UPC at MediaEval 2013 Social Event Detection Task

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

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

ABSTRACT

These working notes present the contribution of the UPC team to the Social Event Detection (SED) task in MediaEval 2013. The proposal extends the previous PhotoTOC work in the domain of shared collections of photographs stored in cloud services. An initial over-segmentation of the photo collection is later refined by merging pairs of similar clusters.

1. INTRODUCTION

These working notes describe the algorithms tested by the UPC team in the MediaEval 2013 Semantic Event Detection (SED) task. The reader is referred to the task description [2] for further details about the study case, dataset and metrics. Our team participated only in Task 1, where all image were to be clustered in events.

The proposed approach is aimed at a light computational solution capable of dealing with large amounts of data. This requirement is specially sensible when dealing not only with large amounts of data, but also with large amount of users. The SED task describes a dataset with photos from different users, so that the events to be detected affect several users. This set up suggests a computational solution to be run on a centralised and shared service on the cloud, in contrast to other scenarios where each user data can be processed on the client side. Any computation on the cloud typically implies an economical cost on the server which, in many cases, is not directly charged on the user, but assumed by the intermediate photo management service. 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 the images.

In addition, the SED task presents an inherent challenge due the incompleteness of the photo metadata. The provided dataset contains real photos with real missing or corrupted information; such as non-geolocalised images, or identical time stamps for the moment when the photo was taken but also uploaded. These situations are common specially when dealing with online services managing photos, which present hetereogenous 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. In our approach, no external data is used, so all submitted runs belong to the *required* type (as specified in the SED overview paper [2]).

These working notes is structured as follows. Section 2 describes the existing PhotoTOC system, which has been adopted as an initial oversegmentation of the dataset. Later, Section 3 presents how the oversegmented clusters are merged considering different metadata sources. The performance of the solution is assessed in Section 4 with the results obtained on the MediaEval SED 2013 task. Finally, Section 5 provides the insights learned and points at future research directions.

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

2. RELATED WORK

The adopted solution is inspired by an original work from Microsoft Research[1] named *PhotoTOC (Photo Table of Contents*. In this previous design, photos are initially sorted according to their creation time stamp and they 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. In particular, a new event is created whenever the criterion show in Equation 1 was satisfied,

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

where PhotoTOC empirically set the configuration parameters to $d = 10$ and $K = log(17)$.

When the time creation is missing in the EXIF metadata, the PhotoTOC uses the file creation time. Whenever a cluster is larger than 23, this event is considered too large and it is split based on color features. This content-based clustering algorithm generates 1/12 the amount of photographs in the large cluster

The main drawback of PhotoTOC approach was the need of an image processing analysis to estimate the content-based similarity. The visual modality was discarded and substituted by the geolocation and textual labels as additional information to the time creation. In addition, in the SED task images from different users were considered taken from different cameras and point of view, all of this driving to a less reliable visual analysis. There is no guarantee either that the empirically set values proposed in PhotoTOC would be useful in another dataset, nor it is clear from the paper how they were estimated.

3. APPROACH

Two solutions have been tested in our submission, both of them having a common starting point in the time-based clustering solution proposed by PhotoTOC.In both solutions, the initial time-based clusters are compared based on associated geolocation, textual labels and user IDs. The first solution relies on weights for each criterion which have been manually tuned, while the second introduces an estimation of the relevance of each feature type.

3.1 User and time-based over-segmentation

The first step in the proposed solution considers the photos of each user separately. The time-based clustering algorithm proposed by PhotoTOC independently optimising configuration parameters K and d with the training dataset provided by MediaEval. The obtained values were $K = log(150)$ and $d = 40$, which clearly differ from the ones proposed in [1]. During this first stage, those images whose *Date taken* matches their *Date uploaded* are not processed, as their time stamp is considered corrupted.

As a result, an over-segmentation of mini-clusters is obtained. Each of them is characterised by its averaged time, averaged geolocation, aggregated set of textual labels and associated user ID. These are the features used in the posterior stages to assess the similarity between the mini-clusters.

3.2 Cluster merges

The set of time-sorted clusters is sequentially analysed in increasing time value. Each cluster is compared with the forthcoming 15 clusters, a time window set to avoid excessive computational time. Two clusters are merged whenever a similarity measure is above an estimated threshold. The submitted runs have considered two options for assessing this similarity: a first one that adopts binary decision based on each criterion which are manually weighted, and a second one where each individual similarity measure is normalised and later fused with a learned weight.

Method 1: Binary decisions and manual weights

This method compares each pair of clusters separately and takes a binary decision for each criterion. The geolocation coordinates are compared with the Haversine distance, the textual label set with the Jaccard Index and the user IDs with a simple binary decision. The three binary decisions are linearly fused with a weighting scheme of 0.2 per geolocation, 0.2 for text and 0.4 for user ID. Two clusters are merged if the fused combination exceeds 0.3.

The binary decision for each criterion is based on a specific similarity threshold learned after optimisation on the training dataset. This process has assumed independence between the different features, so each of them has been treated separately.

Method 2: Weighted fusion of normalised distances

This second solution emerged as a need for a more refined algorithm to combine the different metadata features. In this case the individual and binary decisions are for a single and fused similarity value.

This fusion requires a normalization of the distance values based on the provided training data. This normalization was based after the computation of the distances between 3,000 random pairs of photos selected from the training set and belonging to the same event. The estimated mean and deviation were used to compute the value of the phi function, which is basically a mapping of the z-score between 0.0 and 1.0.

After normalization, it is still necessary to estimate the weight of each modality to be later applied to the linear fusion. These weights were estimated according to the individual gain of each type of features studied Method 1. Results shown in Table 1 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 textual labelling have similar relevance. These experimental values validate the empirical proposal adopted in Method 1.

	Time	Geo	Label	User
Geolocated	0.06	0.28	0.22	0.44
No geolocated	0.08	-	0.30	0.60

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

Finally, the training dataset was used again to estimate the merging threshold for this fused score. The experiments indicated a maximum F1-score for values between 0.3 and 0.6, for which a final threshold of 0.5 was adopted.

4. EXPERIMENTS AND RESULTS

The UPC participated in Challenge 1 with the results shown in Table 2. The more optimised Method 2 corresponds to Run 1, while Runs 2 and 3 correspond to Method 1 with an optimisation with respect to F1 or NMI, respecively. As expected, the values obtained for Method 2 outperform the two runs associated to Method 1.

	F1	NMI	Divergence F1
Method 1 (F1)	0.8798	0.9720	0.8268
Method 1 (NMI)	0.8753	0.9710	0.8220
Method 2	0.8833	0.9731	0.8316

Table 2: UPC results in Challenge 1.

5. CONCLUSIONS

The presented technique has allowed a fast resolution of the photo clustering of images based only on numerical and textual metadata. The obtained results seems reasonable to assist real users in the organisation of shared collections of photographs. However, the authors consider that presented work may still benefit with an optimised set of similarity thresholds adapted to the type of event.

6. REFERENCES

[1] J. C. Platt, M. Czerwinski, and B. Field. Phototoc: automatic clustering for browsing personal photographs. In *Proc. 4th Pacific Rim Conference on Multimedia.* , vol. 1, pp. 6-10 Vol.1, 2003.

[2] 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 Workshop*, Barcelona, Spain, October 18-19 2013.