CERTH at MediaEval 2014 Synchronization of Multi-User Event Media Task

Konstantinos Apostolidis, Christina Papagiannopoulou, Vasileios Mezaris
Information Technologies Institute, CERTH, Thessaloniki, Greece
{kapost, cppapagi, bmezaris}@iti.gr

ABSTRACT

This paper describes the results of the CERTH participation in the *Synchronization of Multi-User Event Media* Task of MediaEval 2014. We used a near duplicate image detector to identify very similar photos, which allowed us to temporally align photo galleries; and then we used time, geolocation and visual information, including the results of visual concept detection, to cluster all photos into different events.

1. INTRODUCTION

People attending large-scale social events collect dozens of photos and video clips with their smartphones, tablets, cameras. These are later exchanged and shared in a number of different ways. The alignment and presentation of the photo galleries of different users in a consistent way, so as to preserve the temporal evolution of the event, is not straightforward, considering that the time information attached to some of the captured media may be wrong (due to different photo capturing devices not being synchronized) and geolocation information may be missing. The 2014 MediaEval Synchronization of Multi-user Event Media (SEM) task tackles this exact problem [1].

2. SYSTEM OVERVIEW

The main goal of our system is the time alignment of photo galleries that are created by different digital photo capture devices, and the clustering of these into event-related clusters. In the first stage, similar photos of the different galleries are identified and are used for constructing a graph, whose nodes represent galleries and edges represent discovered links between them. Time alignment of the galleries is achieved by traversing the graph. After that, we apply clustering techniques in order to split our collection into different events. Figure 1 shows the pipeline of our system.

3. TIME SYNCHRONIZATION

Time synchronization makes use of a Near Duplicate Detector (NDD) that extracts SIFT descriptors from the photos, forms a visual vocabulary and encodes the descriptor-based representation of each photo using VLAD encoding. The nearest neighbours that are returned for a query image are refined by checking the geometrical consistency of SIFT keypoints using geometric coding (GC) [4].

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

We further modified this NDD process to also use color information (HSV histograms), so that near duplicate candidates that are very similar in color are not discarded even if the GC score is relatively low.

We apply the modified NDD on the union of all galleries. Consequently, we filter out identified pairs of near duplicates according to the following rules:

- Reject pairs when geolocation information is available and the location distance of the two photos is greater than a distance threshold.
- Reject pairs when the time difference between the photos is above an extreme time threshold (which indicates that this time difference is unlikely to be due to a time synchronization error alone).

The remaining near duplicate photos belonging to different galleries are considered as links between those galleries.

It is now straightforward to construct a graph whose nodes represent the galleries, and the edges represent these links between galleries. Each edge has a weight which is equal to the number of links between the two galleries. Having constructed the graph, we compute the time offset of each gallery by traversing it, as follows. Starting from the node corresponding to the reference gallery, we select the edge with the highest weight. We compute the time offset of the node on the other end of this edge as the median of the time differences of the pairs of near duplicate photos that this edge represents, and add this node to the set of visited nodes. The selection of the edge with the highest weight is repeated, considering as possible starting point any member of the set of visited nodes, and the corresponding time offset is computed, until all nodes are visited. Alternatively, we can traverse the graph and compute the nodes' time offsets by simultaneously considering the weights of multiple edges.

4. MEDIA CLUSTERING OF EVENTS

Following time synchronization, we cluster all photos to events. Two different approaches are adopted: the first one considers all photo galleries as a single photo collection, exploiting the synchronization results, while the second one first makes a pre-clustering within each gallery separately.

In the first approach, we use the method of [2], resulting in clusters that are time distinct, comprising different events. Subsequently, each of these clusters is split based on the geolocation information. The photos that do not have geolocation information are assigned to the geo-cluster which is more similar according to the color information (e.g. HSV histogram).

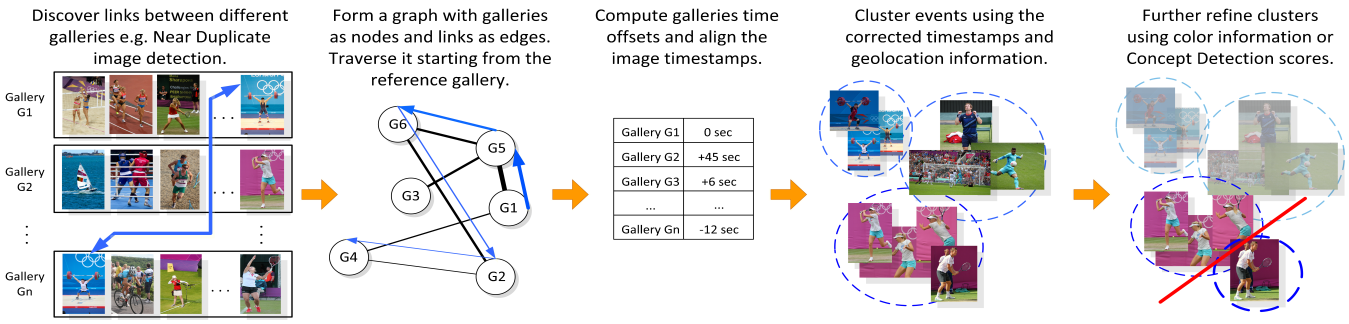

Figure 1: System overview

In the second approach, we detect time gaps between events of each gallery. Specifically, we find the minimum time difference of dissimilar photos which is greater than the maximum time difference of the near-duplicate photos (based on the similarity matrix of GC). The clusters that are formed are merged according to time and geolocation similarity. For the clusters that do not have geolocation information, the merging is continued by considering the time and low-level feature similarity or the time and the concept detector (CD) confidence similarity scores [3].

5. EXPERIMENTS AND RESULTS

We submitted 5 runs in total, combining 3 methods for time synchronization and 3 methods for event clustering:

- *Run1:aNDD-perGallery-mergeCD*: Compute gallery time offsets using our modified NDD. CD scores are used to merge clusters using the second approach of section 4.
- *Run2:aNDD-perGallery-mergeHSV*: Compute gallery time offsets using our modified NDD. HSV histogram similarity is used to merge clusters using the second approach of section 4.
- *Run3:aNDD-concat*: Compute gallery time offsets using our modified NDD. Clustering is performed using the first approach of section 4.
- *Run4:aNDD-multiT-perGallery-mergeCD*: Compute gallery time offsets using our modified NDD and traversal of the graph by simultaneously considering the weights of multiple edges. CD scores are used to merge clusters using the second approach of section 4.
- *Run5:NDD-perGallery-mergeCD*: Compute gallery time offsets using NDD without HSV color information. CD scores are used to merge certain events using the second approach of the section 4.

The results of our approach for all 5 runs, for the Vancouver testset and the London testset are listed in Tables 1 and 2 respectively.

Table 1: Time Synchronization and Clustering metrics for each run for the Vancouver testset.

	run1	run2	run3	run4	run5
Sync. Precision	**0.9118**	**0.9118**	**0.9118**	0.5294	**0.9118**
Sync. Accuracy	**0.7375**	**0.7375**	**0.7375**	0.7014	0.7279
Rand Index	**0.9770**	0.9734	0.9526	0.9601	0.9656
Jaccard Index	0.2581	0.2315	0.2856	0.1782	**0.2861**
F-Measure	0.2052	0.1880	0.2222	0.1512	**0.2225**

Table 2: Time Synchronization and Clustering metrics for each run for the London testset.

	run1	run2	run3	run4	run5
Sync. Precision	0.6111	0.6111	0.6111	0.2222	**0.6389**
Sync. Accuracy	0.7127	0.7127	0.7127	0.6996	**0.7299**
Rand Index	0.9885	**0.9910**	0.9838	0.9829	0.9863
Jaccard Index	0.5051	**0.5614**	0.3232	0.2739	0.4849
F-Measure	0.3356	**0.3596**	0.2443	0.2150	0.3266

6. CONCLUSIONS

This paper presented our framework and results at the MediaEval 2014 Synchronization of Multi-User Event Media Task. Our modified NDD approach gives the best results in time alignment for the Vancouver testset, while the standard NDD yields a slightly better time synchronization for the London testset. In sub-event clustering, the exploitation of consistent timestamps in a gallery and the use of CD confidence scores gives a good performance for the Vancouver testset, whereas HSV histogram similarity seems to give the best clustering results for the London testset.

7. ACKNOWLEDGMENTS

This work was supported by the EC under contracts FP7-287911 LinkedTV and FP7-600826 ForgetIT.

8. REFERENCES

[1] N. Conci, F. De Natale, and V. Mezaris. Synchronization of Multi-User Event Media (SEM) at MediaEval 2014: Task Description, Datasets, and Evaluation. In *Proc. MediaEval Workshop*, 2014.

[2] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal event clustering for digital photo collections. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 1(3):269–288, 2005.

[3] C. Papagiannopoulou and V. Mezaris. Concept-based Image Clustering and Summarization of Event-related Image Collections. In *Proc. Int. Workshop on Human Centered Event Understanding from Multimedia (HuEvent14) of ACM Multimedia (MM14)*, 2014.

[4] W. Zhou, H. Li, Y. Lu, and Q. Tian. SIFT match verification by geometric coding for large-scale partial-duplicate web image search. *ACM Trans. Multimedia Comput. Commun. Appl.*, 9(1):4:1–4:18, Feb. 2013.