Synchronization of Multi-User Event Media (SEM) at MediaEval 2014: 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

ABSTRACT

In this paper we provide an overview of the Synchronization of Multi-User Event Media (SEM) Task that is part of the 2014 MediaEval Benchmark for Multimedia Evaluation. The SEM task is presented this year for the first time in MediaEval, and poses a new challenge, namely the temporal alignment of a series of photo galleries that relate to the same event but have been collected by different users. Besides aligning the pictures on a common timeline, participants are also required to detect the sub-events attended by the users and to group the pictures accordingly. The task is validated on two different datasets related to the 2010 and 2012 Winter / Summer Olympic games, each dataset comprising a variable number of pictures, galleries, and sub-events.

1. INTRODUCTION

Content creation is more and more a collective experience. People attending large social events (a soccer match, a concert), but also personal-scale ones (a wedding, a birthday party) collect dozens of photos and video clips with their smartphones, tablets, cameras, and more recently social cameras. Such information is later exchanged in a number of different ways, including shared repositories, clouds, social networks, etc. In this way, different media galleries are made available to each other, making it possible for any user who attended, or is simply interested to the event, to create his own view of it through summaries, stories, personalized albums [1][2]. However, such a large amount of data turns out to be unstructured and heterogeneous and, even if it would be possible to collect it on the same hard drive, the variability in terms of content, naming, archiving strategies makes it impossible to organize all the event-related material in a simple yet effective manner.

In this respect, a major issue is the need of aligning and presenting the various media galleries captured during an event in a consistent way [3]. As a matter of fact, the time and location information attached to the captured media (timestamp, GPS) can be wrong, inaccurate or even missing (for instance, due to wrong setting of the clock/calendar, different time-zone, modification or removal of tags). Similarly, this is also a common situation in historical events and photo archives, where timestamps and especially location information is rarely available. In some other cases, images might be processed offline for post-production, thus losing the correct temporal information. In such cases, creating a single timeline could turn out to be complicated and challenging, with a concrete risk of representing the event in a misleading way.

2. TASK DESCRIPTION

In our scenario we imagine a number of users (10+) attending the same event and taking photos and videos with different non-synchronized devices (smartphones, handheld cameras, DSLRs, tablets). Each user contributes to the task with one gallery, which includes an arbitrary number of photos, possibly covering just a part of the event, with variable density of acquisitions (single photos are also possible). Assuming that these 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 temporal evolution of the event. Furthermore, considering the high variability in terms of acquisition devices, we cannot expect the clocks of each device of the same user to be synchronized, neither in terms of precision, nor in terms of the time zone set by the users. Furthermore, in some cases, also the location data could be unavailable (not all devices have a GPS onboard), further reducing the available information about the captured event. In view of creating a single timeline, 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 visual similarity.

Therefore, 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 image 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 two different datasets, consisting of a collection of images gathered from Flickr and made available under Creative Commons license. Both datasets refer to well known and structured sport events, namely the Olympic Games held in London in 2012 and the Vancouver Winter Olympic Games of 2010. We have cho-

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain
.

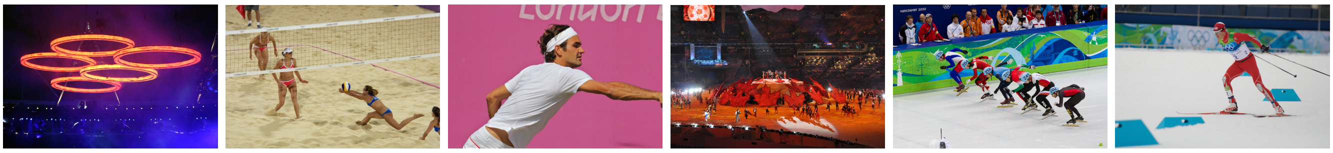

Figure 1: Sample images taken from the two datasets.

sen to work with these two events because on the one hand they exhibit a clear and organized schedule with precise timing. On the other hand they still exhibit a high variability in terms of visual content, due to the common features across different competitions in the same discipline, as well as strong similarities in the environments, in which the pictures are collected, making the synchronization a non-trivial task. As far as this task is concerned, the images within a gallery are consistent in terms of timestamp, and might include the GPS information. Therefore the temporal offsets are at gallery level thus assuming that every user uses one single device for acquisition.

The dataset collected from the London Olympics includes 2124 images, divided into 37 galleries. The first gallery comprises a subset of the data provided in the development set and is defined as the reference gallery. The dataset collected from the Vancouver Winter Olympic Games includes 1351 pictures representing most of the competitions, divided into 35 galleries with a variable number of pictures in each gallery. Also in this case, the first gallery is set as the reference. Fig. 1 shows a few samples of the two datasets.

4. METRICS AND EVALUATION

Two objective metrics will be used to evaluate the results:

- time synchronization error
- sub-event detection error

As far as the first metric is concerned, the goal of the participants is to maximize the number of galleries for which the synchronization error is below a predefined threshold, 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} is the delay between G_i and G_r calculated on the ground truth. 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.

As far the metrics for evaluation are concerned, we have considered for the temporal alignment the precision (Eq. 1) and accuracy (Eq. 2). For the quality of the clustering, we use the Rand Index (RI), as from Eq. 3, the Jaccard index (JI) Eq. 4 , and the F1 score Eq. 5, where P and R represent the Precision and Recall, respectively.

$$Precision = \frac{M}{N-1} = \frac{Card\,(\Delta E_{ir} < \Delta E_{max})}{N-1} \tag{1}$$

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

$$RI = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$JI = \frac{TP}{TP + FP + FN} \tag{4}$$

$$F1 = \frac{2PR}{P + R} \tag{5}$$

Precision measures the number of galleries (M) over the total number of galleries ($N - 1$, excluding the reference), that have been correctly synchronized, namely those galleries, for which the alignment error with respect to the reference gallery, is below a threshold. 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, for instance, the single competitions, or the ceremonies of the different disciplines. 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. In this case, we use the three performance indicators reported above, namely RI, JI, and F1. In the formulation we define a true positives (TP), in case two images related to the same sub-event are associated the same cluster, and the true negative (TN), when two images associated 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. ACKNOWLEDGMENTS

This work was supported in part by the EC under contract FP7-600826 ForgetIT. We would like to thank Anastasia Ioannidou, Alain Malacarne, and Alessio Xompero for their precious help in collecting the images for the dataset.

6. REFERENCES

[1] M. Broilo, G. Boato, and F. De Natale. Content-based synchronization for multiple photos galleries. In *Proceedings - International Conference on Image Processing, ICIP*, pages 1945–1948, 2012.

[2] G. Kim and E. P. Xing. Jointly aligning and segmenting multiple web photo streams for the inference of collective photo storylines. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 620–627, 2013.

[3] 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.