Synchronization of Multi-User Event Media at MediaEval 2015: Task Description, Datasets, and Evaluation

Nicola Conci
DISI - University of Trento
Trento, Italy
nicola.conci@unitn.it

Francesco De Natale
DISI - University of Trento
Trento, Italy
francesco.denatale@unitn.it

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

Mike Matton
VRT
Belgium
mike.matton@innovatie.vrt.be

ABSTRACT

The objective of this paper is to provide an overview of the Synchronization of Multi-User Event Media (SEM) Task, which is part of the MediaEval Benchmark for Multimedia Evaluation. The SEM task was initially presented at MediaEval in 2014, with the goal of proposing a challenge in aligning multiple users' photo galleries related to the same event but with unreliable timestamps. Besides aligning the pictures on a common timeline, participants were also required to detect the sub-events and cluster the pictures accordingly. For 2015 we have decided to extend the task also to other types of media, thus including audio and video information for a more complete and diversified representation of the analyzed event.

1. INTRODUCTION

The ever increasing number of devices for the collection of personal data (smartphones, portable cameras, audio recorders) has lead to the generation of huge amount of data, which can be either stored for personal records or shared among friends, relatives, or social networks. In all cases, being able to arrange such a vast amount of media is of critical importance both for indexing, categorization, and retrieval. This makes it possible for any user who attended, or is simply interested in the event, to recreate the event according to his personal experience, namely through summaries, stories, personalized albums [2][4].

However, it turns out that such a large amount of data is often unstructured and heterogeneous. The strong variability (and sometimes similarity) in terms of content and archiving strategies makes it difficult to manually organize all the event-related material in a simple yet effective manner. In this respect, it would be desirable to find a consistent way of presenting the media galleries captured during an event [7]. This task is not trivial, since timing and location information attached to the captured media (mostly timestamps and GPS) could be inaccurate or missing [3].

This lack of information is even more accentuated in case people use devices that do not have a direct connection to the Internet, thus requiring manual setting of the clock, especially after a battery discharge or replacement. In fact, in the case the temporal information is not represented correctly, there is a concrete risk of a misleading interpretation of the media collection, with high probability of losing part of the semantics of the event, due to a bad alignment along the temporal axis. Under such conditions, videos could be of great help, since they contain both audio and visual information that could be extremely relevant in providing additional details about the ongoing event compared to the sole presence of audio and still pictures.

The SEM task presented in 2014 was dealing only with still pictures and the results provided by the different teams are definitely encouraging. Participating teams competed tackling the problem in different ways. The authors in [5] proposed an approach based on the extraction of visual features (SIFT) to find the image pairs across the galleries that exhibit strong similarity. Then a non-homogeneous linear equation system is constructed to constrain the time offsets between the galleries based on these matching pairs to determine an approximate solution. Sansone et al. [6] relied their implementation on the use of a Markov Random Field to find the best correspondences between the images belonging to two different photo galleries. Zaharieva et al. [8] proposed two multimodal approaches that employ both visual and time information for the synchronization of different images galleries. The first approach relies on the pairwise comparison of images in order to link different galleries, while in the second approach Xmeans clustering is applied, and the time offsets are estimated by calculating the average time differences within the clusters. Apostolidis et al. [1] also proposed a method relying on the combination of different visual features, and using the images exhibiting the strongest similarity to compute the galleries offsets.

2. TASK DESCRIPTION

In our scenario we imagine a number of users attending the same event and taking photos and videos with different non-synchronized devices (smartphones, compact cameras, DSLRs, tablets). Each user contributes to the task with one gallery, which includes an arbitrary number of photos, audio files and videos. Assuming that users would like to merge their photo galleries in a single event-related collection, the best temporal alignment among the galleries should be found, so as to correctly report and preserve the tempo-

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany
.

ral evolution of the event. Furthermore, considering the high variability in terms of acquisition devices, we cannot expect the clocks of each device to be synchronized, neither in terms of precision, nor in terms of the time zone set by the users. In addition, in some cases, also the location data could be unavailable (not all devices have a GPS onboard), further reducing the chances of a correct event reconstruction. In fact, these factors may considerably hinder the quality of the alignment, thus different solutions should be envisaged, encompassing the joint analysis of temporal data, position information, and audio-visual similarity.

The SEM task expects teams to provide the estimated time offset between different galleries of pictures collected by different users and cameras. The goal can be summarised as follows: *given a set of media collections (galleries) taken by different users/devices at the same event, find the best (relative) time alignment among them at gallery level, and detect the significant sub-events over the whole event collection.*

3. DATASETS

For this challenge we make available four different datasets, exhibiting different challenges. The first dataset is related to the *Tour de France 2014*. It consists of images taken during the event and collected from Flickr. The dataset is split into 33 galleries. The dataset covers the entire competition. Some images are also provided with GPS information together with the timestamp. A second dataset concerns the famous exhibition held every year in California, namely *NAMM 2015*. The data-set consists of 420 images and 32 videos, split into 19 galleries. Each user gallery contains a variable number of media (ranging from 12 to 49). All images are downloaded from Flickr, while videos are downloaded from YouTube. The *Spring Party Salesiani 2015* is a dataset collected by the organizers, and recorded during a students' party held in Trento, Italy. It is composed of videos and pictures captured by the attendees during the event. Also in this case a gallery corresponds to the user's device, and media are complemented with the corresponding time-stamps. The last dataset, *Salford Test Shoot* includes 403 audio and 58 video files. Time-codes are available for most of the media. All datasets are provided with the corresponding ground truth, extracted by considering the acquisition time of the media and manually verified to check the consistency with respect to the captured event. The datasets related to the *Tour de France 2014*, *NAMM 2015*, and *Spring Party Salesiani 2015* include material subject to Creative Commons license and are freely available for download [1]. The Salford dataset is instead accessible via the ICoSOLE project website[2].

4. METRICS AND EVALUATION

Each submission will be evaluated in terms of: i) time synchronization error, and ii) sub-event detection error.

Concerning the first one, the goal of the participants is to maximize the number of galleries for which the synchronization error is below a predefined threshold ΔE_{max}, and to minimize the time shift of those galleries. The synchronization error for a gallery G_i with respect to the reference G_r is defined as $\Delta E_{ir} = \Delta T_{ir} - \Delta T^*_{ir}$, where ΔT_{ir} and ΔT^*_{ir} are the delay between G_i and G_r calculated on the participants' submission and ground truth, respectively. The threshold ΔE_{max} depends on the duration of the sub-events in the dataset, and represents the maximum accepted time lapse within which we consider a gallery as reasonably well-synchronized. We use the above quantities in order to estimate the synchronization precision (Eq. (1)) and accuracy (Eq. (2)):

$$Precision = \frac{M}{N-1} = \frac{Card\left(\Delta E_{ir} < \Delta E_{max}\right)}{N-1} \quad (1)$$

$$Accuracy = 1 - \frac{\sum_{i=1}^{N-1} \Delta E_{ir}}{(N-1)\Delta E_{max}} \quad (2)$$

Precision measures the number of galleries (M) over the total number of galleries ($N-1$, excluding the reference), that have been correctly synchronized. With the accuracy we instead evaluate the capabilities of the teams in minimizing the average time lapse calculated over the M synchronized galleries, normalized with respect to the maximum accepted time lapse.

The synchronization task provides a basis for the clustering task. Once the galleries are synchronized, it is possible to cluster the whole event collection to detect sub-events occurring within the entire event. Sub-events are defined in a neutral and unbiased way (e.g., making reference to the calendar/schedule of the event) and coded into the ground truth. We measure the performance of the sub-event clustering over the whole synchronized collection of media. For this, we use the Jaccard index JI and the clustering F1 score (Eq. (3)), where for computing the latter we use P and R, which represent the Precision and Recall, respectively.

$$JI = \frac{TP}{TP + FP + FN}, \qquad F1 = \frac{2PR}{P+R} \quad (3)$$

In the formulation above we declare a true positive (TP) when two images related to the same sub-event are put in the same cluster, and a true negative (TN) when two images belonging to different sub-events are assigned to two different clusters). False positives (FP) occur instead when two images are assigned to the same cluster although belonging to different sub-events.

5. CONCLUSIONS

In this paper we have presented the Synchronization of Multi-User Event Media task held at MediaEval 2015. The competing teams will be evaluated considering four datasets collected by the organizers, and made available online together with the corresponding ground truth. For the evaluation both the synchronization and the clustering performances will be evaluating, by measuring the galleries offset and computing the F1 score, respectively.

Acknowledgments

This work was supported in part by the EC under contract FP7-600826 ForgetIT. We would like to thank Alessio Xompero and Kostantinos Apostolidis for their precious help in collecting and annotating the images for the datasets used in the task.

[1] mmlab.disi.unitn.it/MediaEvalSEM2015

[2] https://icosole.lab.vrt.be/viewer/home

6. REFERENCES

[1] K. Apostolidis, C. Papagiannopoulou, and V. Mezaris. CERTH at MediaEval 2014 Synchronization of Multi-User Event Media Task. In *Proc. MediaEval 2014 Workshop*, CEUR vol. 1263, 2014.

[2] M. Broilo, G. Boato, and F. De Natale. Content-based Synchronization for Multiple Photos Galleries. In *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, pages 1945–1948, 2012.

[3] N. Conci, F. D. Natale, and V. Mezaris. Synchronization of multi-user event media (SEM) at MediaEval 2014: Task description, datasets, and evaluation. In *Proc. MediaEval 2014 Workshop*, CEUR vol. 1263, 2014.

[4] G. Kim and E. P. Xing. Jointly Aligning and Segmenting Multiple Web Photo Streams for the Inference of Collective Photo Storylines. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 620–627, 2013.

[5] P. Nowak, M. Thaler, H. Stiegler, and W. Bailer. JRS at Event Synchronization Task. In *Proc. MediaEval 2014 Workshop*, CEUR vol. 1263, 2014.

[6] E. Sansone, G. Boato, and M.-S. Dao. Synchronizing Multi-User Photo Galleries with MRF. In *Proc. MediaEval 2014 Workshop*, CEUR vol. 1263, 2014.

[7] J. Yang, J. Luo, J. Yu, and T. Huang. Photo Stream Alignment and Summarization for Collaborative Photo Collection and Sharing. *Multimedia, IEEE Transactions on*, 14(6):1642–1651, Dec 2012.

[8] M. Zaharieva, M. Riegler, and M. Del Fabro. Multimodal Synchronization of Image Galleries. In *Proc. MediaEval 2014 Workshop*, CEUR vol. 1263, 2014.