CERTH @ MediaEval 2013 Social Event Detection Task

Emmanouil Schinas, Eleni Mantziou, Symeon Papadopoulos,
Georgios Petkos, Yiannis Kompatsiaris
CERTH-ITI
Thessaloniki, Greece
{manosetro, lmantziou, papadop, gpetkos, ikom}@iti.gr

ABSTRACT

This paper describes the participation of CERTH in the Social Event Detection Task of MediaEval 2013. For Challenge 1, we used the concept of *same event model* to construct a graph of items, in which dense sub-graphs correspond to event clusters. The F1 score and NMI for our best run are 0.7041 and 0.9131, respectively. For Challenge 2, we used an efficient manifold learning method to assign images to specific event types. Our best run for Challenge 2 achieves a F1 score of 0.3344 (0.7163 for the event/non-event case).

1. INTRODUCTION

The paper presents the approaches devised by CERTH for the Challenges of the MediaEval 2013 Social Event Detection (SED) task. Challenge 1 calls for the detection of social events in a set of Flickr images. Challenge 2 calls for the classification of images to a set of event types (incl. non-event). Reuter et al. [5] describe the task in detail.

2. PROPOSED APPROACH

2.1 Overview of Method in Challenge 1

The approach to tackle Challenge 1 is based on a *learned similarity metric* [4], which in the following we call the Same Event Model (SEM). A SEM takes as input the set of per modality dissimilarities between two images, and produces a prediction value in the range [0,1] that indicates the possibility that the two images are of the same event. The SEM is used to construct a graph of same event relationships: pairs of images for which the output of SEM is above some threshold are connected. Eventually, the nodes of the graph are clustered using an efficient community detection algorithm, the Structural Clustering Algorithm for Networks (SCAN) [7] to obtain a set of candidate events. To avoid evaluating the output of the SEM for each pair of images, we introduce a candidate neighbours selection step, as in Reuter and Cimiano [4]: for each item in the collection, we find its nearest neighbours in each modality and only compare it to them. A schematic of our approach is shown in Fig. 1. It is very similar to that of Philip and Cimiano [4], with the difference that they use a learned similarity metric that works on image-event pairs (where the event is represented as the centroid of the images assigned to it), whereas our approach works on image-image pairs, similarly to Petkos et al. [3]. The motivation is that in an image-event approach, it is possible that the centroid may be an inaccurate representation of the event, especially in the case that some images are incorrectly assigned to it. Due to this image-image approach, we also use a different clustering mechanism, graph clustering, instead of the incremental threshold-based clustering of Reuter and Cimiano [4].

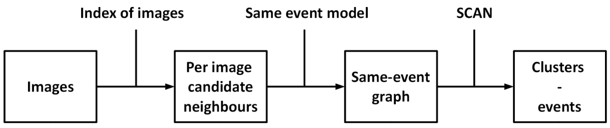

Figure 1: Schematic representation of the approach

2.2 Overview of Method in Challenge 2

For Challenge 2, we made use of SMaL, a **S**calable **Ma**nifold **L**earning framework [2] that is based on Semi-Supervised Learning (SSL) by constructing Laplacian Eigenmaps (LEs) approximately. The main problem in SSL is to build a $n \times n$ similarity matrix between labelled and unlabelled images, which is time consuming for large datasets. SMaL tackles this problem by estimating a smaller covariance matrix, where it is hypothesized that the data $x_i \in \Re^d$ are samples from a distribution $p(x)$. Rotating the data to be as independent as possible, $s = Rx$, can result in a $B \times B$ histogram of bins, using only marginal distributions that approximate the density $p(s)$ of the rotated data. Then, we define eigenfunctions g corresponding to $B \times B$, which can be seen as approximations of the LEs as $n \to \infty$ [1]. An interpolation step to the target dimension k follows and in the end, the $n \times k$ approximate LE vectors are derived. In the final step, a linear classifier is trained using the approximate vectors of the labelled items as input. In our implementation, we opted for the use of linear Support Vector Machine (SVM).

3. EXPERIMENTS

3.1 Runs Description in Challenge 1

For the first challenge we experimented with a variety of classifiers to build SEMs. We obtained the best results using SVMs, but Decision Trees produced comparable results. The inputs for SEM are the following 11 features: user (1 if both images have been uploaded by the same user, 0 otherwise), textual (title, tags and description, similarity computed using BM25 and cosine), taken and upload time, spatial (if exists) and visual information (SURF descriptors

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

	1	2	3	4	5
F1	0.7031	**0.7041**	0.7031	0.7031	0.6434
NMI	**0.9131**	0.9103	**0.9131**	**0.9131**	0.8978
Divergence	**0.6367**	0.6333	**0.6367**	**0.6367**	0.5840

Table 1: Challenge 1 results summary.

aggregated using a VLAD scheme [6], similarity computed using Euclidean distance). We used and evaluated several combinations of the input features. We achieved the best results using textual, temporal, spatial and user features. More specifically the corresponding trained SEM which was used for all submitted runs achieved 99% accuracy on the classification of positive pairs and 99.7% on negative. By using all available features, we get slightly worse results (98.75% for positive and 99.7% for negative examples). Note that even with only the user as a feature the trained model achieves 95% and 99% for positive and negative examples respectively. The SEM for the final submission was trained on the full data of the training set. Regarding the retrieval of the candidate neighbours of each item, the 150 nearest neighbours with respect to the textual features were considered, 200 with respect to time, 50 with respect to location (when it exists), and 100 for visual. For the construction of the items graph we experimented with various values of threshold W affecting the insertion of edges between items. By increasing the threshold, we get a more sparse graph but with a negative impact to the produced results. We have found that a good trade-off between the computational cost and the efficiency of the method is a value of 0.5. For this reason, we used the default value in Runs 1, 2, 3 and 4. In Run 5, we pruned the graph by using $W = 0.9$. At a post-processing step we attempt to merge hubs and outliers to the set of event clusters. Regarding the hubs we attempt to merge each of them with the community with which they have more edges under the condition that this number is greater than a predefined threshold R. In Runs 1, 2 and 5 we used a threshold of 5 links. In Run 3 we make the assignment easier by requiring only 2 links, in contrast with Run 4 where more than 12 links are required. The remaining hubs form single-item events. Regarding outliers we can either follow the same approach as hubs (Run 2) or to consider them all as single-item events (Runs 1, 3, 4 and 5).

3.2 Runs Description in Challenge 2

Different runs were produced with respect to the employed features and classification strategy:

- **Features:** In Run 1, we considered the 1000 most frequent tags, and then applying pLSA using 200 latent topics. In Run 2, we used dense sampling for selecting the keypoints, SIFT features were extracted using codebooks of k=64 visual words and learned using the k-means algorithm. VLAD was performed for the aggregation. The final vectors were power (a=0.5) and L2 normalized and then PCA reduced to 512-dimensional vectors. The same features were also used in Run 5. In Runs 3 and 4, we used both textual and visual features and fused the corresponding LE vectors.
- **Classification strategy:** We used two variants: (a) *max score* (Runs 1, 4 and 5), in which we select the concept of which the detector produced the highest prediction score, (b) *if-max score* (Runs 2 and 3), in which we assign the image to the highest scoring concept, only if that score is higher than a threshold T_E, otherwise assigned to the *no-event* class. T_E was empirically determined by averaging the maximum prediction scores for the images of a validation set ($\approx$ 30% of the training set).

	1	2	3	4	5
F1	0.1105	0.2570	**0.3344**	0.3046	0.2411
F1 Div.	0.0642	0.1516	0.2261	0.2062	0.1494
F1 (E/NE)	0.2201	0.6870	**0.7163**	0.6536	0.5989
F1 Div. (E/NE)	-0.0127	0.1900	0.2157	0.1893	0.1521

Table 2: Challenge 2 results summary.

4. RESULTS AND DISCUSSION

From Table 1, we note that Run 5 leads to the worst results. This indicates that pruning on the edges of the graph has negative impact on the event detection accuracy. The other results seem to be very similar. This leads us to conclude that our method is insensitive to the different parameters involved. For example, Runs 1, 3 and 4 differ regarding the threshold R (5, 12 and 3 respectively) that controls the merging of hubs to adjacent communities. Although this step improves the final results, the threshold value does not seem to affect them significantly.

For Challenge 2, Run 1 has the lowest performance ($F_1 = 0.1105$), which indicates that visual features are more useful than textual. However, their fusion performs best ($F_1 = 0.3344$), revealing a complementary role. We also note that in concepts that are related to each other (e.g. concert and theater), visual features do not perform well. Moreover, threshold T_E used in Runs 2 and 3 is prone to produce many false negatives.

In the future, for Challenge 1 we plan to explore different graph construction strategies and community detection algorithms that consider the edge weights. Also, we plan to use more sophisticated approaches for merging or splitting candidate events based on temporal and spatial information. For Challenge 2, we intend to explore different features for further improving the event type detection accuracy.

5. ACKNOWLEDGEMENTS

This work is supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975.

6. REFERENCES

[1] R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised learning in gigantic image collections. In *NIPS*, pages 522–530. 2009.

[2] E. Mantziou, S. Papadopoulos, and I. Kompatsiaris. Large-scale Semi-Supervised Learning by Approximate Laplacian Eigenmaps, VLAD and Pyramids. In *WIAMIS*, 2013.

[3] G. Petkos, S. Papadopoulos, and Y. Kompatsiaris. Social event detection using multimodal clustering and integrating supervisory signals. In *ICMR'12*, page 23. ACM, 2012.

[4] T. Reuter and P. Cimiano. Event-based classification of social media streams. In *ICMR'12*, page 22. ACM, 2012.

[5] 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 Workshop*, Barcelona, Spain, Oct 18-19 2013.

[6] E. Spyromitros-Xioufis, S. Papadopoulos, I. Kompatsiaris, G. Tsoumakas, and I. Vlahavas. An empirical study on the combination of surf features with vlad vectors for image search. WIAMIS, 2012.

[7] X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger. Scan: a structural clustering algorithm for networks. KDD '07, pages 824–833, New York, NY, USA, 2007. ACM.