Proc. 2nd ACM Workshop on Human Centered Event Understanding from Multimedia (HuEvent'15) at ACM Multimedia (MM'15), Brisbane, Australia, Oct. 2015.

Using Photo Similarity and Weighted Graphs for the Temporal Synchronization of Event-Centered Multi-User Photo Collections

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

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

ABSTRACT

This paper describes a method to temporally align photo collections that have been created during the same event by different users using their own unsynchronized digital photo capture devices. Using multiple similarity measures, we identify pairs of similar photos from different collections. We then temporally align the photo collections by traversing a graph, whose nodes represent the collections, and edges represent the similar photo pairs between collections. Outcome of this process is a set of modified timestamps for the photos, which could be used in applications such as time-based clustering and sub-event detection in multi-user photo collections. We evaluate the proposed synchronization method on benchmark datasets and we compare it to state-of-the-art methods, demonstrating its superiority.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

Keywords

Temporal synchronization; photo collection; weighted graph

1. INTRODUCTION

People attending large-scale social events collect dozens of photos and video clips with their smartphones, tablets and cameras, to be later exchanged and shared in a number of different ways. The metadata that is automatically attached to these media items at capture time, such as timestamps and geo-coordinates, constitute valuable information for the consumption of the visual content in social networking applications, e.g. for generating multi-user timelines and summaries of the captured events, and other similar applications related to social sharing and digital memory preservation. However, the timestamps in particular, despite being among the most valuable information, are not necessarily directly comparable between media items captured by different devices of different users. Time offsets are introduced by user neglect and time-measurement differences across the world (e.g. does everyone bother to introduce summer/winter time changes twice a year, or time-zone changes during every trip, to his digital camera?), or also differences in the way that different classes of photo capture devices handle time (e.g. some mobile phones, depending on their settings, automatically get up-to-date time information from the network provider; other phones are set to not use such network information, and traditional digital cameras typically do not offer such an option). Being able to remove this noise, i.e. accurately aligning photo collections of different users for the same event in chronological order, is important for making time information directly usable in applications.

The alignment and presentation of the photo collections of different users in a consistent way, so as to preserve the temporal evolution of the event, is not straightforward. Besides the capture time information attached to some of the media being possibly wrong (due to the different photo capturing devices not being synchronized, as explained), geolocation information may also be missing, since not all photo capture devices can or are set to record GPS metadata. Furthermore, photos belonging to the same or other events, despite being semantically similar, may be visually dissimilar; and vice versa. For example, different and temporally distant sub-events may take place in the same setting, resulting in very similar photos being captured.

In this paper we propose a synchronization method for multi-user photo collections that are captured during a single, often large, event and we introduce two novelties: 1) the combination of multiple features and similarity measures to identify very similar photos, 2) the employment of a weighted graph-based method that, based on the identified pairs of similar photos, is able to synchronize multiple photo collections even when each individual collection exhibits low coverage of the overall event.

The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 describes our proposed method for the temporal synchronization of photo collections. In Section 4 we experimentally evaluate different design choices of our method and compare it to the most recent related works. Finally, we draw conclusions in Section 5.

2. RELATED WORK

The problem of temporally synchronizing multi-user photo collections is typically addressed at two stages: At the first

HuEvent'15, October 30, 2015, Brisbane, Australia.
© 2015 ACM. ISBN 978-1-4503-3748-9/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2815244.2815246.

stage features are extracted from the photos in order to discover similar photos across different collections. At the second stage, assuming that a sufficient number of similar photo pairs has been identified, an estimation of the temporal offsets between different collections is generated.

The first works identifying and attempting to address the problem of synchronizing photos that carry potentially erroneous timestamps are [4] and [20]. In [4], the authors present a content-based synchronization algorithm that extracts features from the photos using a Color and Edge Directivity descriptor [5] and the SURF descriptor [3]. The difference in the capture times of similar photos (belonging to different collections) is treated as a possible temporal offset between the collections. The temporal offset estimation among multiple collections is calculated by selecting the most frequent value from a histogram of temporal offsets for each collection, and averaging the offsets at a window of one minute around the selected frequent value. In [20], the authors extract color histograms, GIST [13] and Locality-constrained Linear Coding descriptors [19] to describe photos and find similar photo pairs between collections. They employ a sparse bipartite graph to find the informative photo pairs and a max linkage selection competing procedure to prune the false pairs. However, the bipartite graph construction they propose involves solving a series of optimization problems, whose number is proportional to the number of images in every possible photo collection pair, making the overall method computationally expensive. Photo datasets from the Picasa Web album are used for evaluation; nevertheless, only the percentage of aligned galleries is assessed, without considering the accuracy of the synchronization.

In [10], a method is proposed to construct a collective storyline of media found in the social web. The authors focus on segmenting each image in foreground objects and background, so as to assess photo similarity by detecting any instances of the same objects, possibly appearing in different areas of photos with different poses. A user-defined parameter K denotes the number of foreground areas to split each photo to, using a Multiple Foreground Cosegmentation algorithm [9]. They extract color histograms and SIFT local descriptors [11] using dense sampling on the foreground areas. Building a nearest neighbour similarity graph that connects the photo collections to be aligned, they formulate the alignment of the photo collections as an energy minimization problem. Belief propagation on the graph is utilized to solve the problem and achieve temporal alignment. They compare their method to three baseline methods, on a Flickr outdoor recreational activity dataset, demonstrating good alignment. Nevertheless, real-world photo collections are not always centred on objects that can be easily segmented, e.g. photos of a music festival may contain several faces from the audience as well as the musicians faces along with a complex background. This, together with the known imperfection of any segmentation algorithm and the absence of a single segmentation algorithm that can effectively be applied to any kind of photos without parameter tuning (e.g. K) make the method of [10] difficult to automatically apply to any possible photo synchronization scenario.

Several recent works that study the synchronization problem on generic multi-user photo collections have been proposed and evaluated in relation to the MediaEval 2014 SEM benchmarking activity [6]. For example, various different features were used in [21, 1, 12, 16] for photo similarity assessment. Specifically, in [21], the MPEG-7 Color Structure Descriptor and a Joint Composite Descriptor are used. In [1, 12] the SIFT local descriptor is used, while in [16] the SURF local descriptor and HSV color histograms are used. Furthermore, different methods are used for estimating the temporal offsets between collections on the basis of photo similarity assessment results. An Agglomerative Hierarchical Clustering method is proposed in [21] that uses the lowest-level clusters as links between different collections and the highest-level clusters as sub-events. In [1], the authors select the most similar photos between different collections and employ a graph of photo similarities to find paths between each collection and the reference one. In [12], temporal offsets are expressed as a non-homogeneous linear equation system and an approximate solution is calculated. In [16], the authors build a probabilistic graphical model in which each temporal displacement is identified by a set of nearest-neighbour photo pairs across photo collections. They estimate the temporal offsets among photo collections through exact inference.

Looking a bit beyond still images, there are methods in the literature that deal with the temporal synchronization of audio information. In [8], two graph-based approaches for synchronizing multiple audio signals are presented. The graphs are constructed atop the over-determined system resulting from pairwise signal comparison using cross-correlation of audio features. In [7], an approach for the temporal alignment and management of shared audiovisual streams is presented that is based on audio-visual bimodal segmentation. Although such methods cannot deal with photo collection synchronization, they show that temporal synchronization of multi-user media is useful in practice.

Finally, it is worth noting that temporal synchronization of multi-user photo collections and their presentation in a single timeline is an important step prior to any kind of multi-user collection summarization [20, 17] and event clustering [18, 14, 15].

3. PROPOSED METHOD

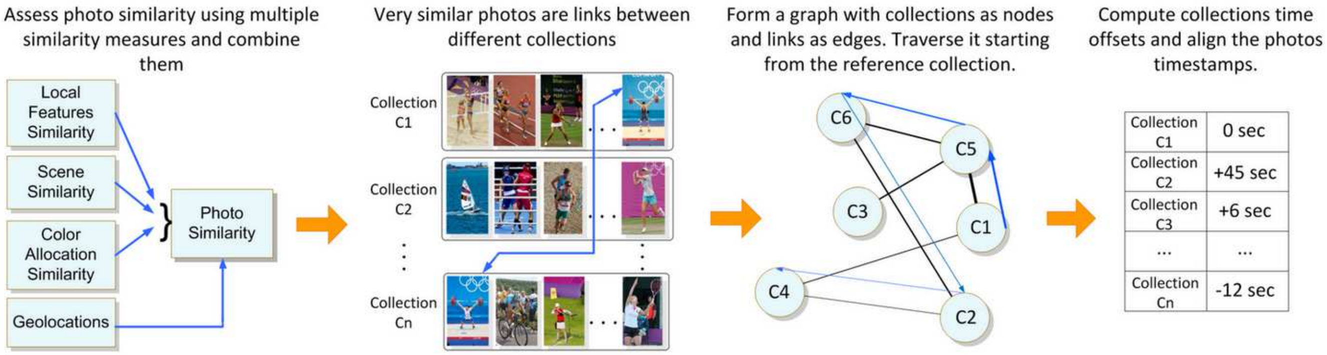
3.1 Method Overview

Figure 1 shows the overview of the proposed method for photo collection temporal synchronization. Initially, similar photos between different collections are identified. We assess photo similarity combining multiple visual similarity measures and also taking into account the geolocation metadata of photos, when available, as discussed in detail in Section 3.2. Then, very similar photos from different collections are considered as links between the photo collections. Subsequently, we construct a graph, whose nodes represent collections and whose edges represent the discovered links between them. Finally, temporal synchronization of the collections is achieved by appropriately traversing the collections' graph, as explained in Section 3.3.

3.2 Photo Similarity Assessment

To identify similar photos of different collections, we combine the information of four similarity measures.

1) Geometric Consistency of Local Features Similarity (GC): We encode the SIFT descriptors [11] extracted from each photo, using VLAD encoding [2]. The nearest neighbours of each photo's VLAD representation are refined by checking the geometrical consistency of SIFT keypoints


Assess photo similarity using multiple similarity measures and combine them
Local Features Similarity
Scene Similarity
Color Allocation Similarity
Geolocations
Photo Similarity
Very similar photos are links between different collections
Collection C1
Collection C2
Collection Cn
Form a graph with collections as nodes and links as edges. Traverse it starting from the reference collection.
C6
C5
C3
C1
C4
C2
Compute collections time offsets and align the photos timestamps.

Collection C1	0 sec
Collection C2	+45 sec
Collection C3	+6 sec
...	...
Collection Cn	-12 sec

Figure 1: Proposed method overview.

for each pair of photos using geometric coding [22]. Output of this process is the W_{GC} matrix, which holds the GC similarity score of each possible pair of n given photos.

2) Scene Similarity (S): We extract the GIST descriptor [13] from each photo. A D_S matrix holds the pairwise Cosine distance of the photos' GIST descriptors. We convert the distance matrix D_S to the similarity matrix W_S as $W_S(i,j) = \exp(-D_S(i,j)), \forall i,j \in [1\ldots n]$.

3) Color Allocation Similarity (CA): We split each photo to three equal-height strips, extracting an 48-bin HSV histogram from each strip. We compute the Cosine distance between the concatenated HSV histogram vectors for each pair of photos, constructing a distance matrix D_{CA}. We convert the distance matrix D_{CA} to the similarity matrix W_{CA} as $W_{CA}(i,j) = \exp(-D_{CA}(i,j)), \forall i,j \in [1\ldots n]$.

4) Physical Location Similarity (PL): We construct a W_{PL} matrix that holds the distances of capture locations for all pairs of photos with geolocation information available. For a pair of photos where geolocation information is not available, we set the respective value in W_{PL} to a negative value so that PL similarity will not be taken into consideration for the specific pair of photos.

We calculate the aforementioned similarity measures on the union of photos of all collections to be synchronized. After calculating the W_{GC}, W_S, W_{CA}, and W_{PL} matrices we combine them into a single similarity matrix W', using the following procedure: for each (i,j) pair of photos, $W(i,j)$ is initially assigned the value of $W_{GC}(i,j)$. If scene similarity ($W_S(i,j)$ value) is significant (above a t_S threshold) and greater than the current value of $W(i,j)$, then $W(i,j)$ is updated with the $W_S(i,j)$ value. The same is subsequently repeated using color allocation similarity ($W_{CA}(i,j)$ and a t_c threshold). The thresholds used in our experiments are discussed in Section 4.

In order to subsequently combine photo capture location distances and visual similarity we do the following: we construct the histogram of all photo's pairwise capture location distances. We estimate a Gaussian mixture model of two Gaussian distributions on this histogram. The Gaussian distribution with the lowest mean ($m1$) presumably signifies photos captured in the same sub-event while the Gaussian distribution with the highest mean ($m2$) presumably corresponds to photos captured in different sub-events. Based on this hypothesis, we weigh each visual similarity value ($W(i,j) \forall i,j \in [1\ldots n]$), so that the total similarity of photos with distance of capture locations lower than $m1$ will be emphasized, while the total similarity of photos with distance of capture locations significantly above $m1$ will be zeroed. To weight the matrix of the visual similarity of photos, W, with the distance of photos capture locations we compute: $W'(i,j) = W(i,j) \cdot a \cdot \exp\left(-\frac{W_{PL}(i,j)^2}{2m1^2}\right), \forall i,j \in [1\ldots n]$ for $a = 1.1$. The a parameter is a factor by which the similarity of photos that exhibit geolocation proximity is increased.

The whole procedure of photo similarity assessment is described in Algorithm 1. The output of this algorithm, the W' matrix, holds the similarity for all pairs of photos, combining the information of all the aforementioned similarity measures. Each value of the W' matrix for photos i and j where i and j do not belong to the same user collection and $W'(i,j) > t$, is treated as a potential link between photo collections. The t parameter is calculated for each pair of collections, by finding the maximum value in the interval $[0.8,\ldots,0.7]$ with step $t_{step} = -0.01$, that allows the discovery of at least three potential links. It should be stressed here that the above procedure does not guarantee that three (or even one) potential links will be found for every collection pair.

Algorithm 1 Similarity Matrices Combination

```
Input: $W_{GC}$, $W_S$, $W_{CA}$, $W_{PL}$, $n$, $t_S$, $t_c$, $a$, $m1$
Output: $W'$
  for $i \leftarrow 1$ to $n$ do
      for $j \leftarrow i+1$ to $n$ do
          $W(i,j) \leftarrow W_{GC}(i,j)$
          if $W_S(i,j) > t_S$ then
              if $W_S(i,j) > W(i,j)$ then
                  $W(i,j) \leftarrow W_S(i,j)$
              end if
          end if
          if $W_{CA}(i,j) > t_c$ then
              if $W_{CA}(i,j) > W(i,j)$ then
                  $W(i,j) \leftarrow W_{CA}(i,j)$
              end if
          end if
          $W'(i,j) \leftarrow W(i,j)$
          if $W_{PL}(i,j) \geq 0$ then
              $W'(i,j) \leftarrow W'(i,j) \cdot a \cdot \exp\left(-\frac{W_{PL}(i,j)^2}{2m1^2}\right)$
          end if
      end for
  end for
```

3.3 Graph-based temporal offset estimation

Having identified potential links for at least some collection pairs, it is relatively straightforward to construct a weighted graph, whose nodes represent the collections, and its edges represent the links between collections. The weight assigned to each edge is calculated as the sum of similarities of the photos linking the two collections. Using this graph, the temporal offsets of each collection will be computed against a user collection which is considered as the reference. Any collection can be used as reference, since we are aiming for relative synchronization; we neither assume nor need one collection to accurately match the true time.

Figure 2 shows the graph constructed for one of the test datasets used in this work. The first collection of this dataset is arbitrarily chosen as the reference collection in this example, and the corresponding node C1 is shown as a yellow circle. Nodes corresponding to collections that have direct links to the reference collection, identified using the procedure of Section 3.2, are depicted as blue rectangles. All other nodes of the graph are depicted as white rectangles. As can be seen from this example, attempting to directly compute the pairwise offsets between the reference collection and each other collection would not work, since only a portion of the collections is directly connected with the reference collection (and this would be the case regardless of which collection was chosen as the reference one, since the graph of Fig. 2 is a sparse graph). Instead, we must find a way to properly traverse the graph in order to estimate an offset for each collection.

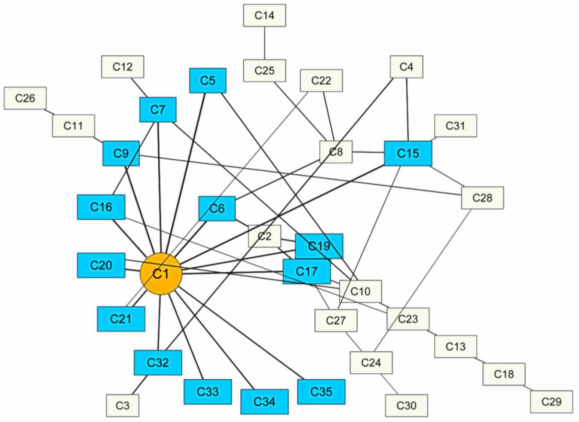


Figure 2: Graph representation of a test dataset.

Given a connected, undirected graph, a spanning tree of that graph is a subgraph that is a tree and connects all the vertices. A single graph can have many different spanning trees. Assuming that the edge weights represent how unfavorable each edge is, a minimum spanning tree (MST) is defined as a spanning tree with weight less than or equal to the weight of every other spanning tree. An MST ensures the traversal of all nodes with minimum effort. In our case, the edge weights represent how favorable each edge is (the higher the better); thus the MST leads to the traversal of the graph along the edges with the highest weights.

We can compute the temporal offset of each collection by traversing the MST of the collections graph as follows: Starting from the node corresponding to the reference collection, 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 use the median of capture time differences as a measure which is less sensitive to extreme offset values of erroneous photo pairs. 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. We denote this graph traversal method as MST-Med in the sequel.

The MST contains the minimum effort path from the reference collection to any other collection. As an alternative way of traversing the graph, we can average the offsets from all possible paths from the reference collection to any collection. We can follow the aforementioned procedure, additionally checking if there are connected nodes in the set of visited nodes every time a node is considered and average the offset computed from the MST path and all other discovered paths. Having computed the offset of all collections, we are able to estimate the aligned timestamps for all photos and finally sort the photos of the entire collection by capture time. We denote this traversal method as MST-Av.

Figure 3 shows the two above graph traversal methods on a toy graph. Assuming the C1 node is the node corresponding to the reference collection, the MST-Med method traverses the graph following the path $C1 \rightarrow C3$, $C1 \rightarrow C2$, $C3 \rightarrow C4$ while the MST-Av method traverses the graph following the path $C1 \rightarrow C3$, $\{C1, C3\} \rightarrow C2$, $C3 \rightarrow C4$. Thus, using the MST-Av method, the C2 collection offset is computed as the average of the $C1 \rightarrow C2$ and $C3 \rightarrow C2$ offsets.

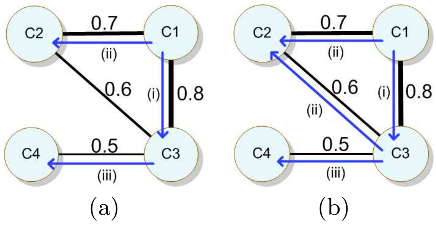


Figure 3: Toy example of graph traversal methods (a) MST-Med and (b) MST-Av.

4. EXPERIMENTS AND RESULTS

4.1 Datasets and Evaluation Framework

For experimentally evaluating the proposed methods we use the MediaEval 2014 SEM task datasets [6], consisting of photos from various users taken during two Olympic Games events. All photos are organized in collections (each collection captured using a single device) and come with timestamps, which are consistent within each collection but may have considerable temporal offsets across different collections. Furthermore, some collections include geolocation information, while others do not. The photos are organized in one training dataset and two test datasets. We strictly followed the experimental setup and evaluation procedure of the MediaEval 2014 SEM task, making our results directly comparable to the results of the task participants.

The training dataset includes a subset of photos downloaded from Flickr, depicting different phases of the London Olympic Games of 2012. It totals 304 photos arranged into 10 collections, each collection containing a variable number of photos. The first test dataset, the *Vancouver* dataset is about the Vancouver Winter Olympics Games, consisting of 1351 photos arranged in 35 collections, while the second, the *London* dataset, is an extension of the training dataset, with a richer variety of competitions, consisting of 2124 photos arranged in 37 collections. As part of the adopted experimental setup, in all datasets the first collection is considered as the reference one. Based on preliminary experiments on the training dataset we empirically set the parameters of the proposed method; specifically, we set the scene similarity significance threshold $t_s = 0.90$, and the color allocation similarity significance threshold $t_c = 0.85$.

The evaluation measures we used, as defined in [6], are: **1) Precision** (P) is the ratio between the number of synchronised collections (M), and the total number of collections ($N-1$, excluding the reference collection) in a testset. A collection is considered to be synchronized if the difference between the estimated timestamps of each of its photos and the corresponding ground truth is lower than a maximum accepted temporal offset ($maxError$). The value of $maxError$ is estimated as in [6], equalling to 1800 seconds. The Precision measure is defined as: $P = M/(N-1)$. **2) Accuracy** (A) is the average temporal offset calculated over the synchronized collections, normalized with respect to the maximum accepted temporal offset ($maxError$). The synchronization error for a collection i with respect to the reference collection r is defined as $\Delta E_{ir} = |\Delta T_{ir} - \Delta T^*_{ir}|$, where ΔT^*_{ir} is the offset between collection i and collection r calculated on the ground truth. Thus, the Accuracy measure is defined as: $A = 1 - (\sum_{i=1}^{M} \Delta E_{ir})/(M \cdot \mathrm{maxError})$

3) Harmonic mean (H). We combine the aforementioned measures using: $H = (2 \cdot P \cdot A)/(P + A)$.

4.2 Comparison of similarity measures

We conducted tests using each one of the GC, S, CA similarity measures together with PL similarity, discussed in Section 3.2, against using the proposed combination of all similarity measures to discover links between different user collections. PL similarity cannot be used alone, since the fact that two photos were taken in the same or close-by location does not suffice for reliably discovering links between photo collections (e.g. during an Olympic Games event, several competitions take place in the same stadium). For synchronization in this set of experiments we used the first of the two graph traversal algorithms of Section 3.3.

The results are shown in Table 1. The first three columns show the evaluation results on the *Vancouver* dataset and the last three on the *London* dataset. We can see from these results that, for instance, the CA+PL similarity measure combination scored $H = 0.865$ for the *Vancouver* dataset, while scoring only $H = 0.257$ for the *London* dataset, indicating that using a single visual similarity measure does not suffice for performing temporal synchronization across significantly different datasets.

4.3 Comparison of graph traversal methods

We continued with experimentally comparing the two graph traversal algorithms of Section 3.3. For this set of experiments we used the combination of all photo similarity measures, which was shown in the previous section to be beneficial. The results are shown in Table 2. It is evident that, out of the two traversal algorithms, the MST-Med performs the best, and for this reason MST-Med is adopted in all subsequent experiments.

Table 1: Results of the proposed method using different combinations of photo similarity measures

	Vancouver			*London*		
Features	P	A	H	P	A	H
GC+PL	0.235	0.776	0.361	0.167	0.694	0.269
S+PL	0.206	0.527	0.296	0.250	0.513	0.340
CA+PL	0.912	0.622	0.865	0.167	0.559	0.257
All	**0.971**	**0.860**	**0.912**	**0.639**	**0.750**	**0.690**

Table 2: Results of the proposed synchronization method using different graph traversal approaches

	Vancouver			*London*		
Features	P	A	H	P	A	H
MST-Med	**0.971**	**0.860**	**0.912**	**0.639**	**0.750**	**0.690**
MST-Av	0.743	0.715	0.729	0.324	0.513	0.340

4.4 Comparison to other methods

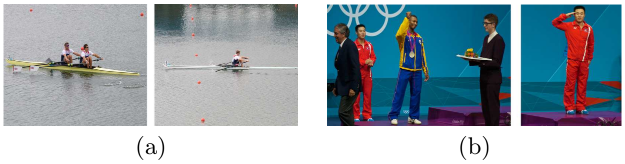
We compare the proposed method against the majority of the literature methods discussed in Section 2 ([4, 21, 1, 12, 16]). For those of the above papers that present more than one variations of their techniques, we report in Table 3 the results of the best-performing technique of each paper. As already mentioned we strictly followed the experimental setup of the MediaEval SEM task and used their datasets and ground truth annotations, which makes our results directly comparable with those published in [21, 1, 12, 16]. For comparing with [4], we re-implemented the method proposed in this paper and tested it under the same experimental setup used for the proposed and all other compared methods. We can see from Table 3 that the proposed method scores the best P and H measures out of all the tested methods. The method presented in [12] scored the best A measure for the *London* dataset, but managed to synchronize only a small subset ($P = 0.15$) of the collections.

Table 3: Comparison between the proposed method and the methods of [4, 21, 1, 12, 16]).

	Vancouver			*London*		
Method	P	A	H	P	A	H
[4]	0.618	0.816	0.703	0.333	0.886	0.484
[21]	0.941	0.792	0.860	0.472	0.875	0.613
[1]	0.912	0.728	0.810	0.611	0.713	0.658
[12]	0.050	0.650	0.093	0.150	**0.920**	0.258
[16]	0.350	**0.860**	0.605	0.250	0.890	0.390
Proposed	**0.971**	**0.860**	**0.912**	**0.639**	0.750	**0.690**

5. CONCLUSIONS

We presented a method to temporally align different user collections of photos captured during the same event. The proposed approach was tested on two benchmark datasets



(a) (b)

Figure 4: Indicative example pairs of photos where S or CA similarity are higher than GC similarity. In (a), GC=52.24% while S=82.39%. In (b), GC=62.52% while CA=83.85%.

and is shown to outperform the most recent literature methods. The superiority of the proposed method is due to two introduced novelties: 1) The combination of multiple photo similarity measures; this allows for different aspects of similarity to be captured (for an indicative example see Fig. 4). 2) A weighted graph-based representation of photo collections that helps us to synchronize photo collections, even those with low coverage of the event. Furthermore, the weighted graph-based representation may be seen as a first step towards multi-user photo collection organization, enabling the application of sub-event clustering and summarization methods even on the same graph, following the synchronization of the photos' timestamps.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

[1] K. Apostolidis, C. Papagiannopoulou, and V. Mezaris. CERTH at MediaEval 2014 Synchronization of Multi-User Event Media task. In *MediaEval Workshop*, Oct. 2014.

[2] R. Arandjelovic and A. Zisserman. All about VLAD. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1578–1585, June 2013.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346 – 359, June 2008.

[4] M. Broilo, G. Boato, and F. G. De Natale. Content-based synchronization for multiple photos galleries. In *IEEE Int. Conf. on Image Processing (ICIP)*, pages 1945–1948, Sep. 2012.

[5] S. A. Chatzichristofis and Y. S. Boutalis. CEDD: Color and edge directivity descriptor. In *Int. Conf. on Computer Vision Systems*, ICVS, pages 312–322, Berlin, Heidelberg, 2008. Springer-Verlag.

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

[7] C. Dimoulas and A. Symeonidis. Syncing shared multimedia through audiovisual bimodal segmentation. *MultiMedia, IEEE*, 22(3):26–42, July 2015.

[8] J. Kammerl, N. Birkbeck, S. Inguva, D. Kelly, A. Crawford, H. Denman, A. Kokaram, and C. Pantofaru. Temporal synchronization of multiple audio signals. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4603–4607, May 2014.

[9] G. Kim and E. P. Xing. On multiple foreground cosegmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 837–844, June 2012.

[10] G. Kim and E. P. Xing. Jointly aligning and segmenting multiple web photo streams for the inference of collective photo storylines. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 620–627, June 2013.

[11] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, Nov. 2004.

[12] P. Nowak, M. Thaler, H. Stiegler, and W. Bailer. JRS at event synchronization task. In *MediaEval Workshop*, Oct. 2014.

[13] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. Journal of Computer Vision*, 42(3):145–175, May 2001.

[14] A. Pigeau and M. Gelgon. Spatio-temporal organization of one's personal image collection with model-based ICL-clustering. In *Int. Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 111–118, Sep. 2003.

[15] A. Pigeau and M. Gelgon. Building and tracking hierarchical geographical & temporal partitions for image collection management on mobile devices. In *ACM Int. Conf. on Multimedia*, pages 141–150, Nov. 2005.

[16] E. Sansone, G. Boato, and M.-S. Dao. Synchronizing multi-user photo galleries with MRF. In *MediaEval Workshop*, Oct. 2014.

[17] P. Sinha. *Automatic Summarization of Personal Photo Collections*. PhD thesis, Long Beach, CA, USA, June 2011.

[18] Y. Sun, W. Wang, X. Gong, C. Chen, X. Yang, and J. Ma. Event clustering of digital media in personal mobile device based on spatio-temporal clustering. In *IEEE Int. Conf. on Broadband Network and Multimedia Technology (IC-BNMT)*, pages 1115–1119, Oct. 2010.

[19] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3360–3367, June 2010.

[20] J. Yang, J. Luo, J. Yu, and T. Huang. Photo stream alignment and summarization for collaborative photo collection and sharing. *IEEE Trans. on Multimedia*, 14(6):1642–1651, Dec. 2012.

[21] M. Zaharieva, M. Riegler, and M. Del Fabro. Multimodal synchronization of image galleries. In *MediaEval Workshop*, Oct. 2014.

[22] 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. on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 9(1):4, Feb. 2013.