwise.

3.3.2 Normalisation of Distances

The linear fusion proposed in Equation 2 requires a normalization of the distance values d_i associated to different type of contextual metadata. These different types may correspond to geographical information, keyword or an identification of the user who uploaded the photo to the cloud. Without such normalisation, the different value ranges of the distances associated to each type of feature would make their comparison biased towards the larger distances.

Distance values are mapped into similarity values through the phi function $\Phi(x)$, which corresponds to the cumulative distribution function (CDF) for a normal distribution. This transformation will map an average distance value of a normal distribution to 0.5, and generate a range of similarity values in the interval $[0, 1]$. Large distances will be transformed into similarity values close to zero, while small distances will correspond to similarities near 1.

$$\bar{s_i} = \Phi\left(d_i, \mu_i, \sigma_i\right) = \frac{1}{2}\left[1 + erf\left(\frac{d_i - \mu}{\sqrt{2\sigma^2}}\right)\right] \tag{3}$$

This normalisation strategy requires the estimation of the average μ and standard deviation σ of the distances for each type of contextual metadata. This estimation is performed with a training process by comparing pairs of photos which correspond to the same event according to the ground truth. We focus on pairs of photos from the same event as we are interested only on the range of distances that correspond to possible merges of the mini-clusters. This way, a 0.5 similarity values is associated to the average distance for the pairs of photos within the same event.

3.3.3 Estimation of Feature Weights

After normalization, it is still necessary to estimate the weight of each feature type w_i to be later applied to the linear fusion. The adopted strategy estimates the weight for each feature according to their relative performance when considered separately for merging. That is, during the training stage, the merging of mini-clusters is tested using a single type of contextual metadata. The experiment is repeated for different merging thresholds, allowing the estimation of the best performance if only one modality is to be considered. The best performance value achieved in each case is used as a weight for the corresponding feature.

In our work, the F1-Score is used as the basic metric to assess the clustering of photos in events. As a consequence, the weights associated to each type of contextual metadata correspond to the normalised best F1-Score achieved by using each feature separately. Equation 4 depicts estimation of w_i based on the F1-Score. The definition of F1-Score can be found in Section 4.2.

$$w_i = \frac{maxF1_i}{\sum_j maxF1_j} \tag{4}$$

4. EXPERIMENTS

4.1 Dataset description

The work presented in this paper is the result of our participation in the MediaEval 2013 Semantic Event Detection (SED) task [19]. The dataset used in that benchmarking is publicly available as the ReSEED Dataset [18].

The full dataset consists of 437,370 pictures from 21,169 events, together with their associated metadata.All the photos were uploaded to Flickr between January 2006 and December 2012. Users published these pictures with different variations of a Creative Commons license, which allows their free distribution, remixing and tweaking. Ground truth events were defined thanks to the *machine tags* that Flickr uses to link photos with events, as presented in [17]. The dataset is already split in two parts: development (train) and evaluation (test). The development dataset includes

306,159 pictures (70%), while the evaluation part consists of 131,211 photos (30%). Training data was used to estimate the parameters for feature normalisation and fusions, as well as the distance thresholds to fuse the mini-clusters. Together with the dataset, an evaluation script is provided to avoid any implementation problem when comparing evaluation metrics from different authors.

In addition, the dataset presents an inherent challenge due the incompleteness and corruption of the photo metadata. Metadata is not complete, as only 45.9% contain geolocation coordinates, 95.6% tag associated, 97.9% a title and 37.9% a textual description. Another source of problems are the identical time stamps between the moment when the photo was taken and when it was also uploaded. These situations are common specially when dealing with online services managing photos, which present heterogeneous upload sources and, in many cases, remove the EXIF metadata of the photos. These drawbacks have been partially managed in the proposed solution, which combines the diversity of metadata sources (time stamps, geolocation and textual labels) in this challenging context.

The reader is referred to [19] for further details about the study case and dataset.

4.2 Metrics

The quality of the system is assessed by comparing the clusters automatically generated by our algorithm with the ground truth events . We have computed the classic *Precision*, *Recall* and *F1-Score* metrics given its popularity [1] [17] as well as adoption in MediaEval 2013 SED task [19].

Given a photo x in the dataset, it is associated to an event e_x by the ground truth annotation, and to a cluster c_x by the automatic classification process. The classification of x can be assessed with the *Precision* (P_x) measure by computing the proportion of documents in the c_x which also belong to the e_x, as presented in Equation 5.

$$P_x = \frac{|c_x \cap e_x|}{|c_x|} \tag{5}$$

Analogously, a complementary *Recall* (R_x) measure is obtained as the proportion of photos from e_x which are classified in the c_x, as shown in Equation 6.

$$R_x = \frac{|c_x \cap e_x|}{|e_x|} \tag{6}$$

The individual P_x and R_x obtained for each document can be averaged through the whole dataset to obtain a global $Precision(P)$ and $Recall(R)$ values, respectively. Finally, these two averages can be combined in the single *F1-Score* (F_1) presented in Equation 7. This value represents the two common properties desired in a clustering algorithm: maximum homogenity within each cluster, while minimising the number of clusters in which photos from each event are spread.

$$F_1 = 2\frac{PR}{P+R} \tag{7}$$

4.3 Estimation of merging thresholds and fusion weights

The contribution of each feature type to the fused similarity function described by Equation 2 is estimated by assessing the F1-Score when merging mini-clusters with a single feature. For this estimation the parameters responsible of the temporal segmentation in mini-clusters were set to $K = log(150)$ and $d = 40$. This way, the result will deliberately several mini-clusters and the potential of each feature may be assessed more clearly.

Figures 4 and 5 show the evolution of F1-Score with respect to the merging threshold for the geolocation and tag features, respectively. In the case of user IDs, instead of learning a distance threshold, the merging criterion simply states that two mini-clusters will be merged if they present the same user ID.

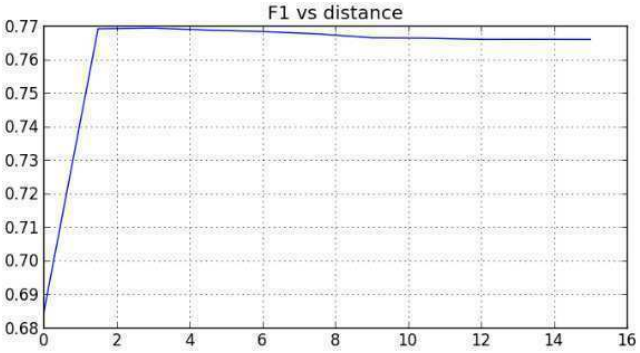

Figure 4: Evolution of the F1-Score with respect to a merging threshold based on geolocations.

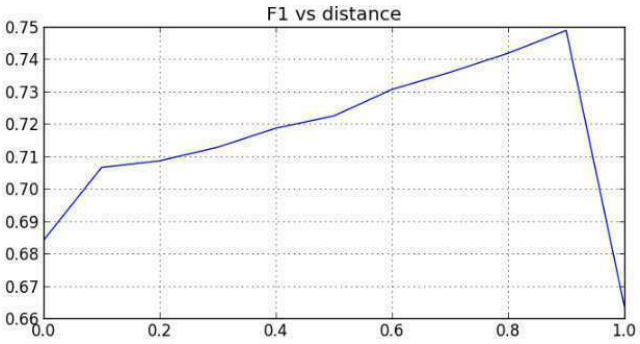

Figure 5: Evolution of the F1-Score with respect to a merging threshold based on tags.

Table 1 contains the normalised weights according to Equation 4, computed by considering the best F1-Scores achieved with each feature type. Weights are also computed for those cases where no geolocation metadata is available, a situation which appears often in 45.9% of the photos. These values indicate that the most important reason for the fusion of two clusters is that both of them belong to the same user ID, while geolocation and tags present a lower and similar relevance.

4.4 Estimation of normalisation parameters

The weights w_i used in the linear fusion of Equation 2 require the estimation of the mean μ_i and standard deviation σ_i for each type of contextual metadata. Such estimation was based after the computation of the distances between

	Geolocated	No geolocated
Geolocation	0.28	-
User ID	0.44	0.60
Tags	0.22	0.30

Table 1: Feature weights for photos with and without geolocation metadata.

1,000 random pairs of photos selected from the training set and belonging to the same event. Table 2 includes the results of this estimation.

	Distance	Mean (μ)	Std (σ)
Geolocation	Haversine	0.164 Km	2.175 Km
Tags	Jaccard	0.526	0.425

Table 2: Mean and standard deviation of distances between 1,000 pairs of photos belonging to a same event.

4.5 Event clustering

The performance of the first over-segmentation, as described in Section 3.2, and its later merge, explained in Section 3.3, has been assessed on the test data partition of ReSEED dataset. The experiments have considered a value of $M = 15$ in the merge stage, which keeps a light computational approach while providing some robustness with respect to the temporal sorting of the mini-clusters.

4.5.1 Qualitative results

Figures 6 and 7 provide two examples of correct events that were detected with the presented techniques. On the other hand, Figures 8 and 9 show cases in which the algorithm failed into a correct event detection.

The example in Figure 6 depicts a music festival where a distinctive quality can be appreciated between the first photo of the series and the rest. This case presents a situation where photos taken from different cameras have been successfully clustered. In the case of Figure 7, the social event of a seminar takes place in two different locations: a classroom and a restaurant. Although the location has changed, the proximity in time keeps the event connected. The two cases depict different challenges in terms of event continuity that have been successfully detected and merged by the algorithm.

The third exampled depicted in Figure 8 presents an example where an event has been incorrectly split in three. This is because this event, which depicts a conference, spans through three different days. This time gap between the three blocks, the lack of geolocation data and the usage of different tags every day prevents the identification of a single event.

An opposite case of undesired merge is depicted in Figure 9. In this case, geolocation data is very similar and time stamps refer to the morning and afternoon of the same day. The ground truth considers two sets as depicting different events, while the algorithm merged them given their closeness. It is difficult for a non-expert on the topic to discern whether these photos are part of the same event.

4.5.2 Quantitative results

Figure 6: Detected event that combines photos of different qualities.

Figure 7: Detected event depicting multiple participants and distinguishable semantic moments.

Table 3 offers quantitative results for event clustering on the ReSEED dataset. Results are provided considering two different pairs of (K, d) parameters. The first column considers the values proposed by the original PhotoTOC system [16], while the second column contains the results with another pair of values empirically set in the current work.

The first observation from the first row in Table 3 is the sensitivity of the algorithm to the pair of (K, d) parameters for temporal clustering. The results obtained with the original configuration are clearly improved by manually tuning them for the ReSEED dataset. If we assume that the authors of the PhotoTOC system tuned their parameters for optimal results for their dataset, we can conclude that the performance of the system is clearly influenced by the choice of these parameters.

If Table 3 is analysed by columns, it shows that, in general, using additional contextual metadata improves performance. All F1 scores are improved when the initial over-segmentation in mini-clusters is merged, but the exception of using the user ID in the second column. This decrease indicates that merging two mini-clusters in a neighbourhood of $M = 15$ based only on on user IDs may decrease performance if these first mini-clusters are already very good. This behaviour should be further studied with a more extensive study on the empirically value set for M.

The last row in Table 3 offers different interpretations upon the convenience of fusing different features. In both

	PhotoTOC [16] K=log(17), d=10	Our work K=log(600), d=14
Time	0.749	0.880
Time+Geolocation	0.802	**0.893**
Time+User ID	0.837	0.875
Time+Tags	0.814	0.883
Time+Fusion	0.822	0.883

Table 3: F1 scores for the different configurations presented in the paper.

Figure 8: An event is incorrectly split in three.

Figure 9: Two photo clusters (upper and lower rows) are incorrectly merged as a single event.

columns the performance of the fused features is not as good as one of the configurations using only one additional contextual data. Nevertheless, while in the first column it is outperformed by adding user information to the time-based clustering, in the second column it is geolocation data which is providing better results. Given the two different outcomes, one may consider the fusion approach as a way to provide some stability to the final solution because, in many real one problems, one may not have a ground truth available for tuning the (K, d) pair not deciding which type of contextual metadata is going to perform best used on its own. For this reason, feature fusion seems to be advisable in this context, although the method considered in this work may be improved by exploring other possibilities.

Among all the considered configurations, the best result is the merging of mini-clusters using only geolocation information. This result indicates the importance of this contextual metadata when combined with time and user information. The success of this configuration is surprising, given that only 27.9% of the pictures contain geographic information [19]. This circumstance raises the interest of predicting the geolocation of those photos that do not contain these type of metadata.

4.6 MediaEval Social Event Detection

The presented work was developed in the framework of the Social Event Detection Task 1 from the MediaEval 2013 benchmark [19]. This forum allowed comparing the results obtained with other state of the art solutions in the field. Table 4 includes the results published by the task organisers for the five teams that obtained better F1-scores among the eleven participants. Results indicate that our light-weight approach offers a state of the art performance, especially in terms of Precision. Notice that the F1-Score value presented in Table 3 slightly improves the results submitted in MediaEval 2013, due to a later optimisation of the (K, d) parameters for temporal clustering.

	F1-Score	Precision
Samangooei et al [20]	0.9454	0.96
Nguyen et al [14]	0.9234	0.98
Our work	**0.8833**	**0.96**
Witsuba et al [23]	0.8720	0.91
Sutanto et al [21]	0.8112	0.86

Table 4: Results of MediaEval 2013 Social Event Detection (Task 1).

5. CONCLUSIONS

This paper has explored the extension of an existing PhotoTOC algorithm for time-based event clustering to the domain of event detection of social events on the web. The initial sets of clusters based on time stamps are assessed in their local neighbourhood for merging. In a second stage, additional contextual metadata common in social media (geolocation, keywords and user ID) are exploited to complement the temporal ones. In both cases, a sequential processing of the data is applied, providing a light solution to the problem and avoiding the extraction of visual features proposed in the original paper of PhotoTOC [16]. This way, the algorithm fits better the low computational requirement of cloud-based services.

The presented experimentation has shown a competitive results when considering the photos from Flickr contained in the ReSeed dataset. Results have proven the sensitive to the parameters that define the temporal clustering to the dataset. While good results may be achieved with timestamps only, including other sources of metadata provides

stability to the system, making it more resilient to changes in the data particularities. When comparing different types of contextual metadata, the study does not provide a clear winner and suggests that a fusion approach between all of them is the safer bet.

One more of the main challenges posed by the social media on the web is the partiality of the available metadata. Future work should focus on an adaptive algorithm that may adjust to the available contextual data and, if necessary, search the missing one whether on the visual content or on the cloud itself. Another research line to improve is a better exploitation of the textual metadata. The Jaccard index is a too simple approach for comparing tags, and ontology-based solutions or text processing techniques should help in a better use of these metadata.

To sum up, the presented technique has allowed a fast resolution of the photo clustering of images based only contextual metadata. This allows a light-weighted solution designed to photo organisation with no visual processing involved, which facilitates its integration on systems with low computation requirements, such as services on the cloud.

Further implementation details can be found in our Python source code[1].

6. ACKNOWLEDGMENTS

This work has been partially funded by the Spanish project TEC2010-18094 MuViPro.

7. REFERENCES

[1] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Proc. of the third ACM international conference on Web search and data mining*. 2010.

[2] D. Martin-Borregon, L. M. Aiello, and R. Baeza-Yates. Space and time clusterization of Social media groups. MSc thesis, Universitat Pompeu Fabra, Barcelona. 2013.

[3] L. Cao, J. Luo, H. Kautz, and T. S. Huang. Annotating collections of photos using hierarchical event and scene models. In *Computer Vision and Pattern Recognition, 2008. IEEE Conference on*, pages 1–8. 2008.

[4] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal event clustering for digital photo collections. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 1(3):269–288, 2005.

[5] M.-S. Dao, G. Boato, F. G. De Natale, and T.-V. Nguyen. Jointly exploiting visual and non-visual information for event-related social media retrieval. In *Proc. of the 3rd ACM conference on International conference on multimedia retrieval*, pages 159–166. ACM, 2013.

[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[7] A. Graham et al. Time as essence for photo browsing through personal digital libraries. In *Proc. of the 2nd ACM/IEEE-CS conference on Digital libraries*, pages 326–335. ACM, 2002.

[1] https://github.com/dmanchon/mediaeval2013

[8] C. Hauff, B. Thomee, and M. Trevisiol. Working notes for the placing task 2013. In *MediaEval 2013 Workshop*, Barcelona, Catalonia.

[9] A. Huang. Similarity measures for text document clustering. In *Proc. of the Sixth New Zealand Computer Science Research Student Conference, New Zealand*, pages 49–56, 2008.

[10] E. Law and L. Von Ahn. Input-agreement: a new mechanism for collecting data using human computation games. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1197–1206. ACM, 2009.

[11] A. C. Loui and A. Savakis. Automated event clustering and quality screening of consumer pictures for digital albuming. *Multimedia, IEEE Transactions on*, 5(3):390–402, 2003.

[12] M. Naaman, S. Harada, Q. Wang, H. Garcia-Molina, and A. Paepcke. Context data in geo-referenced digital photo collections. In *Proc. of the 12th annual ACM Multimedia*, pages 196–203. ACM, 2004.

[13] M. Naaman and R. Nair. Zonetag's collaborative tag suggestions: What is this person doing in my phone? *MultiMedia, IEEE*, 15(3):34–40, 2008.

[14] T. Nguyen, M.-S. Dao, R. Mattivi, and E. Sansone. Event clustering and classification from social media: Watershed-based and kernel methods. In *MediaEval 2013 Workshop*, Barcelona, Catalonia.

[15] G. Petkos, S. Papadopoulos, and Y. Kompatsiaris. Social event detection using multimodal clustering and integrating supervisory signals. In *Proc. of the 2nd ACM International Conference on Multimedia Retrieval*, page 23. ACM, 2012.

[16] J. C. Platt, M. Czerwinski, and B. Field. Phototoc: automatic clustering for browsing personal photographs. In Proc. of Fourth Pacific Rim Conference on Multimedia, volume 1, pages 6–10 Vol.1, 2003.

[17] T. Reuter and P. Cimiano. Event-based classification of social media streams. In *Proc. of the 2nd ACM International Conference on Multimedia Retrieval*. ACM, 2012.

[18] T. Reuter, S. Papadopoulos and V. Mezaris. ReSEED: Social Event dEtection Dataset. In *Proc. of the ACM MultiMedia Systems Conference*. ACM, 2014.

[19] T. Reuter et al. Social Event Detection at MediaEval 2013: Challenges, datasets, and evaluation. In *MediaEval 2013 Workshop*, Barcelona, Catalonia.

[20] S. Samangooei et al. Social event detection via sparse multi-modal feature selection and incremental density based clustering. In *MediaEval 2013 Workshop*, Barcelona, Catalonia.

[21] T. Sutanto and R. Nayak. Admrg @ mediaeval 2013 social event detection. In *MediaEval 2013 Workshop*, Barcelona, Catalonia.

[22] T. Uricchio, L. Ballan, M. Bertini, and A. Del Bimbo. An evaluation of nearest-neighbor methods for tag refinement. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6.

[23] M. Wistuba and L. Schmidt-Thieme. Supervised clustering of social media streams. In *MediaEval 2013 Workshop*, Barcelona, Catalonia.

Event Detection from Social Media: User-centric Parallel Split-n-merge and Composite Kernel

Truc-Vien T. Nguyen
University of Lugano
6900 Lugano, Switzerland
thi.truc.vien.nguyen@usi.ch

Minh-Son Dao
University of Information Technology
Viet-Nam National University HCMC
sondm@uit.edu.vn

Riccardo Mattivi, Francesco G.B De Natale
mmLab - University of Trento, Italy
38123 Povo (TN), Italy
{rmattivi, denatale}@disi.unitn.it

ABSTRACT

In this paper, we present the framework that includes two methods for tackling event classification and clustering challenges defined by MediaEval 2013. For the former, we use supervised machine learning and experiment with Support Vector Machines. First, we present a composite kernel to jointly learn between text and visual features, second, we propose new features for the task, which are derived from Natural Language Processing community and encyclopedic knowledge-Wikipedia. For the latter, a user-centric parallel split-n-merge framework applied for unsupervised clustering social media events is introduced. The purpose of this framework is to cluster social media to events they depict by exploiting and exploring the role of users and the way users interact with data. The output of the proposed framework can be used for event organization/summarization, and as pre-processing stage for event detection and tracking. The methods prove robustness with F1 up to 98% in clustering challenge; the composite kernel yields competitive performance across different event types in classification challenge, and the new features yield significant improvement with respect to state-of-the-art

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Theory, Experimentation

Keywords

social event detection, clustering, classification, user-centric, split-and-merge, user-time image, kernel methods, support vector machines, disambiguation to wikipedia

ICMR 2014 SEWM Workshop, Glasgow, Scotland
.

1. INTRODUCTION

The proliferation of social media has led to an ever-increasing amount of web and multimedia content available on the Web. A large part of this content is related to social events, which are defined as events that are organized and attended by people and are illustrated by social media content created by people [23]. Thus, it is necessary to develop algorithms to support users in the detection and grouping of events into categories. This is commonly referred to as Social Event Detection (SED). The extensive testing and comparison of SED methodologies has been promoted in recent years by the MediaEval Challenge, which is a benchmarking initiative dedicated to evaluating new algorithms for multimedia access and retrieval. MediaEval 2013 recently called for the solution of one of these utmost requirements in SED: "discover event-related multimedia and organize them into event-specific clusters, within a collection of Web multimedia" . This problem is important not only for users who want to organize their data but also for providers who want to analyze data to offer the better tools for their customers under social events perspective. The SED task included two major challenges: unsupervised clustering, and supervised classification of event media. This paper describes social event detection methods that was specially built to meet these challenges at MediaEval 2013 [23]. We also report and discuss the advantages and performance of the methods based on the result evaluated by MediaEval 2013.

In [11], the authors define Social Media Network as an application that can join users worldwide by enabling the ability to create and exchange media via the Internet. Recently, an interesting proposal of Social Life Network (SLN)[9] has been proposed, where all people are kept up to date about real-life events thanks to a very large scale social network, with a major emphasis on multimedia data. In this paper, the authors point out several challenges and opportunities of large-scale social media, such as for instance the management of social events, defined as the "events that are organized and attended by people and are illustrated by social media content created by people"[24]. In [1],[21],[5], several problems related to social media analysis, such as event detection and classification, tracking, summarization, and association were introduced and discussed. Here, a further

confirmation can be found of the fact that automatically analyzing digital content related to social events is the utmost challenge, due to large-scale volume of data coming from different sources and sites, and low reliability of tags and annotations left by users of different communities.

In the context of MediaEval, several research groups brought their best tools to solve some of these problems. Five research groups from different countries participated to the classification challenge, whose goal was to classify a large set of images into event and non-event classes, and then into a set of pre-defined event types. In [28], the authors use a combination of scalable learning framework with a linear classifier based on Support Vector Machines (SVMs). They got 33.44% F1 with a combination of visual and text features. In [30], first LDA Gibbs sampling, then traditional classifiers such as k-Nearest Neighbor (kNN) and decision tree were then used. They achieved 13.1% F1. The system proposed in [10] got 7.48% F1. It computes the similarity between synset representing the tags(c1) and each of the categories(c2). They use Lin similarity measure to compute the semantic relatedness of word senses, then, if any photo encounters same Lin Similarity measure for more than one category, other constraints (Date, Time) were considered. The best results are reported with [2], which are 50% F1. They use textual features from each photo's title, description and keywords, together with GIST features (a feature vector with 960 elements) for each photo. Finally, a Linear Support Vector Classifier is used to classify events into categories. Note that in all of these approaches, there is no work which makes use of kernel methods. Moreover, they take the fusion of text and visual features in a unique learning function. In this paper, first, we propose to use kernels to exploit a wide range of functions for each kind of features, which has not been tried in previous works. Second, with kernel methods, it is very convenient to combine these two kinds of features: text and visual features. In this way, we could find the best method for each feature set and combine them by using various functions.

As far as the clustering is concerned,in [30] the authors use K-means clustering (where the value of k parameter is deduced from training data) and document ranking are used as a semi-supervised method to cluster event-related data. They make use of text information only. In [22], a data-driven three steps approach is applied with text and visual information. This method calculates inter-correlations among clusters to verify the final result. In [28], both text and visual information are used with variety of classifiers (SVM, Decision Trees) to cluster data. In [33], Factorization Machines is used to learn similarity between two time-ordering documents. This method requires a lot of tuning parameters. In [2], propagating geographic locations are applied to compensate the lack of exact location information. Text and visual features are concatenated with weight ratio to feed a linear support vector classifier. In [27], Lucern filter and affinity matrix are constructed with text and visual information. Nevertheless, they recognized at last that visual information makes their result worse. In general, these methods need to use the whole data set for analyzing. Besides, most of them are supervised methods that require a ground-truth for training. Both of these conditions are very difficult to be met in reality. In order to cope with the curse of ground-truth and volume of data, the unsupervised parallel clustering method is introduced, that exploits and explores the most interesting characteristic of social media: the users' role. The contributions of this method are: (1) low computational solution w.r.t large-scale data, (2) parallel computation, and (3) unsupervised clustering with no training data and third-party information requirements.

The structure of the paper is as follow: Section 2 describes our approach for event classification, which makes use of kernel learning; Section 3 introduces a user-centric parallel split-n-merge framework applied for unsupervised clustering social media events; Section 4 reports all experiments and results with our models; finally, Section 5 summarizes the conclusions.

2. SUPERVISED EVENT CLASSIFICATION

In this section we present the machine learning approach to classify events. We also describe the textual features derived from Natural Language Processing (NLP) literature as well as visual features. We can engineer kernels, using one kernel for each feature set and combining them. Thus, we focus on the problem of defining which are the most important features for the task.

2.1 Support Vector Machines and Kernel Methods

In this section we give a brief introduction to support vector machines, kernel methods and kernel spaces, which can be applied to the event classification task.

Support Vector Machines (SVMs) refer to a supervised machine learning technique based on the latest results of the statistical learning theory [32]. Given a vector space and a set of training points, i.e., positive and negative examples, SVMs find a separating hyperplane $H(\vec{x}) = \vec{\omega} \times \vec{x} + b = 0$ where $\omega \in R^n$ and $b \in R$ are learned by applying the Structural Risk Minimization principle [31]. SVMs is a binary classifier, but it can be easily extended to the multi-class case, e.g., by means of the *one-vs-all* method [25]. One strong point of SVMs is the possibility to apply kernel methods [26] to implicitly map data in a new space where the examples are *more easily* separable as described in the next section. Kernel methods [29] are an attractive alternative to feature-based methods since the applied learning algorithm only needs to compute the product between a pair of objects (by means of kernel functions), thus avoiding the explicit feature representation. A kernel function is a scalar product in a possibly unknown feature space. More precisely, The object o is mapped in $\vec{x}$ with a feature function $\phi : \mathcal{O} \rightarrow \Re^n$, where $\mathcal{O}$ is the set of the objects.

The kernel trick allows us to rewrite the decision hyperplane as:

$$H(\vec{x}) = \Big(\sum_{i=1..l} y_i \alpha_i \vec{x}_i \Big) \cdot \vec{x} + b =$$

$$\sum_{i=1..l} y_i \alpha_i \vec{x}_i \cdot \vec{x} + b = \sum_{i=1..l} y_i \alpha_i \phi(o_i) \cdot \phi(o) + b,$$

where y_i is equal to 1 for positive and -1 for negative examples, $\alpha_i \in \Re$ with $\alpha_i \geq 0$, $o_i \; \forall i \in \{1, .., l\}$ are the training instances and the product $K(o_i, o) = \langle \phi(o_i) \cdot \phi(o) \rangle$ is the kernel function associated with the mapping ϕ.

In recent years, kernel methods have attracted much interest to numerous applications in Natural Language Processing and Information Retrieval, due to their ability to

implicitly explore huge amounts of structural features automatically extracted from the original object representation. Kernel engineering can be carried out by combining basic kernels with additive or multiplicative operators or by designing specific data objects (vectors, sequences, and tree structures) for the target tasks.

2.2 Text features

The event detection task is considered as a classification problem, where categories are event types and the problem is framed into a machine learning framework. All the textual information connected to an event is considered, and the extracted features are processed as positive and negative examples. The feature set for our learning framework is described as follow.

1. w_i is text of the title, description, or the tag in each event
2. l_i is the word w_i in lower-case
3. $p1_i$, $p2_i$, $p3_i$, $p4_i$ are the four prefixes of w_i
4. $s1_i$, $s2_i$, $s3_i$, $s4_i$ are the four suffixes of w_i
5. f_i is the part-of-speech of w_i
6. g_i is the orthographic feature that test whether a word contains *all upper-cased, initial letter upper-cased, all lower-cased*.
7. k_i is the word form feature that test whether a token is a word, a number, a symbol, a punctuation mark.
8. o_i is the ontological features. We match w_i with the knowledge base as described in the following.

In our experiments, *run 1* was done without external resources (i.e., ontological features) whereas in *run 2* all the features were used.

2.3 Ontological features

A first important question related to the proposed method is whether machine learning techniques are necessary at all, and whether rich external resources could be injected to achieve better performance. In addition to previous works that mostly employ word clustering, we argue that if cluster features derived from unsupervised learning could bring some improvement, external resource may have similar effect. Given an ontology and knowledge base as a source of text-related knowledge, a text should be matched to the deepest subsumed child class. To this purpose, we used the ontology and knowledge base from KIM [12]. The KIM proton ontology contains about 300 classes, 100 attributes and relations. KIM World Knowledge Base (KB) contains about 77,500 entities with more than 110,000 aliases. Figure 1 shows an excerpt from the KIM ontology. Given a full ontology, we take the deepest subsumed child class that a text matches. For example, if the text "New York" matches with LOCATION, STATE, CITY, then CITY will be chosen since it is the deepest child class in the ontology. If a text matches with many classes in different branches, then a more general class will be chosen. For example, if the text "Washington" matches with PERSON and CITY which lie in two different branches in the ontology, then we choose the class ENTITY as the parent class for both PERSON and CITY.

2.4 Encyclopedic features

Wikipedia is an on-line encyclopedia created through the collaborative effort of millions of contributors. It has grown to be one of the largest online repositories, a multilingual resource with millions of articles available for a large number of languages. Concretely, official Wikipedias have been created for more than 200 languages with varying levels of coverage. The number of entries varies from a few pages to some million articles per language. Recently, Wikipedia has been shown as a valuable pre-processing step for many types of language analysis including measuring semantic similarity between texts [7], text classification [8], named entity recognition [3], relation extraction [15, 17, 18], co-reference resolution [6].

In this work, we employ Natural Language Processing (NLP) techniques, in particular we use Named Entity Recognition (NER) and Disambiguation to Wikipedia (*D2W*). Named entity recognition purposes the detection and classification of text segments into pre-defined categories. For example, given the sentence *"Essex, however, look certain to regain their top spot after Nasser Hussain and Peter Such gave them a firm grip on their match against Yorkshire at Headingley."*, a typical name entity recognizer should identify the two named entities "Nasser Hussain" of type Person, and "Headingley" of type Location.

Entity disambiguation refers to the detection and association of text segments with entities defined in an external repository. Disambiguation to Wikipedia (*D2W*) refers to the task of detecting and linking expressions in text to their referent Wikipedia pages. Figure 2 shows an example of *D2W*. Given a text "John McCarthy, 'great man' of computer science, wins major award.", a *D2W* system is expected to detect the text segment "John McCarthy" and link to the correct Wikipedia page *http://en.wikipedia.org/wiki/John_McCarthy_(computer_scientist)*, instead of other *John McCarthy* who are ambassador, senator or linguist.

We use the *NER* system of [19], which proposes a structural reranking framework for named entity recognition, and we use it to recognize proper names in the content of user's posts/comments. For Disambiguation to Wikipedia, we use the *D2W* system developed in [16, 20], which realizes the disambiguation to Wikipedia in multilingual context. Given a text, the *D2W* can detect and associate text segments with entities defined in Wikipedia. We extract named entities using *NER*, associate them with Wikipedia entries using *D2W* and use them as features in our learning framework.

2.5 Visual features

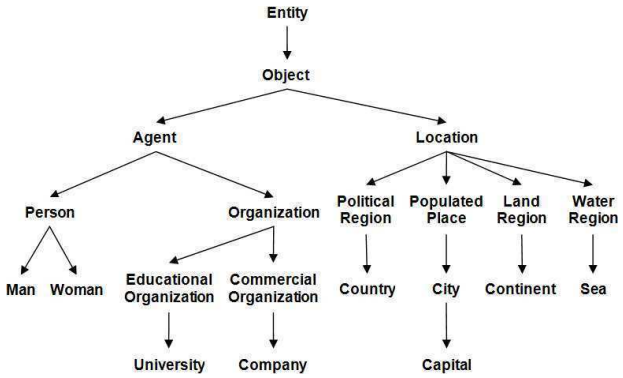

Figure 1: An excerpt from the ontology

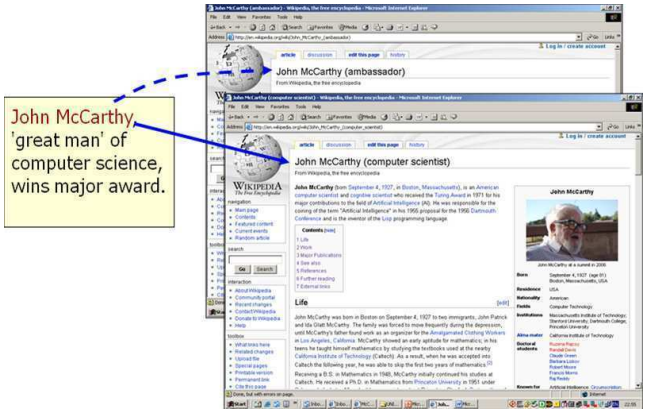

Figure 2: Disambiguation to Wikipedia

For *run 3*, the image feature extraction was performed in a similar manner as in [14], and the SVMs, with the same settings as in [14], was trained with the data available in the SED training set. Since the training set was unbalanced in the number of samples for each class, mainly towards a higher number of samples from the 'non-event' type, we balanced the training set samples used to train our SVM by reducing the number of samples from the 'non-event' class. *Run 4* used the same approach, but the classification followed a two-step classification procedure. First, a classifier was learned with only 'event' and 'non-event' classes, and second another classifier was trained with the remaining eight classes belonging to the different event types. *Run 3* and *4* did not used time information metadata associated with images.

2.6 Combine features

In *run 5*, we used a composite kernel to combine between text and visual features

$$CK = \alpha \cdot K_T + (1 - \alpha) \cdot K_V$$

where α is a coefficient, K_T and K_V is either the kernel applied to text or visual features. Some preliminary experiments on a validation set showed that the composite kernel yields the best performance with $\alpha = 0.5$.

3. USER-CENTRIC PARALLEL SPLIT-N-M -ERGE FRAMEWORK

In this section, we present a set of `user-centric parallel split -n-merge` algorithms, and the framework to cluster data crawled from social networks into different groups according to the events they depict. The whole framework is illustrated in Fig.3. Here, we assume that the data should have the following properties: `user-id`, `datataken`, `dataupload`, `title`, `description`, `tags`, and `URL` of photo. Except for `user-id`, the remaining properties could be NULL (but not all of them are never NULL at the same time).

3.1 User-Time Images

In order to group data belonging to the same user, the user-time image (`UT-image`) is proposed (see Fig.4). Each row of `UT-image` contains all data belonging to one user,

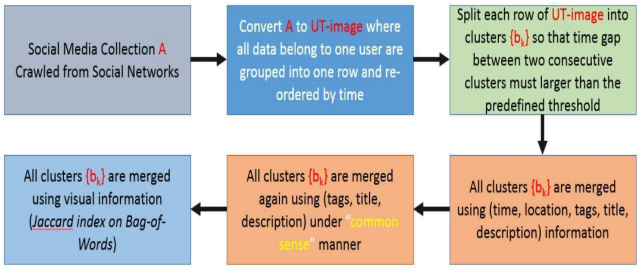

Figure 3: The proposed framework

and is ordered by date ascending. Therefore, `UT-image(i, j)` points to data created by i^{th} user at time j^{th}.

All data whose `time-taken` information is NULL, are grouped together and put at the beginning of each row.

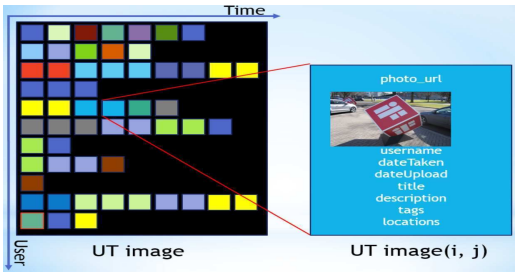

Figure 4: UT image

3.2 User-time-based Split Algorithm

As mentioned in previous sections, users play an important role in social networks. They generate, upload, and share data related to events they looked at or were involved in. Therefore, if data crawled from social networks can be grouped by users, events connected to the same users can be easily detected by clustering data into non-overlap time-ordered chunks. This depends on the obvious assumption that a user cannot attend at the same time two events whose locations are far away each other. Consequently, the temporal gap between two consecutive images taken from the same event is usually smaller than the time gap between two (consecutive) images belonging to two different events, reported by the same user. This observation leads to the first stage of the proposed framework: `user-time-based split` (see Alg.1).

For each row, any data whose `time-taken` information is NULL, is split as one cluster.

3.3 Time-Location-Tag-based Merge Algorithms

Social networks are the virtual place where users in the same community can share and exchange their data. Since people in the same community (e.g., culture, language, education, hobbies, etc.) can give the same "sound and prudent judgment based on a simple perception of the situation or fact"[1], they could probably tag the same event with similar words. Besides, with the support of high-tech devices (e.g., camera, smartphone, etc.), most recent media have time-stamp and possibly location (e.g., GPS) information. These observations are good clues to build the second and third

[1]www.merriam-webster.com

Algorithm 1 user-time-based split algorithm

```
1: procedure UTS(in A, in α, out B)
2:     B ← ∅;
3:     convert the original data A into UT-image;
4:     r ← number of row of UT-image;
5:     for i=1 to r do
6:         c ← number of column of row i^th of UT-image;
7:         for j=1 to c do
8:             t_j ← time-taken-of-UT-image(i,j);
9:             t_{j+1} ← time-taken-of-UT-image(i,j+1);
10:            if | t_j - t_{j+1} | ≧ α then
11:                split data at column j^th;
12:                B ← B ∪ new-cluster;
13:            end if
14:        end for
15:    end for
16:    return B;
17: end procedure
```

stages of the proposed framework to merge the clusters that belong to the same event: `time-location-tag-based merge` (see Alg.2) and `common-sense-based merge` (see Alg.3).

Algorithm 2 time-location-tag-based merge algorithm

```
1: procedure TLTM(in-out B, in α, in β, in γ, )
2:     for each cluster b_k in B do
3:         create time-taken-boundary T_k;
4:         create location-union L_k;
5:         create document D_k from tags, title, and description;
6:     end for
7:     do
8:         with any pair of cluster (b_k, b_l) ⊂ B do
9:         merging if 2/3 following conditions are hold
10:        {
11:        Tdistance(T_k, T_l) ≦ α;
12:        Ldistance(L_k, L_l) ≦ β;
13:        JaccardIndex(D_k, D_l) ≧ γ;
14:        }
15:        if did merge then
16:            update time-taken-boundary T_k;
17:            update location-union L_k;
18:            update document D_k ;
19:        end if
20:    while (can merge)
21:    return B;
22: end procedure
```

The `time-taken-boundary` T_k of cluster b_k is created by storing the period of time $(T_k.starttime, T_k.endtime)$ so that $\forall i : T_k.endtime \geqq b_k.\texttt{time-taken}_i \geqq T_k.starttime$.

The `location-union` L_k of cluster b_k is created by storing all non-empty (longitude, latitude).

The `document` D_k is built by applying several NLP techniques (e.g., Stemming, tokenization, etc.)[2] to filter and store meaningful words from tags, title, and description properties of b_k.

`Tdistance`$(T_k, T_l) \leq \alpha$ is TRUE if $(T_k \neq \emptyset \wedge T_l \neq \emptyset) \wedge ((0 \leq T_k.starttime - T_l.endtime \leq \alpha) \vee (0 \leq T_l.starttime - T_k.endtime \leq \alpha) \vee (T_l \cap T_k \neq \emptyset))$.

`Ldistance`$(L_k, L_l) \leq \beta$ is TRUE if $\exists l_k^i \neq \varnothing \wedge l_l^j \neq \varnothing$: `Haversine-distance`[3]$(l_k^i, l_l^j) \leq \beta$.

The Alg.3 is built based on the fact that there should be some major "keywords" that are selected with higher frequency by users who were involved in or are interested to the same event (e.g,. name or acronym of a conferences attended, name of musical group or singer in a concert, etc.). This algorithm will increase the chance of merging the clusters that have large "noise" in tags and cannot be successfully handled by `JaccardIndex` in Alg.2.

[2]http://nlp.stanford.edu/software/

[3]$en.wikipedia.org/wiki/Haversine\_formula$

Algorithm 3 common-sense-based merge algorithm

```
1: procedure CMM(in-out B, in γ)
2:     for each cluster b_k in B do
3:         process tf-idf on D_k and select the most common keywords to create ND_k set;
4:     end for
5:     do
6:         with any pair of cluster (b_k, b_l) ⊂ B do
7:         merging if JaccardIndex(ND_k, ND_l) ≧ γ;
8:         process tf-idf on ND_k and select the most common keywords and update ND_k set;
9:     while (can merge)
10:    return B;
11: end procedure
```

3.4 Visual-based Merge Algorithm

In [4], the authors proved that images related to an event of a given type share some common visual features that are characteristic for that event type. Therefore, the third stage of the proposed framework is `visual-based merge` (see Alg.4): two image sets belonging to two clusters are merged if they share a subset of common visual features.

Algorithm 4 visual-based merge algorithm

```
1: procedure VFM(in-out B, in θ)
2:     for each cluster b_k in B do
3:         BoW_k ← ∅;
4:         for each image img_k^i in b_k do
5:             calculate dense-RGB-SIFT;
6:             generate bag-of-words BoW_i^k;          ▷ 4096 words
7:             BoW_k ← BoW_k ∪ BoW_k^i;
8:         end for
9:     end for
10:    do
11:        with any pair of cluster (b_k, b_l) ⊂ B do
12:        merging if JaccardIndex(BoW_k, BoW_l) ≧ θ;
13:    while (can merge)
14:    return B;
15: end procedure
```

3.5 Parallel Split-n-Merge Scheme

Each algorithm of the proposed framework can be divided into a separate routine that can run independently, this parallelism is also present within some of the algorithms. For instance, in Alg.1, each row of `UT-image` can be treated as a separate thread. Thus, the processing time could be reduced thanks to parallel programming. For merging, we could divide set B into N subsets B_k, then Alg.2, 3, or 4 can apply for each set B_k. The results of all threads will be merged and divide again to $N/\#threads$ subsets. This progress will loop until no further merge can be applied. With the right policy, then, the proposed framework can help clustering social media events not only in off-line but also in on-line mode. This can cope with an important emerging problem nowadays: managing social media streams that require real-time processing.

4. EXPERIMENTAL RESULTS

The proposed framework has been tested and evaluated by using the datasets and evaluation tools offered by MediaEval 2013, Social Events Detection task [24].

4.1 Classification Results

Experimental setup

Our experiments aim at investigating the effectiveness of features and the combination of kernels for the event classification task. For this purpose, we combine the kernels over textual and visual features. Diverse features are applied individually for each type and in combination together. We consider our task as a classification problem where categories are event types. All the text belonging to an event is taken to extract features, which are used to create positive and negative samples.

Our learning framework is applied to cope with the **challenge 2**: *"For each image in the dataset decide whether the image depicts an event or not (in the latter case assign the no-event label to it)."* and *"For each image in the dataset that is not labelled as no-event, decide what type of event it depicts."*.

Since only five groups participated in the **challenge 2**, the proposed framework is compared to methods introduced by these groups: ADMRG[30], CERTH-1[22], QMUL[2], and VIT[10]. All te groups shared the same datasets and evaluation tools offered by MediaEval 2013. The comparison result is denoted in Table 1. In general, the proposed framework gained a promising result comparing to others.

We use the data in the MediaEval 2013 evaluation campaign corpus provided by the organizers. This data portion includes 27.754 photo instances, corresponding to eight event types. Every photo is assigned one of the eight event types: Concert, Conference, Exhibition, Fashion, Non_event, Other, Protest, Sports, and Theater_dance.

The data are processed using GATE platform[4] for tokenization, POS tagging and basic word features. We used Support Vector Machines to train and test our binary classifier. Here, event classification is formulated as a multiclass classification problem. The *One Vs. Rest* strategy is employed by selecting the instance with largest margin as the final answer. For experimentation, we use 5-fold cross-validation with the svm-light tool[5].

Results

We notice that in the comparison table 1, we gain better results than most of other teams, respectively. For challenge 2, the classification event vs. non-event is acceptable in almost every run, as well as the detection of some classes. Table 3 shows the results on each relation type in the best run, which combines text and visual features. The third and the fifth columns present the results without and with our new features.

Table 2 shows all runs of our approach. The first run does not use any visual feature, as well as third-parties information as compulsory required by MediaEval 2013 - SED task. The second row presents the results of only text features, but, injected with our new features. The third row describes the results in combination with visual features, but without new features. The fourth rows describe the results with visual information and with our new features, which are derived from *NER* and *D2W*. The integration of new features yields a good improvement of about 1.78% and 1.28%, respectively.

Obviously, we have followed the supervised machine learning for challenge 2, so it could not be learnt efficiently with only 36 positive instances of the class "fashion", it may be better if we used rule-based instead. Moreover, it is not trivial to provide a good detection on the class "other events", which is a rather undefined class. In the combination between text and visual features, the composite kernel did a good job with 5 classes out of 9 above 55%. In general, the proposed method proves to be very competitive, although there is still room for improvement, as we can try the feature set with other learning machines, or we can combine them in a learning framework to achieve better performance.

	F1	**Divergence F1**
Proposed Method	44.95	34.08
(with visual info)	42.20	31.45
ADMRG [30]	13.1	2.1
CERTH-1 [28]	33.44	22.61
QMUL [2]	50.00	NA
VIT [10]	7.48	NA

Table 1: Comparison Results on the test set

Run	**Pre**	**Rec**	**F1**
1 - compulsory run w/t visual info	33.72	71.48	45.83
2 - w/t visual info with *D2W*	35.31	71.51	47.61
3 - with visual info	50.46	57.12	53.58
4 - with visual info and *D2W*	52.46	57.04	54.86

Table 2: Cross-validation results of challenge 2

Event		**F1**		**F1**
conference		61.36		32.71
fashion		6.67		28.57
concert	without	58.66	with	60.23
non_event	NER	93.21	NER	94.62
sports	and	17.46	and	18.03
protest	D2W	61.75	D2W	71.52
other		7.91		15.27
exhibition		17.48		19.28
theater_dance		55.26		57.22

Table 3: Cross-validation results of challenge 2 on the best run with visual features-Without and with new features NER and D2W

4.2 Clustering Results

The proposed framework is applied to cope with the **challenge 1**: *"Cluster the entire dataset of all images included in the test set according to events they depict"*. The major difficulty here is the missing information about the number of clusters. Another challenge consisted in the fact that not all of properties' information are fully provided. For example, geographical information (45.9%), tags (95.6%), title (97.9%), and description (37.9%) w.r.t 437, 370 pictures assigned to 21, 169 events.

[4] http://gate.ac.uk/

[5] http://svmlight.joachims.org/

The proposed framework is compared to methods introduced by nine groups: ADMRG[30], CERTH-1[22], CERTH-2[28], ISMLL[33], QMUL[2], SOTON[27], TUWIEN[34], UPC[13], and VIT[10]. Also in this case, all the datasets and evaluation tools are those offered by MediaEval 2013 to all participants. The comparative results are shown in Table 4. In general, the proposed framework achieved promising result as compared to others.

	F1	**NMI**	**Divergence F1**
Proposed Method	0.9320	0.9849	0.8793
(with visual info)	0.9508	0.9931	0.9020
ADMRG [30]	0.8120	0.9540	0.7580
CERTH-1 [28]	**0.7041**	0.9103	0.6333
	0.7031	**0.9131**	**0.6367**
CERTH-2 [22]	**0.5701**	0.8739	0.5025
	0.5698	**0.8743**	**0.5049**
ISMLL [33]	0.8784	0.9655	NA
QMUL [2]	0.7800	0.9400	NA
SOTON [27]	0.9461	0.9852	0.8864
TUWIEN [34]	0.6900	0.8500	NA
UPC [13]	0.8833	0.9731	0.8316
VIT [10]	0.1426	0.1802	0.0025

Table 4: Comparison Results

Run	**F1**	**NMI**	**Divergence F1**
1 - compulsory run w/t visual info	0.9234	0.9829	0.8705
2 - w/t visual info	0.9316	0.9848	0.8788
3 - w/t visual info	0.9320	0.9849	0.8793
4 - with visual info	**0.9508**	**0.9931**	**0.9020**

Table 5: Each run with different parameters

Table 5 shows all runs of the proposed framework. The first run does not use any visual as well as third-parties information, as compulsory required by MediaEval 2013 -SED task. At the first run, the proposed method did gain a better result ($F1 = 0.9320$) compared to CERTH-1 ($F1 = 0.5698$), CERTH-2 ($F1 = 0.7031$), ADMRG ($F1 = 0.8110$), and QMUL ($F1 = 0.5900$) though most of them were using supervised methods and required some parameters to be manually tuned. The first run used only Alg.1 and 2 with $\alpha = 24$ hours, $\beta = 5km$, $\gamma = 0.2$. The second run was as the first one, except for $\alpha = 8$ hours and $\beta = 2km$. The third run used Alg.1, 2, 3, with same parameters as the second one. The last run was equal to the third one, with additional visual information $\theta = 0.3$ (i.e. Alg.4). The most interesting point is that results (e.g. F1, NMI, Div F1) of the proposed method increase at each step, while others don't. For example, in Table 4 CERTH-1 and CERTH-2 cannot get the best F1, NMI, and Div F1 at the same time when changing their parameters.

5. CONCLUSION

The `user-centric parallel split-n-merge` framework is introduced for unsupervised event-based clustering of social media. Series of simple algorithms are built based on characteristics of user's role (e.g., common sense, habits of taking, uploading and sharing data) in social networks. Major advantages of the proposed framework are the low computational complexity, easy implementation, parallelizability, generalization (less tuning parameters). The experimental results showed that the proposed framework can beat other methods not only on for accuracy but also for complexity and real-time processing.

In the future, the parallel stage will be investigated thoroughly and tested on cloud-computing to examine the ability of real-time processing. Moreover, a dictionary of (placename, longitude, latitude) will be built in order to get better results in location-based merging. Visual information will also be analyzed carefully to discover the optimal scheme to increase the qualification of the proposed framework.

For event classification, while it is straightforward the use of supervised machine learning to classify events, no previous works have tried kernel methods to combine text with visual features. As each kind of feature has its own characteristics, kernel methods offer nice properties to design a kernel function for each feature set, and combining them together. The combination has proved its robustness with a significant improvement in performance (from 45.83% to 53.58% with basic features, and from 47.61% to 54.86% with our new features).

As the data are obtained from social networks, basic natural language features, such as the word itself, prefixes, suffixes, and part-of-speech tag cannot guarantee a good performance. However, as we see, encyclopedic knowledge such as Wikipedia, could provide a great additional resource. We proposed new features that are derived from named entities (*NER* task) and Wikipedia entries (*D2W* task). Our study illustrates that those new features clearly provide an improvement with respect to the base model. Most interestingly, we showed that they provide improvements both with and without visual features. This makes clear that encyclopedic features are very useful for event classification task. The features can be used together with other learning algorithms to yield better results. Our composite kernel, which combines both kinds of features, can outperform the state-of-the-art.

6. REFERENCES

[1] A. Aggarwal and O. Rambow. Automatic detection and classification of social events. In *Empirical Methods in NLP*, pages 1024–1034. Association for Computational Linguistics, 2010.

[2] M. Brenner and E. Izquierdo. Mediaeval 2013: Social event detection, retrieval and classification in collaborative photo collections. In *MediaEval 2013*. MediaEval, 2013.

[3] W. Dakka and S. Cucerzan. Augmenting wikipedia with named entity tags. In *Proc. of IJCNLP*, 2008.

[4] M. Dao, J. Boato, and F. DeNatale. Discovering inherent event taxonomies from social media collections. In *ICMR*. IEEE, 2012.

[5] W. Dou, X. Wang, W. Ribarsky, and M. Zhou. Event detection in social media data. In *VisWwk Workshop on Interative Visual Text Analytics*. IEEE, 2012.

[6] T. Finin, Z. Syed, J. Mayfield, P. McNamee, and C. D. Piatko. Using wikitology for cross-document entity coreference resolution. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*, pages 29–35. AAAI Press, 2009.

[7] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[8] E. Gabrilovich and S. Markovitch. Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. *J. Mach. Learn. Res.*, 8:2297–2345, Dec. 2007.

[9] A. Gupta and R. Jain. Social life networks: A multimedia problem? In *Int. Conf. on Multimedia*. ACM, 2013.

[10] I. Gupta, K. Gautam, and K. Chandramouli. Vit@mediaeval 2013 social event detection task: Semantic structuring of complementatry information for clustering events. In *MediaEval 2013*. MediaEval, 2013.

[11] A. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Journal of Business Horizons - Elsevier*, 53(1):59–68, January-February 2010.

[12] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff. Semantic annotation, indexing, and retrieval. *Web Semant.*, 2:49–79, December 2004.

[13] D. Manchon-Vizuete and X. Giro-i Nieto. Upc at mediaeval 2013 social event detection task. In *MediaEval 2013*. MediaEval, 2013.

[14] R. Mattivi, J. Uijlings, F. G. De Natale, and N. Sebe. Exploitation of time constraints for (sub-)event recognition. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, pages 7–12, New York, NY, USA, 2011. ACM.

[15] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[16] T. V. T. Nguyen. Disambiguation to Wikipedia: A Language and Domain independent approach. In *Proc. of the 9th Asia Information Retrieval Societies Conference (AIRS)*, Singapore, December 2013.

[17] T. V. T. Nguyen and A. Moschitti. End-to-end relation extraction using distant supervision from external semantic repositories. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 277–282, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[18] T. V. T. Nguyen and A. Moschitti. Joint distant and direct supervision for relation extraction. In *Proc. of the 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, November 2011.

[19] T. V. T. Nguyen and A. Moschitti. Structural reranking models for named entity recognition. *Intelligenza Artificiale*, 6, December 2012.

[20] T. V. T. Nguyen and M. Poesio. Entity disambiguation and linking over queries using encyclopedic knowledge. In *Proc. of the 6th workshop on Analytics for Noisy Unstructured Text Data*, Mumbai, India, December 2012.

[21] A. Nurwidyantoro and E. Winarko. Event detection in social media: A survey. In *ICT for Smart Society (ICISS)*, pages 1–5. IEEE, 2013.

[22] D. Rafailidis, T. Semertzidis, M. Lazaridis, M. Strintzis, and P. Daras. A data-driven approach for social event detection. In *MediaEval 2013*. MediaEval, 2013.

[23] T. Reuter, S. Papadopoulos, V. Mezaris, P. Cimiano, C. de Vries, and S. Geva. Social event detection at mediaeval 2013: Challenges, datasets, and evaluation. In *Proceedings of MediaEval 2013*, Barcelona, Spain, October 2013.

[24] T. Reuter, S. Papadopoulos, V. Mezaris, P. Cimiano, C. de Vries, and S. Geva. Social event detection at mediaeval 2013: Challenges, datasets, and evaluation. In *MediaEval 2013*. MediaEval, 2013.

[25] R. M. Rifkin and T. Poggio. *Everything old is new again: a fresh look at historical approaches in machine learning*. PhD thesis, Massachusetts Institute of Technology, 2002.

[26] K. Robert Müller, S. Mika, G. Rätsch, K. Tsuda, , and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.

[27] S. Samangooei, J. Hare, D. Dupplaw, M. Niranjan, N. Gibbins, P. Lewis, J. Davies, N. Jain, and J. Preston. Social event detection via sparse multi-modal feature selection and incremental density based clustering. In *MediaEval 2013*. MediaEval, 2013.

[28] E. Schinas, E. Mantziou, S. Papadopoulos, G. Petkos, and Y. Kompatsiaris. Certh@mediaeval 2013 social event detection task. In *MediaEval 2013*. MediaEval, 2013.

[29] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[30] T. Sutanto and R. Nayak. Admrg@mediaeval 2013 social event detection. In *MediaEval 2013*. MediaEval, 2013.

[31] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer–Verlag, New York, 1995.

[32] V. N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

[33] M. Wistuba and L. Schmidt-Thieme. Supervised clustering of social media streams. In *MediaEval 2013*. MediaEval, 2013.

[34] M. Zeppelzauer, M. Zaharieva, and M. Del Fabro. Unsupervised clustering of social events. In *MediaEval 2013*. MediaEval, 2013.

Social Event Detection at MediaEval: a three-year retrospect of tasks and results

Georgios Petkos
CERTH-ITI
Thessaloniki, Greece
gpetkos@iti.gr

Symeon Papadopoulos
CERTH-ITI
Thessaloniki, Greece
papadop@iti.gr

Vasileios Mezaris
CERTH-ITI
Thessaloniki, Greece
bmezaris@iti.gr

Raphael Troncy
EURECOM
Sophia Antipolis, France
raphael.troncy@eurecom.fr

Philipp Cimiano
CITEC, University of Bielefeld
cimiano@cit-ec.uni-bielefeld.de

Timo Reuter
CITEC, University of Bielefeld
treuter@cit-ec.uni-bielefeld.de

Yiannis Kompatsiaris
CERTH-ITI
Thessaloniki, Greece
ikom@iti.gr

ABSTRACT

This paper presents an overview of the Social Event Detection (SED) task that has been running as part of the MediaEval benchmarking activity for three consecutive years (2011 - 2013). The task has focused on various aspects of social event detection and retrieval and has attracted a significant number of participants. We discuss the evolution of the task and the datasets, we summarize the set of approaches pursued by participants and evaluate the overall collective progress that has been achieved.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Social Event Detection, MediaEval, Multimedia

1. INTRODUCTION

The wealth of content uploaded by users on the Internet is often related to different aspects of real world activity. This presents an important mining opportunity and thus there have been many efforts to analyze such data. For instance, web content has been used for applications such as detecting breaking news [19] or landmarks [11]. A very interesting field of work in this direction involves the detection of social events in multimedia collections retrieved from the web. With social events we mean events which are attended by people and are represented by multimedia uploaded online by different people. Instances of such events are concerts, sports events, public celebrations or even protests. Figure 1 displays three broad categories of events (news, personal, entertainment) and several sample event types and images for each of them.

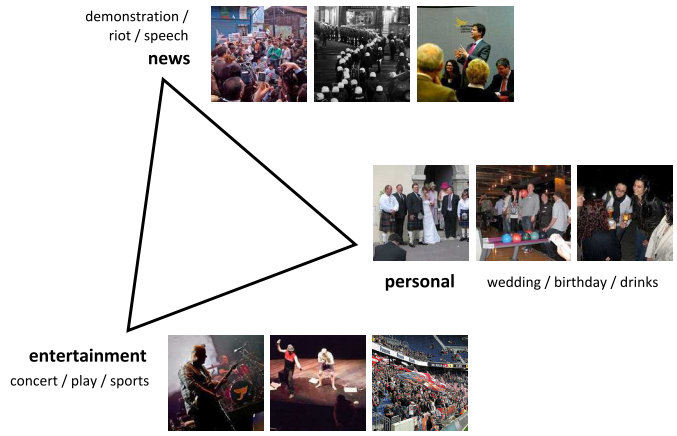

Figure 1: Broad event categories and sample images.

Indicative of the growing interest in the topic of detection of social events in web multimedia is that a relevant task has been organized in the last three years as part of the well-known MediaEval benchmarking activity. In this paper, we discuss the evolution of the task and the datasets in these three years, we summarize the set of approaches pursued by participants, and evaluate the overall collective progress that has been achieved.

The rest of the paper is structured as follows. In the next section we present the task objectives, used datasets and evaluation measures through the three years. Section 3 provides an overview of the pursued approaches and summarizes obtained results. Finally, Section 4 concludes the paper and discusses the directions to which the task and relevant research may turn to in the future.

ICMR 2014 SEWM Workshop, Glasgow, Scotland
.

Year	Challenge	Dataset
2011	Find events related to two categories: (a) soccer matches in Barcelona & Rome, (b) concerts in Paradiso & Parc del Forum	73,645 Flickr photos from five cities, May 2009
2012	Find events related to three categories: (a) technical events (e.g. exhibitions) in Germany, (b) soccer events in Hamburg and Madrid, (c) Indignados movement events in Madrid	167,332 Flickr photos from five cities, 2009-2011
2013	(a) Cluster photo collection into events, (b) attach YouTube videos to the discovered events	437,370 Flickr photos around upcoming or last.fm events, 2006-2012 and 1,327 YouTube videos around the events defined by the photos
	Categorize photos into eight event types or non-event	57,165 Instagram photos around event keywords, 27-29 April & 7-13 May 2013

Table 1: Overview of SED task from 2011 to 2013.

2. CHALLENGE DEFINITIONS, DATASETS AND EVALUATION

In the following, we review the task definitions, the used datasets and evaluation measures in the three years that the Social Event Detection task has been a part of the MediaEval benchmarking activity. At the end of the section, we provide a short discussion about the evolution of the task and the datasets. Table 1 provides a summary of the task challenges and datasets over the three years.

2.1 SED 2011

2.1.1 Challenges

The SED 2011 task had two challenges. In both, participants were provided with a set of images collected from Flickr (Section 2.1.2) and were asked to surface events of a particular type at particular locations. For each event, participants needed to find the set of relevant photos.

More particularly, the first 2011 challenge reads: *"Find all soccer events taking place in Barcelona (Spain) and Rome (Italy) in the test collection"*. Soccer events, for the purpose of this task, may include not only soccer games but also social events centered around soccer (e.g. celebration of winning the cup; as opposed to, for example, a single person playing with a soccer ball out in the street, which is not a *social* soccer event under the task's definition). For instance, the retrieved photos of such an event may include photos of a game being played, photos of fans inside the stadium during/a bit before/a bit after some game or photos of fans leaving the stadium after the end of a game. Examples of images that are relevant to soccer events are given in Fig. 2(a).

The second challenge is very similar and reads as follows: *"Find all events that took place in May 2009 in the venue named Paradiso (in Amsterdam, NL) and in the Parc del Forum (in Barcelona, Spain)"*. Some examples of relevant images can be seen in Fig. 2(b) and (c).

There are two differences between the two challenges. In the first challenge, both a topical (soccer) and a location criterion are defined for the events of interest, whereas in the second only a location criterion is defined (although the type of events that is held in these venues is easy to discover). Additionally, the specificity of the location of interest is different in the two challenges. These differences were deliberately opted for, in order to examine how the solutions of the participants would be affected.

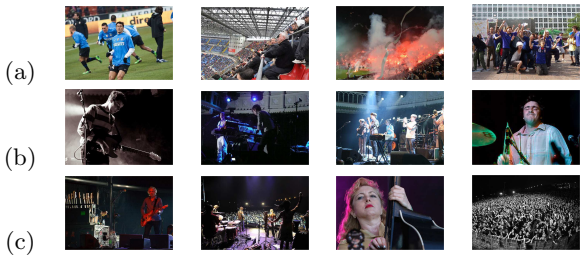

Figure 2: Example images of (a) soccer events, (b) concert events in Paradiso, Amsterdam, (c) concert events in Parc del Forum, Barcelona.

For both challenges, participants were allowed to use data from external resources (such as Wordnet, Wikipedia, or even visual concept detectors trained on external collections), provided that they did not relate to specific images of the test dataset (or any images given for specifying the sought events), and that their development and use did not benefit from any knowledge of the task's dataset and challenge definitions. Also, participants were asked to perform a baseline run without visual information (of course, the use of visual information in addition to the various image metadata were encouraged in subsequent runs).

2.1.2 Dataset

The dataset for the 2011 task consisted of 73,645 photos and was created by issuing appropriate queries to the Flickr web service through its web-based API. The collected photos represent the complete set of geotagged photos that were available for five different cities (i.e., Amsterdam, Barcelona, London, Paris and Rome, based on the geotags) and were taken in May 2009, further augmented with a few non-geotagged photos for the same cities and time period [27]. However, before providing the XML photo metadata archive (including any tags, geotags, time-stamps, etc. for the photos) to the task participants, the geotags were removed for 80% of the photos in the collection (randomly selected). This was done in order to simulate the frequent lack of geotags in photo collections on the Internet (including the Flickr collection) and to make the task more challenging (full knowledge of geotagging information would help a lot): since most images found on the web are not geotagged, participants would also need to consider tag and/or visual information for finding the complete set of relevant events and

images.

2.1.3 Ground truth and evaluation

The evaluation of the submissions to the 2011 Task was performed with the use of the ground truth event-media associations. As an aid, the cluster-based event detection framework of [17] was employed in generating this ground truth. Two evaluation measures were used:

- Harmonic mean (F-score) of Precision and Recall for the retrieved images. This measures only the goodness of the retrieved photos, but not the number of retrieved events, or how accurate the correspondence between retrieved images and events is.
- Normalized Mutual Information (NMI). This compares two sets of photo clusters (where each cluster comprises the images of a single event), jointly considering the goodness of the retrieved photos and their assignment to different events.

Both employed evaluation measures receive values in the range [0, 1], with higher values indicating a better agreement with the ground truth results.

2.2 SED 2012

2.2.1 Challenges

The challenges of the SED 2012 task were quite similar to those of the previous year: again a collection of images collected from Flickr (Section 2.2.2) was provided and participants were asked to find events of a particular type at particular locations (for each event, participants needed to provide the set of relevant photos). In contrast to the first year, however, the 2012 task had three challenges.

More particularly, the first challenge reads: *"Find technical events that took place in Germany in the test collection."* Technical events, for the purpose of this task, are public technical events such as exhibitions and fairs. The annual CeBIT exhibition, taking place in Hannover, is a good (but of course, not the only) example of such an event.

The second challenge reads: *"Find all soccer events taking place in Hamburg (Germany) and Madrid (Spain) in the test collection".*

The third challenge reads: *"Find demonstration and protest events of the Indignados movement occurring in public places in Madrid in the test collection."* The Spanish Indignados movement centers around a series of demonstrations and other protests taking place all over Spain in 2011-2012, which relate to the financial crisis outbreak as well as national politics in general.

As in the first year, variation in the challenges was deliberately introduced. First, the theme and location of queries was quite different between challenges. Additionally, the notion of "technical events" of the first task, although instantiated with a set of examples, was still somewhat vague and unclear and it was interesting to see how participants dealt with this. Most importantly, in contrast to the events that challenges one and two were concerned with, the events that were of interest to the third challenge were typically not scheduled, well-organized events (e.g., a technical fair that is typically announced several months before it actually takes place, or similarly a football game that is scheduled several days in advance) but rather spontaneous gatherings organized via social media channels.

Finally, as in the previous year, participants were allowed to use data from external resources, provided that they did not relate to specific images of the test dataset, and were asked to perform a baseline run that did not use any visual information.

2.2.2 Dataset

A collection of 167,332 photos (more than twice as many as in the 2011 edition of this task) was created by issuing appropriate queries to the Flickr web service through its web-based API. The collected photos were all licensed under a Creative Commons licence, and were captured between the beginning of 2009 and the end of 2011 (specifically, 51,019 photos captured in 2009, 53,080 in 2010 and 63,233 in 2011) by 4,422 unique Flickr users. Like in the previous year's dataset, all photos were originally geo-tagged; however, before providing the XML photo metadata archive (including any tags, geotags, time-stamps, etc.) to the task participants, the geotags were removed for 80% of the photos in the collection (randomly selected) in order to simulate a more realistic analysis scenario (as in SED 2011).

2.2.3 Ground truth and evaluation

The evaluation of the submissions to the 2012 SED task was performed with the use of ground truth that in part came from the EventMedia associations [27] (for challenge 1), and in part was the result of a semi-automatic annotation process carried out with the help of the CrEve tool [33] (for all three challenges). The two evaluation measures that were used in the first year, namely the F-score and NMI, were used in 2012 as well.

2.3 SED 2013

2.3.1 Challenges

The 2013 task had significant differences to the two previous years' tasks. Whereas in the previous years a single dataset that includes both event and non-event photos was provided and the challenges asked for the retrieval of events matching specific criteria, in 2013 two datasets were provided, and two new distinct challenges were defined.

More particularly, the first challenge reads: *"Produce a complete clustering of the image dataset according to events."* That is, the first challenge asked for a clustering of all images in the relevant dataset, according to the events that they depict. This comes in contrast to the challenges in the first two years, where a) not all images in the collection were related to some event and b) specific criteria were defined for the events of interest. Importantly, the target number of events was not given in this new challenge and therefore it had to be discovered from the data.

Also, there was an extension to Challenge 1 that introduced for the first time the use of video content. The description of this extension was the following: *"Assign all videos into the event sets you have created for the images in Challenge 1".* Participants were expected to use their created event clusters and assign the videos to them. As in the main task, here we also requested a complete assignment of the videos to events.

The second challenge reads as follows: *"Classify media into event types".* A second dataset was provided and the task was a) to decide for each image whether it depicts an event or not and b) for those images identified as depict-

ing some event, to identify the type of event. Essentially, this is a classification task that requires learning how event-related photos look like (both in terms of visual content and accompanying metadata). Eight event types were defined, and methods were expected to automatically decide to which type (if any) an unknown media item belongs.

The submissions to both challenges in 2013 were subject to the same conditions as those of the previous year, i.e. data from external resources could be used, provided that they did not relate to specific images of the test dataset. Also, participants of the first challenge were asked to perform a baseline run without exploiting visual information.

2.3.2 Datasets

The dataset for Challenge 1 consists of 427,370 pictures from Flickr and 1,327 videos from YouTube together with their associated metadata. The pictures were downloaded using the Flickr API, had an upload time between January 2006 and December 2012 and corresponded to 21,169 events. The events were determined by people using *last.fm* and *upcoming* machine tags, as described in Reuter et al. [21], and include sport events, protest marches, BBQs, debates, expositions, festivals or concerts. All of them are published under a Creative Commons license allowing free distribution. As it is a real-world dataset, there are some features (capture/upload time and uploader information) that are available for every picture, but there are also features that are available for only a subset of the images: geographic information (45.9%), tags (95.6%), title (97.9%), and description (37.9%). 70% of the dataset iwas provided for training, accompanied by its ground truth clustering. The rest was used for evaluation purposes.

The dataset for Challenge 2 is comparable to that of Challenge 1 except for the fact that the pictures were gathered from Instagram using the respective API. The training set was collected between 27th and 29th of April 2013, based on event-related keywords, and consisted of 27,754 pictures (after cleaning). The test set was collected between the 7th and 13th of May 2013 and consisted of 29,411 pictures. There are eight event types in the dataset: music (concert) events, conferences, exhibitions, fashion shows, protests, sport events, theatrical/dance events (considered as one category) and other events (e.g. parades, gatherings). As in the dataset for Challenge 1, some metadata were not present for all pictures: 27.9% of the pictures have geographic information, 93.4% come with a title and almost all pictures (99.5%) have at least one tag.

2.3.3 Evaluation and ground truth

The ground truth for both challenges was created by human annotators. It should also be noted that for the datasets of the second challenge in particular, several borderline cases were completely removed. The results of event-related media item detection were evaluated using three evaluation measures:

- F-score. This is applicable to both the first and the second challenge. It should be noted that for the second challenge, it was used for evaluating both for the classification of images into event types (F_{cat}) and the classification of event / non-event photos ($F_{E/NE}$).
- Normalized Mutual Information (NMI). This is applicable only to the second challenge.
- Divergence from a Random Baseline. All evaluation measures were also reported in an adjusted measure called *Divergence from a Random Baseline* [5], indicating how much useful learning has occurred and helping detect problematic clustering submissions (applicable to both C1 and C2).

2.4 Evolution of SED

The tasks in the first and the second year were quite similar. In both, the datasets contained both event and non-event images and the task was to retrieve sets of images that represent events matching given criteria. The task changed significantly in the third year, though: participants were asked to separately detect if images are related to some event (and if yes to what type) and to cluster event-related images in order to produce a set of events. In some sense, the problem presented in the first two years is split in two sub-problems (minus the retrieval / filtering that is required in the first two years). Thus, it can be said that there are two distinct eras in the evolution of the task, one that includes the first two years and one that includes the third.

Additionally, the datasets became larger from year to year. They also became richer in that over the years, with video data and an additional social media source (Instagram) made available in the 2103 edition.

3. APPROACHES

In this section we provide an overview of approaches followed by the participants. As discussed in the previous section, the SED task can be split into two distinct eras. In the first, the task was defined by asking for groups of photos, each of which represents an event that matches some criterion (e.g. soccer events in Madrid), whereas in the second, the task is split in two parts: a clustering and a classification part. Naturally, the approaches pursued by participants differ significantly between these two eras and thus it makes sense to present them independently.

3.1 SED 2011-2012

At a very high level, there are two types of approaches pursued by participants in the first two years:

1. A list of event descriptions that match the required criteria are fetched from online event directories (e.g. last.fm and Eventful) and subsequently the images in the provided datasets are matched to these descriptions.
2. A sequence of filtering or classification (in order to match the provided criteria) and clustering steps within the provided datasets is used to obtain the required events, without looking at external event directories.

Most approaches fall into the second class. For instance, the approaches described in [7, 12] belong to the first class, whereas the approaches described in [14, 16, 22, 29, 31, 28] belong to the second class.

Of course, there are important differences between the methods in each of these classes. For example, regarding the two methods that utilize external event directories, the essential difference is the way that matching takes place: in [7] photos were matched to event descriptions using Lucene queries, whereas [12] uses a probabilistic approach.

Some methods in the second class also utilize external sources, similarly to the methods falling into the first class, but they use sources that may assist in enriching the event-matching criteria. For instance, [1, 7, 22] use external sources such as the Google Geocoding API, DBPedia or Freebase to expand the representations of either locations or types of events so that more efficient filtering / classification can be achieved.

Other than that, methods in the second class differ in the set and sequence of filtering and clustering operations that they apply. Reasonably, the most common clustering criteria are time and location, as a unique combination of time and location clearly identifies a distinct event. For instance, in [16], a classifier applied at the first step assigns a city name to each item (either using geotags, if available, or textual information) and at the next step, all images that are related to the same city and occur on the same day are placed in a cluster/ event. Similarly, [22] forms groups of images related to distinct locations and then applies the Quality Threshold clustering algorithm on each group based only on time. To cater for the problem of missing location (e.g, when there is no metadata that can be used to assign a photo to a location), some approaches perform a post-processing step that applies reasonable heuristic rules to match such images to appropriate clusters. A different clustering strategy [4] first examines the images that belong to each user independently, clusters them using time and then combines the clusters produced using the other features.

Of particular interest is the approach in [24], where there is not a sequence of different clustering steps on an individual modality each time. Instead, there is a single clustering step that takes into account all modalities at once. To achieve this, the authors utilize a learned similarity metric that takes as input the set of modality-specific distances between a pair of items and predicts if that pair of items belong to the same event. Subsequently, the predicted intra-class relationships are organized in a graph in which nodes represent photos and the existence of an edge indicates a positive prediction of this "same event" model. The final events are produced by running a graph clustering algorithm on this graph. Additionally, in order to make the approach computationally feasible for larger datasets, a "candidate neighbor selection" step is used; i.e. the predictions of the "same event model" are evaluated between each photo in the dataset and its best matches according to each modality.

Different approaches achieved the best results in each of these first two years. The overall results for the first year are listed in Table 2. There were seven submissions and a different approach achieved the best results in each of the two challenges. In the first challenge, which involved the retrieval of soccer events, the best results were achieved by [16]. As mentioned before, this approach performed an early classification of photos to cities and then performed a partitioning of photos into buckets containing same day and same city photos. In the second challenge, which involved the retrieval of concert events at particular venues, the best results were achieved by [12] and [7] (one is best in terms of F-score and the other in terms of NMI). Interestingly, both these approaches follow the first high level approach that was mentioned before, i.e. they match the photos to event descriptions retrieved from online event directories. This indicates that despite the fact that such approaches may, in general, be limited only to events that are listed in online directories, they may also be quite effective.

	Challenge 1		**Challenge 2**	
	F-score	**NMI**	**F-score**	**NMI**
[1]	68.70	0.410	33.00	0.500
[7]	-	-	68.67	**0.678**
[12]	59.13	0.247	**68.95**	0.6171
[14]	10.13	0.026	12.44	-0.01
[16]	**77.37**	**0.630**	64.00	0.379
[22]	58.65	0.475	66.05	0.644
[29]	64.90	0.236	50.44	0.448

Table 2: SED 2011 results.

	Challenge 1		**Challenge 2**		**Challenge 3**	
	F-score	**NMI**	**F-score**	**NMI**	**F-score**	**NMI**
[31]	2.15	0.020	29.99	0.200	47.58	0.310
[28]	**84.58**	**0.724**	**90.76**	**0.850**	**89.83**	**0.738**
[24]	18.66	0.187	74.64	0.674	66.87	0.465
[2]	-	-	72.66	0.65	-	-
[4]	70.15	0.601	-	-	60.96	0.446

Table 3: SED 2012 results.

In the second year, there were five submissions. A summary of the results for the second year can be found in Table 3. In general, the results achieved in the first challenge are worse than those achieved in the other two and this is most likely due to the fact that the term "technical events" is a bit fuzzy. Also, the results for challenge 2 are better than those for challenge 3, and again, this is most likely due to the fact that soccer events are much more clear and uniform than the Indignados events. The best approach for all challenges was presented by [28]. It involves a city classification step and subsequently, for each city, topic detection with the use of LDA. Importantly, a manually constructed topic representing the topic of each of the three challenges was added to the results of LDA. Then, using the topic models learned, the photos that are relevant to the query of each challenge were retrieved. Events were identified by finding, for each topic and city of interest, the days for which the number of photos was above some threshold. Finally, a simple post-processing step that merges and splits events using some simple heuristic rules is performed.

3.2 SED 2013

In the third year, the two challenges had distinctly different objectives. In the following we discuss the approaches that the participants used for each of them separately.

The objective of the first challenge is similar in some sense but also has a significant difference to those of the previous two years. In particular, within SED 2013 all images in the collection were assumed to belong to some event and a complete clustering was required. This means that no filtering step was required. Since the photos in the collection were related to a set of heterogeneous metadata, this is essentially involved a multimodal clustering problem and therefore some form of fusion. There were 11 submissions and they mainly differed in the way that clustering and fusion is performed.

	Challenge 1		Challenge 2	
	F-score	**NMI**	F_{cat}	$F_{E/NE}$
[20]	0.570	0.873	-	-
[23]	**0.946**	**0.985**	-	-
[25]	0.704	0.910	0.334	0.716
[13]	0.883	0.973	-	-
[15]	0.932	0.984	**0.449**	**0.854**
[32]	0.780	0.940	-	-
[26]	0.812	0.954	0.131	0.537
[30]	0.878	0.965	-	-
[18]	0.236	0.664	-	-
[6]	0.142	0.180	-	-
[3]	0.780	0.940	0.332	0.721

Table 4: SED 2013 results.

Some approaches opt for a sequence of unimodal clustering operations. Again, the most common approach is to cluster by location and time. For instance, [20] first clusters items by location and then further clusters each initial cluster by time. Subsequently, they compute a per-modality weighted similarity measure between each non-geotagged image (that could not be clustered in the first step) and each of the clusters; and the initial clusters are expanded. There are also approaches that first consider a per-user clustering by time and then merge clusters by some fused similarity measure [13, 15].

There are again some approaches [25, 30] that perform fusion using a learned similarity model. In particular, [25] follows a graph-based approach similar to [24], whereas [30] uses it as part of a Quality Threshold clustering algorithm that is modified in a pseudo-incremental manner in order to make it applicable to a large dataset.

There are also a couple of approaches that have introduced some quite different and interesting aspects. In particular, [18] applies a Chinese Restaurant Process to cluster the photos. It computes a fused similarity metric as a linear combination of per-modality similarities using as weight the probability of two photos that have the same value in that modality to belong to the same cluster. They then use the merged similarity metric to compute the probability of assigning each photo to each cluster as part of an incremental and stochastic cluster assignment process. Another interesting approach is presented in [6], where textual features are used to compute an appropriate semantic similarity measure based on WordNet.

The overall results for the third year are listed in Table 4[1]. The best performing approach is that of [23]. It computes one affinity matrix per modality and then averages them to obtain an aggregate one that is used as part of either a DBScan or spectral clustering procedure. Additionally, to make computation of each affinity matrix feasible for large collections, a candidate neighbour selection step, similar to that of [24], is used. It is also important to note that due to the fact that the complete clustering challenge is somewhat easier than last years' challenges, and does not require the additional process of filtering/classification, in general the results obtained in this year are better that in the previous two in terms of absolute values of the evaluation measures.

In the second challenge, there were five submissions. All of them adopt a direct classification procedure, using an SVM classifier. The main difference between the methods pertains to the set of features used. Of interest is the approach in [25], where scalable Laplacian Eigenmaps are used in order to obtain in a semi-supervised manner the representation of the photos that is fed into the classifier. It is also interesting that [6] utilizes semantic similarity features. The best performing approach in the second challenge was [15], which also uses an SVM classifier, but introduces a very rich set of textual features, including also a set of ontological features.

4. CONCLUSIONS AND OUTLOOK

This paper presented an overview of the Social Event Detection task that has been part of the popular MediaEval benchmarking activity in the last three years. The task has two distinct eras; the one covers the first two years, whereas the other covers the third. In the first era, the challenge involved a single type of challenge: given a collection of images, to return sets of images that represent social events that match some specific criteria. In the second era, there was a deliberate decision to explicitly split the problem in parts: a clustering and a classification task, thus encouraging participants to explore a different approach with a distinct number of steps. We have seen that a large variety of interesting approaches has been used to deal with the challenges. For instance, we have seen approaches that utilize external event directories, perform complete clustering of collections, utilize different techniques to match images or sets of images to topics and locations, etc.

To conclude this paper, we discuss the outlook for the SED task and the problem of social event detection in general. As mentioned, the Social Event Detection task has been one of the more popular tasks in the MediaEval benchmarking activity. In particular, the number of participants in the third year was remarkable. Moreover, it has been encouraging that rather distinct approaches that have some clearly novel features have appeared. Therefore, it makes sense to continue the challenge and thus to further strengthen the relevant community. Indeed, the fourth edition of the task is currently being prepared. Due to the larger number of participants in the first challenge of the third year, it is planned to continue the complete clustering challenge. On the other hand, the photo classification challenge will be most likely discontinued, due to the relatively limited participation to it. Additionally, there are plans for bringing back the problem of event retrieval, this time as a distinct challenge. There are also plans for introducing another new challenge, focusing on summarization and presentation of clusters of images related to events.

Moving on with the discussion on the possible future directions in the field of social event detection, one first thing to note is that, so far, all versions of the SED task and all relevant work that has appeared elsewhere, have not tackled the challenge of detecting social events in a completely "into the wild" scenario. This means that there has not been an attempt to collect a really random (and large) collection of images from the web, without any prior knowledge about whether the images in it represent some social event or not, and to detect social events using it. Previous approaches, both as part of the SED task and other work, have used datasets that had a large ratio of event to non-event pho-

[1]The Divergence from Random Baseline was not included for the sake of uniformity with the first two years.

tos. This is because they have been crawled either using machine tags or appropriate spatio-temporal criteria. Alternatively, some approaches have utilized event directories and matched new content to event descriptions from these directories, e.g. once the time and location of some event is known, one may query Flickr for photos matching these criteria. However, such approaches are also limited and can only enrich already known events. Clearly though, due to the fact that a set of photos that has been really collected without any prior knowledge would typically have a very low percentage of event-related photos, a different approach than anything we have seen so far is required to deal with the problem of social event detection "into the wild".

The first step towards this direction could be the development of an accurate approach for classification of images as being or not related to some event. This is one of the reasons why in the third year a relevant challenge was organized. Some of the results were promising, however in order to deal with the complete scenario, even higher accuracy is required. To try to give a more quantitative feeling about this we will mention that during early experimentation for collecting the data for the second challenge of the third year, it was found that only roughly $1-2\%$ of images collected from a random stream were related to events. The current best achieved accuracy for characterizing an image as non-event is slightly lower that 90%, thus in a dataset of 1000 images, around 10-20 of them will in fact be event related, but roughly 100 of them will be classified as such, resulting in a very unclean set of images that will be further considered as being event-related. It should also be noted that improvement of the methods for identifying event-related images may have a benefit also on collection mechanisms; in particular, once some images have been identified with high confidence as event-related, they may be used to improve the collection of other event-related images by specifying appropriate search criteria.

Thus, it appears that the identification of event-related images and the generic "into the wild" scenario are two possible directions of future work in the problem of social event detection. Another possibility is the use of external sources in order to improve the results obtained from an event-agnostic approach. It is quite reasonable that, although event directories may contain only part of the real world events, they should be of value in order to refine the events identified from e.g. a clustering approach. Finally, results so far have relied mostly on metadata, rather than on image content; thus, novel approaches that make a more extensive use of visual features may surface in the future.

5. ACKNOWLEDGMENTS

This work was supported by the EC under contracts FP7-287975 SocialSensor, FP7-318101 MediaMixer and FP7-287911 LinkedTV.

6. REFERENCES

[1] M. Brenner and E. Izquierdo. Mediaeval benchmark: Social event detection in collaborative photo collections. In Larson et al. [9].

[2] M. Brenner and E. Izquierdo. Qmul @ mediaeval 2012: Social event detection in collaborative photo collections. In Larson et al. [10].

[3] M. Brenner and E. Izquierdo. Mediaeval 2013: Social event detection, retrieval and classification in collaborative photo collections. In Larson et al. [8].

[4] M. Dao, T. Nguyen, G., and F. De Natale. The watershed-based social events detection method with support from external data sources. In Larson et al. [10].

[5] C. de Vries, S. Geva, and A. Trotman. Document clustering evaluation: Divergence from a random baseline. 2012.

[6] I. Gupta, K. Gautam, and K. Chandramouli. Vit@mediaeval 2013 social event detection task: Semantic structuring of complementary information for clustering events. In Larson et al. [8].

[7] T. Hintsa, S. Vainikainen, and M. Melin. Leveraging linked data in social event detection. In Larson et al. [9].

[8] M. Larson, X. Anguera, T. Reuter, G. Jones, B. Ionescu, M. Schedl, T. Piatrik, C. Hauff, and M. Soleymani, editors. *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, October 18-19, 2013*, volume 1043 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[9] M. Larson, A. Rae, C. Demarty, C. Kofler, F. Metze, R. Troncy, V. Mezaris, and G. Jones, editors. *Working Notes Proceedings of the MediaEval 2011 Workshop, Santa Croce in Fossabanda, Pisa, Italy, September 1-2, 2011*, volume 807 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2011.

[10] M. Larson, S. Schmiedeke, P. Kelm, A. Rae, V. Mezaris, T. Piatrik, M. Soleymani, F. Metze, and G. Jones, editors. *Working Notes Proceedings of the MediaEval 2012 Workshop, Santa Croce in Fossabanda, Pisa, Italy, October 4-5, 2012*, volume 927 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.

[11] Y. Li, D. Crandall, and D. Huttenlocher. Landmark classification in large-scale image collections. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 1957–1964. IEEE, 2009.

[12] X. Liu, B. Huet, and R. Troncy. Eurecom @ mediaeval 2011 social event detection task. In Larson et al. [9].

[13] D. Manchon-Vizuete and X. Giró i Nieto. Upc at mediaeval 2013 social event detection task. In Larson et al. [8].

[14] M. Morchid and G. Linares. Mediaeval benchmark: Social event detection using lda and external resources. In Larson et al. [9].

[15] T. Nguyen, M. Dao, R. Mattivi, E. Sansone, F. De Natale, and G. Boato. Event clustering and classification from social media: Watershed-based and kernel methods. In Larson et al. [8].

[16] S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, and A. Vakali. Certh @ mediaeval 2011 social event detection task. In Larson et al. [9].

[17] S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, and A. Vakali. Cluster-based Landmark and Event Detection on Tagged Photo Collections. *IEEE Multimedia*, 18(1):52–63, February 2011.

[18] A. Papaoikonomou, K. Tserpes, M. Kardara, and T. Varvarigou. A similarity-based chinese restaurant process for social event detection. In Larson et al. [8].

[19] S. Phuvipadawat and T. Murata. Breaking news detection and tracking in twitter. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 3:120–123, 2010.

[20] D. Rafailidis, T. Semertzidis, M. Lazaridis, M. Strintzis, and P. Daras. A data-driven approach for social event detection. In Larson et al. [8].

[21] T. Reuter and P. Cimiano. Event-based classification of social media streams. In *Proceedings of the 2nd ACM Intern. Conf. on Multimedia Retrieval*, page 22. ACM, 2012.

[22] M. Ruocco and H. Ramampiaro. Ntnu@mediaeval 2011 social event detection task. In Larson et al. [9].

[23] S. Samangooei, J. Hare, D. Dupplaw, M. Niranjan, N. Gibbins, P. Lewis, J. Davies, N. Jain, and J. Preston. Social event detection via sparse multi-modal feature selection and incremental density based clustering. In Larson et al. [8].

[24] E. Schinas, G. Petkos, S. Papadopoulos, and Y. Kompatsiaris. Certh @ mediaeval 2012 social event detection task. In Larson et al. [10].

[25] M. Schinas, E. Mantziou, S. Papadopoulos, G. Petkos, and Y. Kompatsiaris. Certh @ mediaeval 2013 social event detection task. In Larson et al. [8].

[26] T. Sutanto and R. Nayak. Admrg @ mediaeval 2013 social event detection. In Larson et al. [8].

[27] R. Troncy, B. Malocha, and A. Fialho. Linking Events with Media. In *Proc. Open Track of the Linked Data Triplification Challenge at I-SEMANTICS'10*, Graz, Austria, September 2010.

[28] K. Vavliakis, F. Tzima, and P. Mitkas. Event detection via lda for the mediaeval2012 sed task. In Larson et al. [10].

[29] Y. Wang, L. Xie, and H. Sundaram. Social event detection with clustering and filtering. In Larson et al. [9].

[30] M. Wistuba and L. Schmidt-Thieme. Supervised clustering of social media streams. In Larson et al. [8].

[31] M. Zeppelzauer, M. Zaharieva, and C. Breiteneder. A generic approach for social event detection in large photo collections. In Larson et al. [10].

[32] M. Zeppelzauer, M. Zaharieva, and M. del Fabro. Unsupervised clustering of social events. In Larson et al. [8].

[33] C. Zigkolis, S. Papadopoulos, G. Filippou, Y. Kompatsiaris, and A. Vakali. Collaborative Event Annotation in Tagged Photo Collections. *Multimedia Tools and Applications*, 2012.