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.

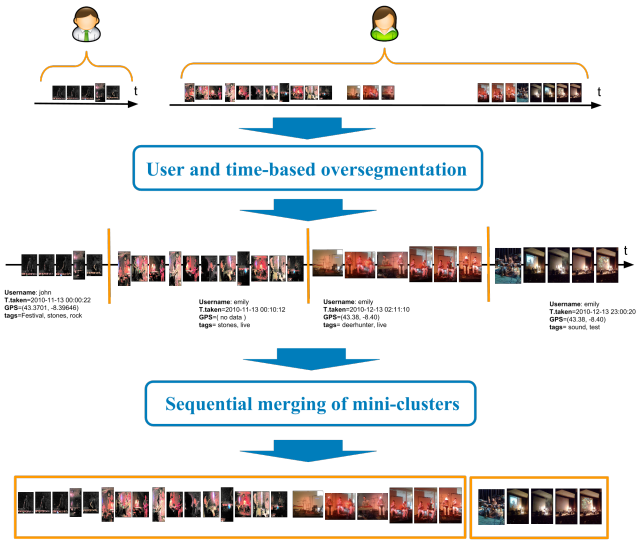

Figure 1: System architecture.

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.

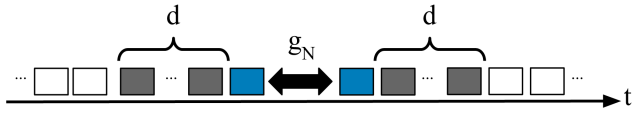

Figure 2: A new event boundary is created when time difference g_N exceeds the average time differences in the neighbourhood defined by d.

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.

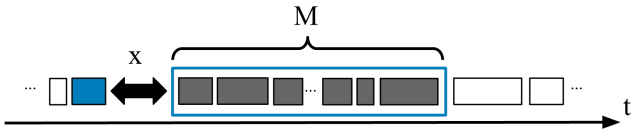

Figure 3: Each mini-cluster is compared to the following M mini-clusters, and merged if their relative distance x is below a certain threshold .

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.

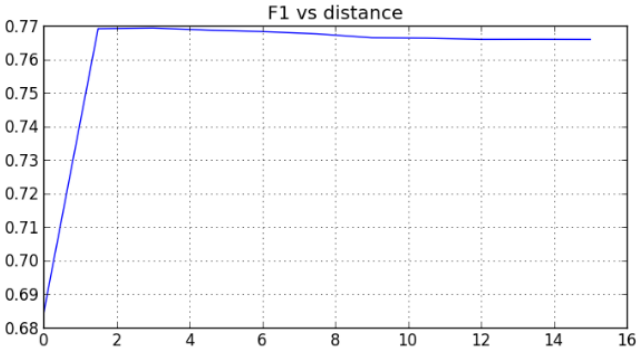

Figure 4: Evolution of the F1-Score with respect to a merging threshold based on geolocations.

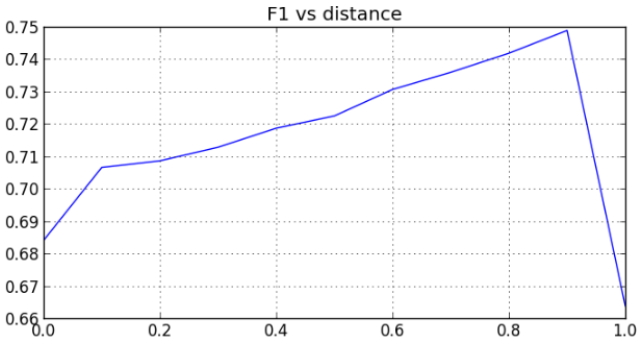

Figure 5: Evolution of the F1-Score with respect to a merging threshold based on tags.

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

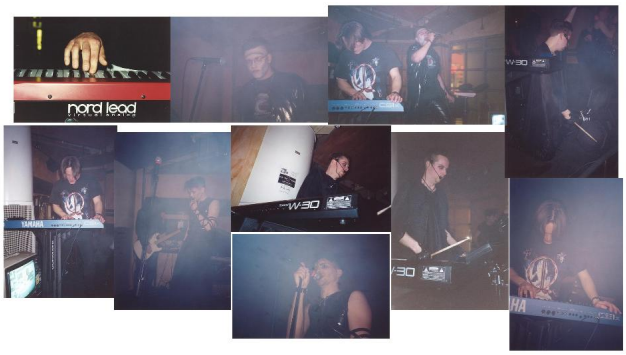

Figure 6: Detected event that combines photos of different qualities.

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.

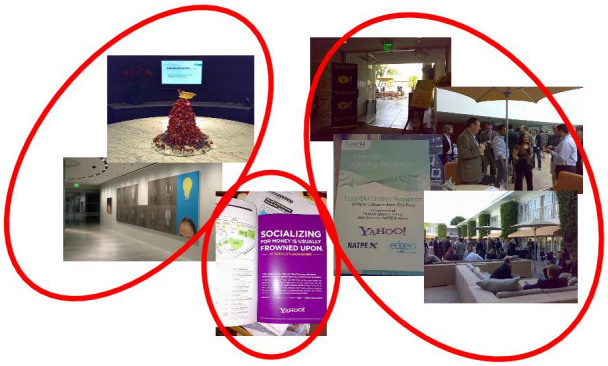

Figure 8: An event is incorrectly split in three.

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.