CERTH at MediaEval 2015 Synchronization of Multi-User Event Media Task

Konstantinos Apostolidis
CERTH-ITI
Thermi 57001, Greece
kapost@iti.gr

Vasileios Mezaris
CERTH-ITI
Thermi 57001, Greece
bmezaris@iti.gr

ABSTRACT

This paper describes the results of our participation to the Synchronization of Multi-User Event Media Task at the MediaEval 2015 challenge. Using multiple similarity measures, we identify pairs of similar media from different galleries. We use a graph-based approach to temporally synchronize user galleries; subsequently we use time information, geolocation information and visual concept detection results to cluster all photos into different sub-events. Our method achieves good accuracy on considerably diverse datasets.

1. INTRODUCTION

People attending large 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 media 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 and geolocation information may be missing. The 2015 MediaEval Synchronization of Multi-user Event Media (SEM) task tackles this exact problem [2].

2. METHOD OVERVIEW

The proposed method temporally aligns user galleries that are created by different digital capture devices, and clusters the time-aligned photos into event-related clusters. In the first stage, we assess media similarity by combining multiple similarity measures and by taking into account the geolocation metadata of photos. Similar media of the different galleries are identified and are used for constructing a graph, whose nodes represent galleries and edges represent the discovered similarities between media items of different galleries. Synchronization of the galleries is achieved by traversing the minimum spanning tree (MST) of the graph. Finally, we apply clustering techniques to split the media to different sub-events. Figure 1 illustrates the proposed method.

3. MEDIA SIMILARITY ASSESSMENT

To identify similar photos of different galleries, we combine the information of four similarity measures [1]:

1. Geometric Consistency of Local Features Similarity (GC): We check the geometric consistency of SIFT keypoints for each pair of photos, using geometric coding [6]. The GC similarity can discover near-duplicate photos.
2. Scene Similarity (S): We calculate the pairwise cosine distances between the extracted GIST descriptor [4] of each photo. High S similarity indicates photos captured at similar scenery (indoor, urban, nature).
3. Color Allocation Similarity (CA): We divide each image to three equal, non-overlapping horizontal strips, and extract the HSV histogram of each. We calculate the pairwise cosine distances between the concatenation of the HSV histograms. High CA similarity indicates photos with similar colors.
4. DCNN Concept Scores Similarity (DCS): We use the Cafe DCNN [3] and the googleNet pre-trained model [5] to extract concept scores for photos. We use the Euclidean distance to calculate pairwise distances between concept scores vectors of photos. High DCS similarity indicates semantically similar photos.

We calculate the aforementioned similarity measures on the photos of all galleries to be synchronized. We combine the information of all similarity measures, using the following procedure: initially, the similarity $O(i,j)$ of photos i and j is set equal to $GC(i,j)$. Then, if $S(i,j) > t_s$ and $S(i,j) > GC(i,j)$, $O(i,j)$ is updated as $O(i,j) = S(i,j)$. The same update process is subsequently repeated using CA similarity and DCS similarity (and the respective t_c, t_d thresholds).

Subsequently, we weigh each similarity value so that the similarity of photos with distance of capture locations lower than a m threshold is emphasized, while the similarity of photos with distance of capture locations significantly above this threshold is zeroed. Similar photos that belong to different user galleries are treated as potential links between these galleries.

To identify similar audio files of different galleries, we perform cross correlation of audio data, degraded to 11KHz sampling rate. For video files, we select a frame for each second of video and resize it to 1 pixel width. To identify similar video files of different galleries, we perform cross correlation of the horizontally concatenated resized frames.

The t_s, t_c and t_d thresholds are empirically calculated over the training dataset. The m threshold is calculated by estimating a Gaussian mixture model of two Gaussian distributions on the histogram of all photo's pairwise capture location distances. The Gaussian distribution with the low-

MediaEval 2015 Workshop, September 14-15, 2015, Wurzen, Germany

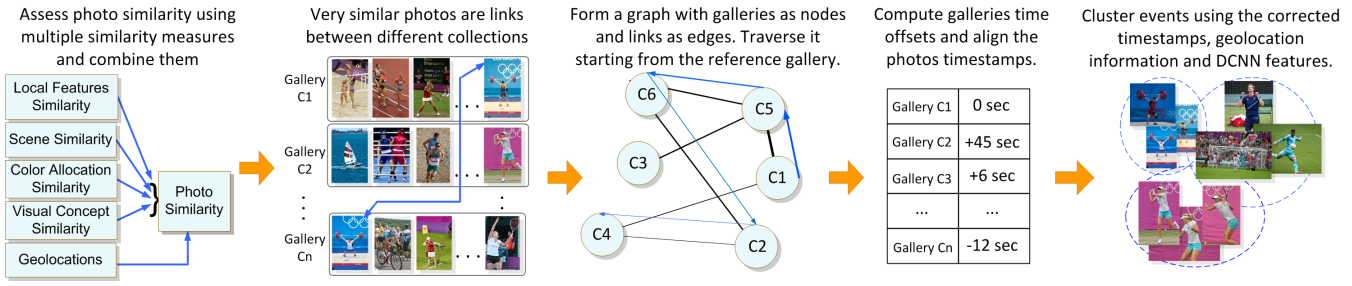

Figure 1: System overview

est mean (m) presumably signifies photos captured in the same sub-event.

4. TEMPORAL SYNCHRONIZATION

Having identified potential links for at least some gallery pairs, we construct a weighted graph, whose nodes represent the galleries, and its edges represent the links between galleries. The weight assigned to each edge is calculated as the sum of similarities of the photos linking the two galleries. Using this graph, the temporal offsets of each gallery will be computed against the reference gallery.

We compute the temporal offset of each gallery by traversing the minimum spanning tree (MST) of the galleries graph. This procedure ($MSTt$) can be summarized as follows: Starting from the node corresponding to the reference gallery, we select the edge with the highest weight. We compute the temporal offset of the node on the other end of this edge as the median of the capture time differences of the pairs of similar photos that this edge represents. We add this node to the set of visited nodes. The selection of the edge with the highest weight is repeated, considering any member of the set of visited nodes as possible starting point, and the corresponding temporal offset is again computed, until all nodes are visited. This process is explained in more detail in [1].

The MSTt method calculates the offsets using only the shortest path from a visited node to any given node. We also explore a variation of the MStt process as an alternative way of computing temporal offsets ($MSTx$): before traversing the MST of the graph, we detect fully-connected triplets of nodes and we average the offset of the shortest path with the alternative path in each triplet, only if the difference of the two paths is lower than $maxDiff$ threshold. Utilizing in this $MSTx$ process some additional information that the $MSTt$ method ignores, we expect to achieve better accuracy in time synchronization.

5. SUB-EVENT CLUSTERING

After time synchronization, we cluster all photos to sub-events. Two different approaches were adopted. In the first approach (MPC), we apply the following procedure: At the first stage, we split the photo's timeline where consecutive photos have temporal distance above the mean of all temporal distances. At the second stage, geolocation information is used to further split clusters of photos.During the third stage, clusters are merged using time and geolocation information. For the clusters that do not have geolocation information, the merging is continued by considering visual similarity. In the second approach (APC), we augment the DCNN feature vectors with the normalized time information and cluster the media using Affinity Propagation.

6. RESULTS

We submitted 4 runs in total, combining the 2 methods for temporal synchronization and the 2 methods for sub-event clustering. The results of our approach for all datasets and all four runs are listed in Table 1. From the reported results, it is clear that our method achieved good accuracy but only managed to synchronize a small number of galleries, particularly in the *TDF14* dataset. In sub-event clustering, the MPC method scored a slightly better F-score (column F1) for two of the datasets. The $MSTt$ and $MSTx$ methods performed the same because $maxDiff$ was set too low ($maxDiff = 10$), which allowed only very small adjustments, thus degenerating the $MSTx$ method to $MSTt$.

Table 1: Proposed method results.

Dataset	Run	Precision	Accuracy	F1
NAMM15	$MSTt$+APC	0.833	0.908	0.226
	$MSTt$+MPC	0.833	0.908	0.348
	$MSTx$+APC	0.833	0.908	0.226
	$MSTx$+MPC	0.833	0.908	0.348
TDF14	$MSTt$+APC	0.125	0.845	0.113
	$MSTt$+MPC	0.125	0.845	0.001
	$MSTx$+APC	0.125	0.845	0.113
	$MSTx$+MPC	0.125	0.845	0.001
STS	$MSTt$+APC	0.424	1.000	0.123
	$MSTt$+MPC	0.424	1.000	0.164
	$MSTx$+APC	0.424	1.000	0.123
	$MSTx$+MPC	0.424	1.000	0.164

7. CONCLUSIONS

In this paper our framework and results at the MediaEval 2015 Synchronization of Multi-User Event Media Task are presented. Better fine-tuning of the algorithm parameters is required to achieve consistently good performance on diverse datasets. As a future work, we are considering extending the algorithm to automatic parameter selection (which could lead to select more links between galleries, thus improving precision), experiment with different values of $maxDiff$, and apply a more sophisticated method to combine different similarity measures.

8. ACKNOWLEDGMENTS

This work was supported by the European Commission under contract FP7-600826 ForgetIT.

9. REFERENCES

[1] K. Apostolidis and V. Mezaris. Using photo similarity and weighted graphs for the temporal synchronization of event-centered multi-user photo collections. In *Proc. 2nd Workshop on Human Centered Event Understanding from Multimedia (HuEvent'15) at ACM Multimedia (MM'15)*, Brisbane, Australia, Oct. 2015.

[2] N. Conci, F. De Natale, V. Mezaris, and M. Matton. Synchronization of Multi-User Event Media (SEM) at MediaEval 2015: Task Description, Datasets, and Evaluation. In *Proc. MediaEval Workshop*, Wurzen, Germany, Sept. 2015.

[3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *ACM Int. Conf. on Multimedia*, Nov. 2014.

[4] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[6] 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.