Automatic synchronization of multi-user photo galleries

E. Sansone, K. Apostolidis, N. Conci, G. Boato, V. Mezaris, F.G.B. De Natale

***Abstract*—In this paper we address the issue of photo galleries synchronization, where pictures related to the same event are collected by different users. Existing solutions to address the problem are usually based on unrealistic assumptions, like time consistency across photo galleries, and often heavily rely on heuristics, limiting therefore the applicability to real-world scenarios. We propose a solution that achieves better generalization performance for the synchronization task compared to the available literature. The method is characterized by three stages: at first, deep convolutional neural network features are used to assess the visual similarity among the photos; then, pairs of similar photos are detected across different galleries and used to construct a graph; eventually, a probabilistic graphical model is used to estimate the temporal offset of each pair of galleries, by traversing the minimum spanning tree extracted from this graph. The experimental evaluation is conducted on four publicly available datasets covering different types of events, demonstrating the strength of our proposed method. A thorough discussion of the obtained results is provided for a critical assessment of the quality in synchronization.**

***Index Terms*—Markov Networks, Weighted Graph, Multimodal, Multimedia Synchronization, Events.**

I. Introduction

The proliferation of multimedia capturing devices and ubiquitous connectivity have increased the possibility of sharing photos and other multimedia content over the Internet through social networks, shared spaces, cloud storage systems, and various other multimedia sharing services. The design of methods and systems to efficiently manage and organize this ever-increasing amount of data represents a great challenge for the multimedia research community [1].

A situation that frequently occurs in this context is the need of users to share media captured at an event they attended (a concert, a ceremony, a sport or public event, etc.) with others and/or within a social community interested in that event. The final goal of sharing would be the creation of a shared repository, where all the contributors, and possibly users who did not attend in person, can find an enriched view of the event by observing it from multiple perspectives, with finer granularity and better completeness. Data could then be exploited in various contexts including, for example, the creation of storyboards, automatic summaries, and personalized albums. An illustration of the problem is depicted in Fig. 1. However, in order to achieve this, data has to be organized consistently, so as to allow browsing across different dimensions, e.g. sorting based on the time of acquisition, clustering by location, and/or ranking by visual similarity according to a given query.

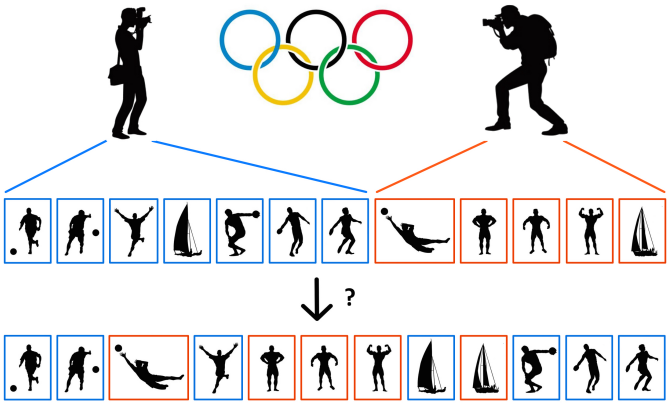

Fig. 1: Illustration of the problem of synchronizing multi-user photo galleries.

Several approaches dealing with different facets of the general problem of multimedia data organization have appeared in the literature. These include techniques for clustering, summarization, visualization, and semantic analysis. Clustering techniques, e.g. [2], aim at partitioning a collection of multimedia data into clusters of similar or related items. Summarization techniques, on the other hand, e.g. [3], [4], aim at producing a concise yet informative version of the original dataset. Visualization methods, e.g. [5], seek to present multimedia data, be it original collections or clustered/summarized versions of them, for an enhanced user experience. Lastly, semantic analysis and annotation methods have seen great progress in the last few years. In fact, thanks to a better knowledge about *how* humans perceive and interpret the multimedia data (e.g., by means of event detection [6], [7], [8], person detection and identification [9], [10], photo tagging [11], [12]), they provide valuable support to data organization according to a human-centered perspective.

In order to ensure an efficient organization of media galleries, and in particular when dealing with *multi-user* collections, the temporal dimension is very critical. In such scenarios each gallery contains photos from a single user/device, therefore the timestamps of photos within a gallery are consid-

Emanuele Sansone, Nicola Conci, Giulia Boato, and Francesco G.B. De Natale are with the Department of Information Engineering and Computer Science - DISI, University of Trento, 38123, Italy; e-mail: {e.sansone, conci, boato, denatale}@unitn.it.

Konstantinos Apostolidis and Vasileios Mezaris are with the Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), Thermi 57001, Greece; e-mail: {kapost, bmezaris}@iti.gr.

ered to be consistent, while the temporal relation of photos from different galleries is affected by an unknown offset. A consistent time annotation allows easily aligning photos belonging to different galleries along a common axis, thus making it possible to locate and cluster important sub-events, match different users' views of the same happening, and ultimately create a unique and consistent flow of information, the *story*. Therefore, the lack of temporal data annotation or its limited accuracy/reliability may cause major problems in media management. This is indeed the most common situation: although we can reasonably assume the coherence of the timestamps provided by a single device within the time-frame of an event, nothing or little can be said about the absolute correctness of such time-based annotations, and thus we cannot guarantee the coherence of timestamps generated by different devices. Possible sources of error include wrong or missing setting of the camera clock, the use of multimedia capturing devices with different time zone settings, missing timestamp information, post-processing of the media with modification or removal of timestamps. These situations prevent a correct synchronization of the different galleries provided by different users, which is indeed necessary for an effective organization of the data.

In this scenario it is necessary to estimate the temporal offset among the clocks of the different users' devices, taking into account all the available information: the metadata of the photos, when available, often stored in the EXIF header, together with the visual content of the images. As far as the EXIF metadata is concerned, we consider date and time of acquisition, together with the geo-location, according to the information provided by the GPS receiver. Note that the availability of such information is not a pre-requisite for synchronization, and in particular the GPS coordinates are often missing, while timestamps are considered consistent only for the photos captured by the same device (referred to, as a gallery).

In this paper, we present a method to achieve an automatic synchronization of multi-user galleries, which outperforms the methods found in the literature. Photo similarity is assessed by extracting Deep Convolutional Neural Network (DCNN) features. Similar photos are then used to define a graph of galleries. Each pair of galleries in the graph is finally synchronized using a probabilistic graphical model. Experimental results over different event types show that the proposed method is able to achieve very good synchronization precision and accuracy.

The paper is structured as follows: Section II reviews the related work in this research area. Section III presents the method we propose. Section IV reports the results of experimental evaluation and comparisons on four benchmark datasets, also including a thorough commentary on the complexity of the proposed method. Concluding remarks are provided in Section V.

II. Related Work

Different pieces of information can be used by a single user to properly arrange his photo galleries, and the temporal data is generally ranked among the most useful elements, as demonstrated in [13], [14]. In [15], [16], [17] the timestamps provided in the metadata of digital photos are used for photo gallery organization purposes, as for example alignment and summarization. Such approaches are shown to be very effective on galleries captured by a single device, where the provided timestamps are consistent. Extending the photo organization use case to multiple cameras (as in the case of shared repositories of photo galleries, where different users share their pictures and videos of the same event), the overall consistency of the timestamps becomes questionable. This is due to the fact that different users may have different time settings (time zones, wrong clock setting) in their devices, and, as a consequence, directly using these timestamps can introduce noise in any photo organization process.

The literature reports some works that, although not matching exactly the problem we are addressing, do deal with multi-user photo galleries and introduce some notion of information alignment. In [18], storyline graphs are produced from web communities' photos, for supporting photo recommendation applications. In [19], still photos are temporally aligned with video footage, while in [20] photos of different blog posts are grouped using natural language processing of the text found in the blogs. The above methods focus on summarizing media collections and, when performing media alignment with the use of photo timestamps, they treat them as if they were consistent across different photo galleries. A different photo-sequencing problem is addressed in [21]: given a set of images depicting the same object, temporal ordering is achieved by computing the optimal similarity transformation of all images to the first image in the sequence. The warped images can then be animated in the computed temporal order, as a single video sequence. However, this method makes the strong assumption that all pictures depict the same rigid object. Looking a bit beyond images, the alignment of different time-series, which are typically obtained by repeated measurements of a biological, chemical or physical process, is a well studied problem (e.g. the method in [22]). However, such methods rely on finding common (exact-duplicate) sub-sequences between two data time-series, and their application to our temporal photo alignment problem is therefore not straightforward.

Previous methods that deal with our problem include [23], [24], [25], [26], [27]. In [23], a collective storyline of photos is constructed by exploiting the visual information. Each image is segmented in foreground objects and background, so as to assess photo similarity by detecting instances of the same objects across galleries. A user-defined parameter K denotes the number of foreground areas in which a photo can be split, using a Multiple Foreground Co-segmentation algorithm [28]. However, real-world photo collections are not always centered on objects that can be easily segmented, e.g., photos of a music festival may contain faces from the audience, as well as musicians faces along with a complex background. An example is provided in Fig. 2, where we show that the segmentation results obtained by applying the algorithm proposed in [28] for a selection of photos in the real-life datasets used in this paper. From these examples it is clear that this approach is not suitable for our synchronization scenario,

due to the imperfections of this (and any other) segmentation algorithm. Broilo et al. [24] attempt a content-based alignment approach to compute the estimated delay among the photos taken with different cameras by finding the suitable pairs of similar photos across the different galleries of the same event; the delay among the galleries is estimated based on the time delays between the selected pairs of pictures. In order to operate properly, however, the method relies on the hypothesis that different photographers take photos that capture the same sub-events. This condition turns out to be very restrictive, especially when considering a real and unsupervised scenario, where no control on the acquisition process can be imposed. Along with visual data, geo-location information can also be used for alignment purposes, such as in the approach by Yan et al. [25], in which a bipartite kernel sparse representation of visual and spatial similarities is used for photo stream alignment and summarization. However, the framework of [25] requires considerable computational power and cannot scale well beyond approximately a hundred images. Several recent works that study the synchronization problem of generic multi-user photo galleries have been proposed and evaluated in relation to the MediaEval SEM (Synchronization of multi-user Event Media) benchmarking activity that was organized in 2014 [29] and 2015 [30]. Concerning the features used for similarity assessment of the images, Zaharieva et al. [31] propose exploiting the MPEG-7 Color Structure Descriptor [32] and a Joint Composite Descriptor. In [33], [34], [35] the SIFT local descriptor is used, while in [36] the SURF local descriptor and HSV color histograms are employed. In [33], the most similar photos between different galleries are detected and a graph of photo similarities is employed to find paths between each gallery and the reference one. In [34], temporal offsets are expressed as a non-homogeneous linear equation system, and an approximate solution is calculated. In [36], a probabilistic graphical model is built, in which each temporal displacement is identified by a set of nearest-neighbor photo pairs across the galleries. The method in [37] is an extension of the method in [33] that can also handle galleries that include video and audio. Finally, in [35] a probabilistic algorithm is employed, where in each run a hypothesis is calculated for the set of time offsets with respect to a reference gallery. The final set of time offsets is calculated as the medoid of all hypotheses. The most relevant works mentioned in this section are summarized in Table I.

In this work we build on our preliminary methods [38] and [36], which we combine and extend in several ways. Specifically, the novelties of the proposed method compared to our older methods (or a straightforward combination thereof), and to the methods of the literature are:

i) The use of DCNN-based features to assess the semantic similarity of photos. By replacing the hand-crafted features of [38], [36] with DCNN-based features, we take advantage of the significant progress made in deep learning. To our knowledge, this is the first work that uses this kind of features for the specific problem of multi-user photo gallery synchronization.
ii) The combination of graph-based representation of galleries and a probabilistic model to perform the synchronization. Graph-based representations in image and video analysis problems are not new (e.g. in [39] and [40] the problems of video-shot and image scene categorization, respectively, are formed as graph partitioning tasks; similarly, the photo synchronization problem in [38] is transformed to a graph-search). Also, the use of a probabilistic model for photo synchronization was first briefly sketched in [36]. However, by extending and combining these two notions we obtain significant benefits: a) we remove the constraint of a fixed reference gallery, by employing the probabilistic model on top of the constructed graph of galleries (in contrast to [36], which made the restrictive assumption that a single reference gallery was specified in advance, and all other galleries shared similar photos with it). b) We relieve the probabilistic graphical model from the need of using visual features, which are only employed at the graph construction stage to assess the visual similarity between photos and to define the potential temporal offsets. c) Compared to [36], in which all possible offsets are taken into account, we treat as potential candidates in the probabilistic graphical model only the finite number of values that correspond to the time differences between similar photos across each pair of galleries in the graph. Overall, the changes to the model lead to a speed-up of three orders of magnitude in the inference stage, compared to [36]. d) Adopting the aforementioned probabilistic model, we perform global optimization on the offsets available for each pair of galleries (as opposed to [38]).
iii) The construction of an overall photo synchronization method that is free of thresholds (as opposed to both [36] and [38]) and thus is able to yield good results in diverse datasets.

These elements guarantee the effectiveness of the proposed approach on all datasets, as will be discussed in Section IV.

Fig. 2: Examples of the segmentation algorithm of [28] applied on a selection of photos from the datasets we use for our experiments.

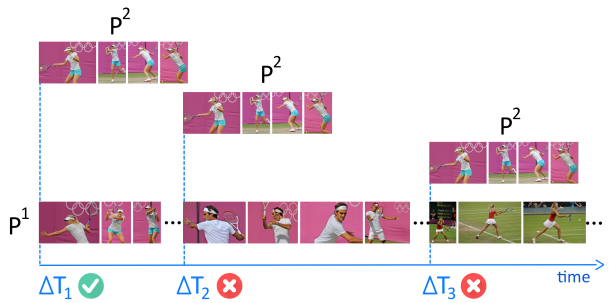

Fig. 3: Example of galleries taken from one of the datasets used in this work. On the bottom, the reference gallery. On the top, the second gallery that has to be synchronized with respect to the reference according to three possible hypotheses (i.e., three possible offsets). In this case, $\Delta T^* = \Delta T_1$.

III. Temporal Synchronization of Multi-User Photo Galleries

A. Problem Statement

Let us assume we are given two photo galleries, namely $P^1 = (I_1^1, I_2^1, \ldots, I_{N_1}^1)$ and $P^2 = (I_1^2, I_2^2, \ldots, I_{N_2}^2)$, with the associated sequence of timestamps $T^1 = (t_1^1, t_2^1, \ldots, t_{N_1}^1)$ and $T^2 = (t_1^2, t_2^2, \ldots, t_{N_2}^2)$. In this case, I_l^i is the l-th photo of the i-th gallery and t_l^i is the corresponding timestamp. Considering that each gallery is acquired by a single device, we can assume that the relative time differences between the elements in P^1 (and similarly for P^2) are correct. Conversely, we do not know if the two sets T^1 and T^2 have the same absolute reference time or if they have been affected by an unknown offset. Therefore the problem consists of estimating the value of this temporal offset, called ΔT^*, and estimating one (in general, different) such value for each different pair of galleries. A visual representation of the problem is shown in Fig. 3.

B. General Approach

Figure 4 illustrates the main stages of the proposed approach for the automatic temporal synchronization of multi-user photo galleries. Initially, similar photos across different galleries are identified. Visual similarity is evaluated exploiting the features extracted from a pre-trained Deep Convolutional Neural Network (DCNN), as explained in Section III-C. The most similar photos from different galleries are considered as links between different pairs of photo galleries. Subsequently, we construct a graph where nodes represent galleries, and edges represent the discovered links between them, as explained in Section III-D. Temporal synchronization of the galleries is achieved exploiting a probabilistic graphical model. As explained in Section III-E, for this purpose visual similarity is modeled through potential functions, and the max-sum algorithm is used for the temporal offsets estimation. In order to get a fast response without sacrificing the accuracy, we apply a coarse-to-fine optimization approach.

C. Photo Similarity Assessment

As can be ascertained from Section II, local descriptors (e.g. SIFT, SURF) are most often used to produce a representation of the visual content of the photos. Driven by the success of DCNNs in large scale image classification tasks, DCNN-based visual representations are quickly gaining ground and are used in image retrieval tasks as well, exhibiting state-of-the-art performance ([41], [42], [43], [44], [45], [46]).

To identify similar photos of different galleries, we adopted the method proposed in [42], which uses DCNN-based features to assess the similarity of photos. We choose to use the GoogLeNet network [47], due to its fast response in the testing phase. The network is pre-trained on ImageNet data using the Caffe deep learning framework [48], and is publicly available on the Caffe model zoo[1]. In the 22-layer deep GoogLeNet network, there are 9 inception layers sequentially connected. According to [47], any input image is resized to 224×224 pixels and segmented to equal and non-overlapping regions. An inception layer extracts convolutional filters responses from each such region of the input image, which can be used as features to describe the visual content of the specific region (the number of regions that an image is segmented to, as well as the size of each region and of the corresponding feature vector, are reported in Table II). In this way we generate a single collection of L2-normalized filter responses from a selected DCNN layer (i.e., one collection for all regions of all images, and all filters of that layer), and we perform k-means clustering in this collection to obtain a vocabulary of 256 words. Then, each filter response vector is assigned to its nearest visual word, enabling the subsequent encoding of all the filter responses from all regions into one feature vector using VLAD encoding [49]. Finally, the VLAD descriptors are normalized by intra-normalization [50].

Taking into consideration the evaluation carried out in [42], we choose to experiment with features from the "conv2/norm2", "inception3a/output", "inception4a/output" and "inception5a/output" layers as these layers scored the best results in different datasets. For the sake of completeness we also conducted two additional experiments: we used directly the final classifier layer ("loss3/classifier", without any further encoding of its output) and we also implemented a fusion

[1]https://github.com/BVLC/caffe/wiki/Model-Zoo

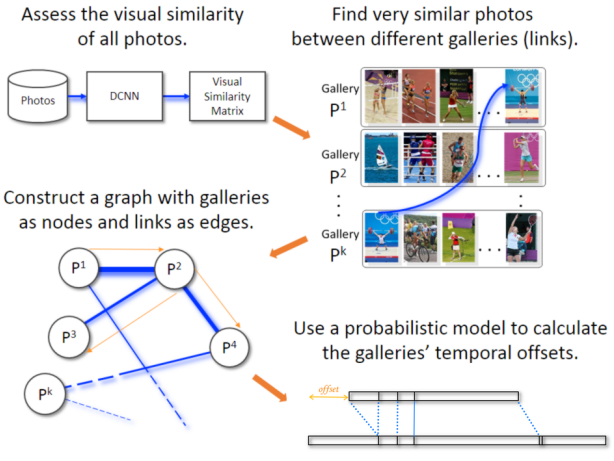

Fig. 4: Overview of the proposed method.

TABLE I: Review of the most relevant works presented in Section II. For each reference, the objective, the features used and the datasets adopted for validation are reported. The indication *personal collection* refers to non-standard datasets retrieved from the Internet or from private material. The last column reports the main limitation of each approach.

Refs.	Objective	Features	Dataset	Notes
[15][16][17]	Summarization	Visual, Metadata, Text	CeWe 2009, Collection from Facebook	Galleries from a single camera, timestamps assumed reliable
[18][19][20]	Alignment, summarization	Visual, Text	Collection from Flickr, YouTube	Timestamps assumed reliable
[21]	Temporal ordering	Visual	Personal collection	All images need to depict the exact same object from different viewpoints
[23]	Synchronization	Visual	Collection from Flickr	The same objects, which can be easily segmented, need to be depicted in the different galleries
[24]	Synchronization	Visual	Personal collection	The same sub-events must be captured in the different galleries
[25]	Synchronization	Visual + GPS	Personal collection	Difficult to scale due to high computational complexity
[31]	Synchronization	MPEG-7 CSD	MediaEval2014	Synchronization is based on pairwise comparison of galleries (limited scalability with respect to the number of galleries)
[34]	Synchronization	Visual (SIFT)	MediaEval2014	Synchronization is based on pairwise comparison of galleries (limited scalability with respect to the number of galleries)
[33]	Synchronization	Visual (SIFT)	MediaEval2014	-
[36]	Synchronization	Visual (SURF+HSV)	MediaEval2014	A reference gallery that covers most of the event must be provided
[38]	Synchronization	Visual (SIFT+GIST+HSV)	MediaEval2014	-
[35][37]	Synchronization	Visual + Audio	MediaEval2015	-

TABLE II: Feature sizes using different inception layers and the final classifier layer of GoogLeNet, according to [47], and the resulting feature vector size following VLAD aggregation for the inception layers' features.

GoogLeNet Layer Name	Number of regions	Region Size	Layer feature size	Resulting feature vector size using a vocabulary of 256 centers
conv2/norm2	28 x 28	8 x 8	192	49152
inception3a/output	28 x 28	8 x 8	256	65536
inception4a/output	14 x 14	16 x 16	512	131072
inception5a/output	7 x 7	32 x 32	832	212992
loss3/classifier	1	224 x 224	1000	N/A

approach on features of all the above selected layers (except the "loss3/classifier"), obtained by averaging the corresponding similarity matrices. The resulting feature vector size after VLAD aggregation, for the different layers we test, is reported in Table II. The preliminary experiments with these different layers, which lead to the final choice adopted in our method, are reported in Section IV-B.

We construct the visual similarity matrix $\boldsymbol{W}$ of all images in a collection as:

$$\boldsymbol{W}(i,j) = \exp\{-D(V_i, V_j)\},\ \forall i,j \in \{1,\ldots,N\} \quad (1)$$

where V_i and V_j are the VLAD vectors of I_i and I_j photos, respectively, D is the Euclidean distance function and N is the total number of images in the collection. Each pair of photos I_i^k and I_j^l, where $k \neq l$ (i.e., the photos do not belong to the same gallery), is treated as a potential link between the k and l galleries. To define the number of links that we utilize for each pair of galleries we introduce two alternative approaches:

- *exact* approach: for each pair of galleries we keep the $\lfloor \alpha N \rfloor$ most significant links, where N is the total number of images in the collection (and α is a constant).
- *coverage* approach: we define graph coverage as the ratio of connected galleries pairs to the total number of possible gallery pairs. The latter is equal to $k(k-1)/2$, where k is the number of galleries in the collection. We start by selecting the most significant link (i.e., the pair of photos with the highest visual similarity) and we compute the current graph coverage. We repeat this process, and when we reach certain values of graph coverage we stop collecting new links for the pairs of galleries, which have at least $\lfloor \alpha N \rfloor$ links. These values of graph coverage are arbitrarily set starting from 10% (the first one) up to 100% with a step of 10%. The intuition behind this procedure is that we select only the strong links when available, but we also explore weaker links for the pairs of galleries that do not exhibit strong links, in order to be able to synchronize as many galleries as possible.

For all the experiments conducted in Section IV, we set $\alpha = 0.1$ based on a set of preliminary evaluations that indicated good balance between processing speed and detection accuracy.

D. Galleries Graph Construction

Having identified potential links for at least some gallery pairs, it is relatively straightforward to construct a weighted graph, whose nodes represent the galleries, and edges represent the links between galleries. The weight assigned to each edge is calculated as the median of the similarity values

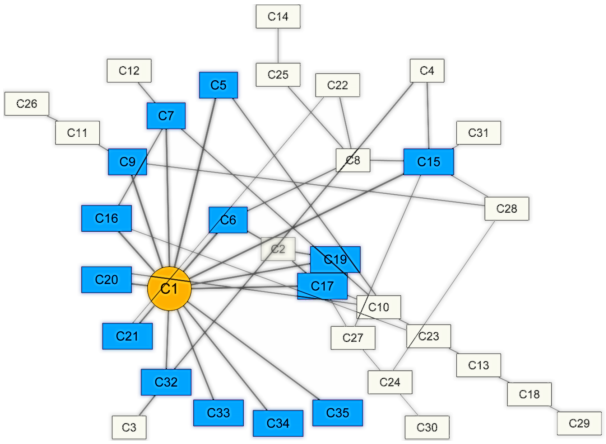

Fig. 5: Illustration of the galleries graph for the *Vancouver* dataset used in our experiments.

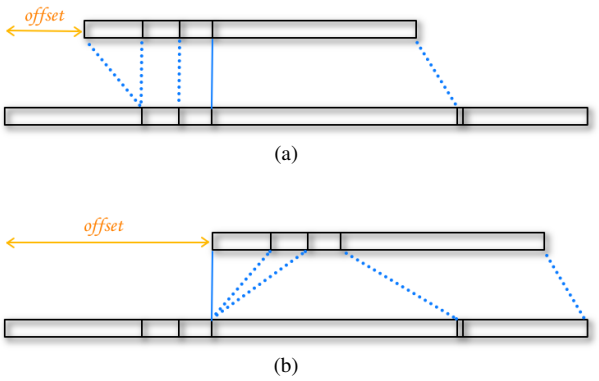

Fig. 6: Examples of correspondences between two galleries with different offsets. In each subfigure the lower stripe is the reference and the upper is the gallery to be synchronized; black vertical lines correspond to pictures, dotted and continuous blue lines are correspondences between photos.

of the photos pairs that serve as links between the two galleries. We follow this approach for calculating the edges' weights, in contrast to [38], based on preliminary experiments, which showed that this can slightly improve the performance (compared to [38]) in three out of the four datasets employed in this work (see Section IV-A for details on the employed datasets). Finally, using this graph, the temporal offsets of each gallery will be computed against one (random) gallery that is considered as the reference. Any gallery can be considered as the reference one, since we are aiming for relative synchronization; we neither assume nor need one gallery's timestamps to accurately match the true time.

It should be stressed here that both the *exact* and *coverage* approaches discussed in Section III-C do not guarantee the discovery of links between the reference gallery and every other gallery, as shown in Fig. 5. In the figure, representing the galleries graph of one of the datasets used for evaluation, the first gallery is arbitrarily chosen as the reference gallery, and the corresponding node $C1$ is shown as an orange circle. Nodes corresponding to galleries that have direct links to the reference gallery, identified using the procedure described in Section III-C, are depicted as blue rectangles. All other nodes of the graph are depicted as white rectangles. Attempting to directly compute the pairwise offsets between the reference gallery and each other gallery would not work, since only a portion of the galleries is directly connected with the reference gallery (regardless of which gallery was chosen as the reference, since the graph of Fig. 5 is not fully connected). Instead, we must find a way to properly traverse the graph in order to estimate an offset for each gallery.

E. Temporal Offsets Estimation

Given the graph of galleries, the simplest approach to estimate the temporal offsets would be to use the method presented in [38] and traverse the minimum spanning tree (MST) of the galleries graph, synchronizing each pair of galleries using the median of the temporal offsets of all links connecting the two galleries. However, in order to improve the performance, we propose here a more elaborate approach based on probabilistic graphical models that substitutes the use of median, while still traversing the graph in the order specified by the MST (i.e., selecting pairs of galleries for offset estimation).

Probabilistic graphical models allow to handle dependencies between random variables and to simultaneously take into account the uncertainty of inference. More specifically, we adopt undirected graphs, namely Markov Random Fields (MRFs), to model the similarities between photos as mutual dependence properties.

From the minimum spanning tree, we compute the offset between galleries by finding the best correspondences of photos. In Fig. 6 black stripes and black vertical segments are used to symbolize galleries and photos, respectively. In each sub-figure, the lower stripe corresponds to the reference gallery, while the upper stripe is the gallery to be synchronized. The two examples depict two different situations with different offsets. Each of these offsets is defined according to a pair of visually similar images, which correspond to a potential link discovered at the previous stage (and visualized as continuous blue line). All other images are compared against their closest reference neighbors given that offset (shown as dotted blue lines). The set of all image comparisons is called the set of correspondences. It is evident from Fig. 6, that the first offset produces a better set of correspondences than the second one, due to a better satisfaction of the temporal constraints imposed by the acquisition timestamps of images in both galleries. This definition of offsets, and thus correspondences, considerably differs from the model in [36]. In fact, the possible offsets in [36] are obtained by discretizing the temporal axis and by considering only those time instants that produce different sets of correspondences. Here, the offsets are defined based on the potential links obtained in the previous stage and the set of correspondences are defined according to each of these possible offsets. As mentioned previously, the advantage of this definition compared to the one proposed in [36] is twofold: (i) a speedup of the inference stage due to a decreased set of possible solutions; (ii) the removal of the discretization

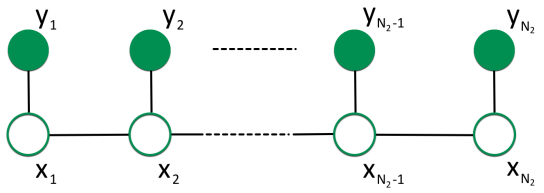

Fig. 7: Markov network with observed and hidden nodes.

parameter, which required in [36] two different types of synchronization (a coarse and a fine synchronization).

Notice that the number of images in each gallery in input to the graphical model is generally smaller compared to their original versions, since only visually similar images are now considered. In order not to complicate the notation, we keep denoting with N_1 and N_2 the size of the processed galleries (namely, the ones provided by the graph construction stage, see Section III-D). Let us now assume that gallery P^1 is the reference and that P^2 has to be synchronized with respect to P^1. We start by defining the set of all possible temporal offsets as $\{\Delta T_m : m = 1, \ldots, Q\}$ (Q is the cardinality of the set); these offsets are determined by the potential links (see Section III-C) of the two galleries, namely by the pairs of visually similar photos that have been found (the difference of timestamps between the photos in each pair is considered a possible temporal offset). Then, by defining the sequence of correspondences between the two galleries, and given the offset ΔT_m, as $\mathbf{x}^{\Delta T_m} = (x_1^{\Delta T_m}, \ldots, x_l^{\Delta T_m}, \ldots, x_{N_2}^{\Delta T_m})$, where $x_l^{\Delta T_m}$ represents the photo in the reference gallery associated with photo I_l^2, the synchronization can be formulated as an optimization problem. In other words,

$$\Delta T^* = \arg\max_{\Delta T_m} f(\mathbf{x}^{\Delta T_m}) \tag{2}$$

where $f : X \rightarrow \mathbb{R}$ is the function that assigns a similarity score to each sequence of correspondences and $X = \{\mathbf{x}^{\Delta T_m} : m = 1, \ldots, Q\}$ is the set of all possible candidate solutions. Therefore, the optimization consists of finding the best set of correspondences in order to maximize the similarity between galleries.

By introducing the sequence of observed nodes $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_{N_2})$, where y_i refers to photo I_i^2 in P^2, and considering the sequence of latent variables $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_{N_2})$ with event space X, it is possible to define an undirected graphical model, as shown in Fig. 7. The edges of the graph serve two purposes: edges between nodes in $\mathbf{x}$ and $\mathbf{y}$ are used to compare images across the two galleries, while edges between nodes belonging to the sequence $\mathbf{x}$, are used to catch the temporal correlation between images within the same gallery.

Function f in (2) is defined as the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$. The dependence properties defined by the graph in Fig. 7 permit to factorize the distribution, namely considering a potential function for each maximal clique, where in our case a maximal clique consists of just a single edge in the graph:

$$f(\mathbf{x}) \doteq p(\mathbf{x}|\mathbf{y}) = \frac{1}{Z} \prod_i \psi(x_i, x_{i+1}) \prod_k \phi(x_k, y_k). \tag{3}$$

In (3), $\psi(x_i, x_{i+1})$ and $\phi(x_k, y_k)$ are the potential functions associated to the edge connecting x_i and x_{i+1} and to the edge connecting x_k and y_k, respectively, while Z is the normalization constant.

1) Definition of the potential functions: The potential $\phi(x_k, y_k)$ in (3) quantifies the dissimilarity between pairs of photos across the two galleries according to localization information. Intuitively, photos acquired from the same location are more likely to refer to the same event. Conversely, photos acquired in completely different locations are more likely to be related to different events.

When available, the GPS information stored in the photo headers is used to compare a pair of photos on the basis of the spatial distance between their acquisition locations. GPS coordinates, expressed in terms of latitude and longitude, can be transformed into spherical coordinates by assuming that the Earth is spheric. In fact, if u_I^G and v_I^G are the latitude and the longitude coordinates for photo I and u_I^S, v_I^S and w_I^S are used to identify the correspondent spherical components, then it is possible to apply the following transformation:

$$\mathbf{z}_I = \begin{cases} u_I^S = R \cos\left(\frac{2\pi u_I^G}{360}\right) \cos\left(\frac{2\pi v_I^G}{360}\right) \\ v_I^S = R \cos\left(\frac{2\pi u_I^G}{360}\right) \sin\left(\frac{2\pi v_I^G}{360}\right) \\ w_I^S = R \sin\left(\frac{2\pi u_I^G}{360}\right) \end{cases} \tag{4}$$

where $\mathbf{z}_I = (u_I^S, v_I^S, w_I^S)$ is introduced for notation compactness and R is the average radius of the Earth. The orthodromic distance between the acquisition locations of two photos I_1 and I_2 can be therefore computed using the following relation:

$$D_G(I_1, I_2) = 2R sin^{-1}\left(\frac{D(\mathbf{z}_{I_1}, \mathbf{z}_{I_2})}{2R}\right) \tag{5}$$

where $D(\mathbf{z}_{I_1}, \mathbf{z}_{I_2})$ is the Euclidean distance between $\mathbf{z}_{I_1}$ and $\mathbf{z}_{I_2}$. In case images are not geo-tagged, the distance in (5) is set to 0 and the localization information does not therefore contribute to the estimation of the offset.

The potential $\phi(x_k, y_k)$ in (3) is defined as:

$$\phi(x_k, y_k) = \exp\left\{ -\gamma \frac{D_G(x_k, y_k)}{D_G^{max}(k)} \right\} \tag{6}$$

where x_k and y_k are the observed and hidden nodes of the MRF model, respectively, $D_G^{max}(k) = \max_{x_k}\{D_G(x_k, y_k)\}$ is the normalization term and γ is a real positive parameter used to scale the importance of the orthodromic distance to offset estimation.

The potential $\psi(x_i, x_{i+1})$ in (3) is instead used to account for the temporal dimension. In particular, it considers pairs of photos for both galleries and measures the quality of the alignment given a particular offset. The measure is based on the distance D_T using a metric, namely the l_1-norm:

$$D_T(x_i, x_{i+1}) = \left\| \begin{bmatrix} t_{y_i} \\ t_{y_{i+1}} \end{bmatrix} - \begin{bmatrix} t_{x_i} \\ t_{x_{i+1}} \end{bmatrix} \right\|_1 \tag{7}$$

where t_{y_i} and $t_{y_{i+1}}$ denote the timestamps of the i-th and $i+1$-th image in P^2, while t_{x_i} and $t_{x_{i+1}}$ are the timestamps of the corresponding images in P^1. If $t_{y_j} = t_{x_j}$ and $t_{y_i} = t_{x_i}$, then the image pairs are perfectly aligned and $D_T(x_i, x_{i+1}) = 0$. The potential $\psi(x_i, x_{i+1})$ can then be defined similarly to (6):

$$\psi(x_i, x_{i+1}) = exp\left\{-\delta \frac{D_T(x_i, x_{i+1})}{D_T^{max}(i)}\right\} \tag{8}$$

where $D_T^{max}(i) = \max_{x_i, x_{i+1}} \{D_T(x_i, x_{i+1})\}$ is the normalization term and δ is a real positive parameter used to weight the temporal distance component.

It is worth mentioning that the potentials $\phi(x_k, y_k)$ and $\psi(x_i, x_{i+1})$ are defined so that the objective function in (3) is maximized when the correspondences between photos across the two galleries have both strong location similarity and precise temporal alignment. The contribution to the offset estimation of these two terms can be weighted by setting differently the values of the potentials parameters, namely γ, δ.

2) Parameter estimation: Starting from a training set of galleries provided with the ground truth related to the synchronization offset, it is possible to estimate the parameters of the model described in the previous subsections. In fact, if $\boldsymbol{\theta}$ is the parameters vector, namely $\boldsymbol{\theta} \doteq [\gamma, \delta]^T$, and $\mathbf{x}^*$ is the sequence of visually similar photos between a pair of training galleries given the true offset (which can be obtained by using the stages preceding the MRF model), then the optimal parameters $\boldsymbol{\theta}^*$ are estimated as follows:

$$\begin{aligned} \boldsymbol{\theta}^* &= arg\, max_{\boldsymbol{\theta}} \{p(\mathbf{x}^*|\mathbf{y})\} \\ &= arg\, max_{\boldsymbol{\theta}} \left\{ \frac{exp\{-\sum_i \theta_i h_i(\mathbf{x}^*)\}}{\sum_{\tilde{\mathbf{x}}} exp\{-\sum_i \theta_i h_i(\tilde{\mathbf{x}})\}} \right\} \end{aligned} \tag{9}$$

where the second equality in (9) is obtained by substituting (6) and (8) into (3) and by exploiting the following relations:

$$\begin{aligned} h_1(\mathbf{x}) &\doteq \sum_{k=1}^{N_2} \frac{D_G(x_k, y_k)}{D_G^{max}(k)} \\ h_2(\mathbf{x}) &\doteq \sum_{i=1}^{N_2-1} \frac{D_T(x_i, x_{i+1})}{D_T^{max}(i)}. \end{aligned} \tag{10}$$

It is worth mentioning that, since the objective in (9) is convex, the optimal parameters vector $\boldsymbol{\theta}^*$ can be estimated through standard approaches to convex optimization like the gradient descent algorithm. Another important aspect is that the model can be easily integrated with other types of features when available, including textual tags or data crawled from social networks. The wider the set of features, the more robust the model can be in estimating the offset.

3) Offset estimation: After the parameters of the potentials are estimated, the offset is found by computing the optimal sequence of states over the set X, namely finding:

$$\begin{aligned} \mathbf{x}_{max} &= arg\, max_{\mathbf{x}} \{p(\mathbf{x}|\mathbf{y})\} \\ &= arg\, max_{\mathbf{x}} \left\{ \prod_i \psi(x_i, x_{i+1}) \prod_k \phi(x_k, y_k) \right\} \end{aligned} \tag{11}$$

where the second equality holds since the normalization constant Z in (3) does not affect the maximization. The expression above can be solved efficiently through exact inference thanks to the tree-structured shape of our graphical model. In particular, the max-sum algorithm is adopted: firstly, the Markov network has to be converted into an equivalent factor graph; secondly, messages are propagated from the leaves to the root, which can be chosen arbitrarily, and finally the maximization is performed. Fig. 8 shows the exact flow of messages heading towards the root x_{N_2}, highlighted by blue arrows.

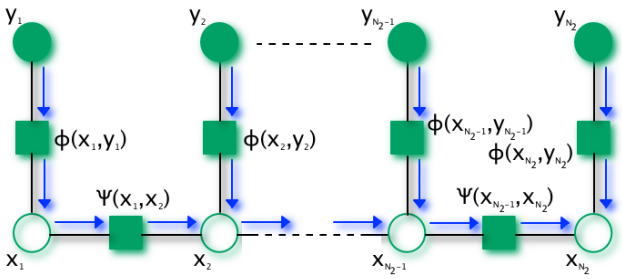

Fig. 8: Equivalent factor graph.

By following the approach in [51] and by introducing $\mu_{a\to b}$ as the message propagating from a generic point a of the graph to any other point b, it is possible to define the complete set of messages in the following way:

$$\begin{aligned} \mu_{y_k \to \phi(x_k,y_k)}(y_k) &\doteq 0 \\ \mu_{\phi(x_k,y_k)\to x_k}(x_k) &\doteq \ln\left(\phi(x_k, y_k)\right) \\ \mu_{x_i \to \psi(x_i,x_{i+1})}(x_i) &\doteq \mu_{\psi(x_{i-1},x_i)\to x_i}(x_i) + \\ &\quad + \mu_{\phi(x_i,y_i)\to x_i}(x_i) \\ \mu_{\psi(x_{i-1},x_i)\to x_i}(x_i) &\doteq \ln\left(\psi(x_{i-1}, x_i)\right) + \\ &\quad + \mu_{x_{i-1}\to\psi(x_{i-1},x_i)}(x_{i-1}) \end{aligned} \tag{12}$$

where $k = 1, \ldots, N_2$ and $i = 1, \ldots, N_2 - 1$.

Once all messages arrive to the root x_{N_2}, the best state, and consequently the best offset, is computed through the maximization:

$$\begin{aligned} x_{N_2}^{max} = \arg \max_{x_{N_2}} \big\{ &\ln\left(\phi(x_{N_2}, y_{N_2})\right) + \\ &+ \mu_{\psi(x_{N_2-1},x_{N_2})\to x_{N_2}}(x_{N_2}) \big\}. \end{aligned} \tag{13}$$

IV. Experimental Results

In this section we present our experimental results. In Section IV-A we describe the datasets and the evaluation procedure we adopt. In Section IV-B we investigate the impact of different approaches used to identify the most similar images between different galleries and the impact of features extracted from different layers of the GoogLeNet DCNN used to assess visual similarity. Subsequently, in Section IV-C we evaluate the effectiveness of the proposed method and compare it to state-of-the-art methods from the literature. Finally, in Section IV-D we discuss time complexity and storage requirements of the proposed method.

A. Datasets and Evaluation Framework

For the experimental validation of the proposed methods we employ the two datasets of the MediaEval 2014 SEM task

TABLE III: Datasets used for evaluation.

	London	*Vancouver*	*NAMM15*	*TDF14*
Number of photos	1351	2124	420	2471
Number of galleries	35	37	19	33

[29], consisting of photos from various users taken during two Olympic Games events, and two datasets of the MedialEval 2015 SEM task [30], consisting of photos from an exhibition and a cycling event, for a total of four different datasets[2]. All photos for each of these datasets are organized in galleries (each gallery captured using a single device) and come with timestamps, which are consistent within each gallery but may have considerable temporal offsets across different galleries. Furthermore, some galleries include geo-location information, while others do not. We strictly followed the experimental setup and evaluation procedure of the MediaEval SEM tasks, making our results directly comparable to the results reported by the tasks participants.

The first test dataset, *Vancouver*, is about the Winter Olympics Games held in 2010, consisting of 1351 photos arranged in 35 galleries, while the second, *London*, is about the Olympic Games held in 2012 and consists of 2124 photos arranged in 37 galleries. The third dataset concerns the famous exhibition held every year in California for music merchants (*NAMM15*), consisting of 420 images and 32 videos, split into 19 galleries with each user gallery containing a variable number of media. In the context of this work we deal only with the images included in the dataset, for simplicity (but our algorithm can be easily extended to take into account videos files as well). Finally, the fourth dataset is related to the Tour de France event held in 2014 (*TDF14*). The dataset is split into 33 galleries and covers the entire competition. As part of the adopted experimental setup, in all datasets the first gallery is considered as the reference one. Table III summarizes the four datasets that we use.

As far as the evaluation metrics are concerned, we have adopted the *precision* and *accuracy* to assess the quality in synchronization, as defined in [29], and briefly reported hereafter for convenience.

1) Precision (P) corresponds to the ratio between the number of synchronized galleries (M_{syn}), and the total number of galleries ($M-1$, excluding the reference gallery) in the dataset. A gallery is considered to be synchronized if the difference between the estimated timestamps of its photos and the corresponding ground truth is lower than a maximum accepted temporal offset ($maxError$). The value of $maxError$ is set to 1800 seconds as per the guidelines in [29]. The Precision measure is defined as:

$$P = \frac{M_{syn}}{M-1}. \tag{14}$$

2) Accuracy (A) is the average temporal offset calculated over the synchronized galleries, normalized with respect to the maximum accepted temporal offset ($maxError$). The synchronization error for gallery P^i with respect to the reference gallery r is defined as $\Delta E_{ir} = |\Delta T_{ir} - \Delta T^*_{ir}|$, where ΔT_{ir} and ΔT^*_{ir} are the offset between gallery i and gallery r obtained from the method and the ground truth, respectively. Thus, the Accuracy measure is defined as:

$$A = 1 - \frac{\sum_{i=1}^{M_{syn}} \Delta E_{ir}}{M_{syn} \cdot \text{maxError}}. \tag{15}$$

3) Harmonic mean (H). We combine the aforementioned measures according to:

$$H = (2 \cdot P \cdot A)/(P + A). \tag{16}$$

In the subsequent experiments, we also report the average and standard deviation of each of the aforementioned evaluation metrics, across the four datasets used.

B. Impact of Different Design Choices

In order to understand the impact of different approaches for constructing the galleries graph, as discussed in Section III-C, we conducted experiments on:

1) The different DCNN layers to extract features from ("loss3/classifier", "conv2/norm2", "inception3a/output", "inception4a/output" and "inception5a/output"), as well as their combination by means of late fusion,
2) The two alternative approaches for identifying the most similar images per gallery pair (*exact*, *coverage*).

The results are shown in Table IV, where we report the synchronization precision and accuracy for each dataset, as well as the averages across all datasets for all three metrics defined in Section IV-A (precision, accuracy, harmonic mean). The best results for each column of this table are shown in bold. According to these results, the "loss3/classifier" layer's features exhibit the worst performance, since this kind of features assesses the purely semantic (rather than visual) similarity of images. In contrast, the best overall results are achieved when using the "inception3a/output" layer's features and the *exact* approach: the harmonic mean reaches 81.7%, the precision also receives its highest value (at 80.3%), and accuracy achieves a score of 83.8%. Considering these results, we can conclude that, in our implementation, the "inception3a/output" layer and the *exact* approach, out of those discussed in Section III-C, represent the best choices for evaluating photo similarity.

C. Comparison with the State-of-the-Art

We compare the proposed method against the most relevant approaches of the literature [24], [31], [34], [33], [35], [37] and also against the preliminary works [36], [38] that the proposed method is based on. To demonstrate the superiority of the MRF-based technique for the calculation of the galleries' offsets, we also include in the comparison: i) a modified version of our earlier method of [38] using the DCNN-based features; ii) a similarly modified version of [36]; iii) a straightforward combination (cascade) of [38] and [36], in which the MST-based procedure of [38] is replaced by the

[2]The MediaEval Development Sets and Test Sets include material subject to Creative Commons license and are freely available for download at http://mmlab.disi.unitn.it/MediaEvalSEM2014 and http://mmlab.disi.unitn.it/MediaEvalSEM2015.

TABLE IV: Results of using the different DCNN-based features (discussed in Section III-C) and the different links identification approaches (discussed in Section III-D).

GoogLeNet Layer Name	Links identif. approach	*Vancouver*		*London*		*NAMM15*		*TDF14*		Average and standard deviation across all datasets		
		P(%)	A(%)	P(%)	A(%)	P(%)	A(%)	P(%)	A(%)	P(%)	A(%)	H(%)
loss3/ classifier	*coverage*	26.5	94.0	36.1	81.1	72.2	67.5	25.0	75.0	40.0±22.1	79.4±11.2	49.7±14.4
	exact	5.9	56.9	44.4	64.3	77.8	76.9	43.8	45.2	43.0±29.4	60.8±13.3	46.3±27.6
conv2/ norm2	*coverage*	91.2	82.8	25.0	77.1	**83.3**	92.8	62.5	84.1	65.5±29.6	84.2±6.5	71.0±23.4
	exact	76.5	63.4	**77.8**	78.8	**83.3**	91.6	**65.6**	72.9	75.8±7.4	76.7±11.8	76.0±8.7
inception3a	*coverage*	94.1	90.5	41.7	79.5	**83.3**	90.8	**65.6**	81.5	71.2±22.9	85.6±5.9	76.6±16.8
	exact	**97.1**	83.7	75.0	84.3	**83.3**	93.4	**65.6**	73.7	**80.3**±13.3	83.8±8.0	**81.7**±9.4
inception4a	*coverage*	94.1	86.7	30.6	79.3	**83.3**	91.0	**65.6**	75.4	68.4±27.8	83.1±7.0	72.9±21.1
	exact	85.3	82.1	38.9	85.2	**83.3**	85.3	**65.6**	70.7	68.3±21.5	80.8±6.9	72.4±14.7
inception5a	*coverage*	**97.1**	**96.0**	36.1	77.1	77.8	90.7	50.0	78.1	65.2±27.4	85.5±9.3	72.6±21.5
	exact	73.5	76.7	41.7	84.3	**83.3**	92.8	56.3	**92.8**	63.7±18.5	86.6±7.8	72.2±13.3
Fusion approach	*coverage*	**97.1**	88.4	25.0	**95.0**	**83.3**	**96.9**	**65.6**	75.7	67.8±31.3	**89.0**±9.6	73.0±24.4
	exact	91.2	79.2	75.0	84.6	77.8	87.5	**65.6**	79.2	77.4±10.6	82.6±4.1	79.6±5.6

TABLE V: Comparison between the proposed method and the methods of [24], [31], [34], [33], [36], [38], [35], [37].

Method	*Vancouver*		*London*		*NAMM15*		*TDF14*		Average and standard deviation across all datasets		
	P(%)	A(%)	P(%)	A(%)	P(%)	A(%)	P(%)	A(%)	P(%)	A(%)	H(%)
[24]	61.8	81.6	33.3	88.6	50.0*	73.5*	25.0	80.6	43.9±17.6	81.1±6.2	55.5±15.9
[31]	94.1	79.2	47.2	87.5	-	-	-	-	70.7±33.2	83.4± 5.9	73.6±17.2
[34]	5.0	65.0	15.0	**92.0**	-	-	-	-	10.0± 7.1	78.5±19.1	17.6±11.6
[33]	91.2	72.8	61.1	71.3	77.9*	90.4*	9.4	79.3	61.4±37.0	78.5± 8.7	62.7±31.8
[35]	-	-	-	-	40.0	78.0	5.0	**92.0**	22.5±24.5	85.0± 9.9	31.2±30.6
[37]	94.1	76.0	61.1	65.6	83.3	90.8	12.5	84.5	62.8±36.2	79.2±10.9	64.0±30.1
[36]	35.0	86.0	25.0	89.0	44.5*	82.4*	15.6	83.2	30.0±12.5	**85.2**± 3.0	43.2±13.7
[38]	**97.1**	86.0	63.9	75.0	50.0*	71.5*	21.9	90.8	59.6±30.9	80.8± 9.1	64.5±23.0
[36] modified (inception3a)	5.9	**93.4**	5.6	73.5	47.1	85.0	9.4	73.2	17.0±20.1	81.3±9.8	24.7±24.1
[38] modified (inception3a,*exact*)	85.3	56.3	47.2	74.6	**88.9***	88.7*	**65.6**	84.6	73.1±20.9	76.1±14.4	72.7±14.1
[38] and [36]	8.8	54.7	19.4	67.4	58.8	66.8	25.0	80.6	28.0±21.6	67.4±10.6	36.5±19.8
Proposed (inception3a,*exact*,MRF)	**97.1**	83.7	**75.0**	84.3	83.3*	**93.4***	**65.6**	73.7	**80.3**±13.3	83.8±8.0	**81.7**±9.4

* Note: since the 19-th gallery in the *NAMM* dataset contains only videos, and for simplicity in this work we deal only with images included in the dataset, we regard this gallery as non-correctly synchronized in all the experiments we conducted.

original MRF-based method of [36] in order to estimate the temporal offsets. For the literature works that reported results of different variants of the algorithm, we include in Table V only the best results. As already mentioned, we strictly followed the experimental setup of the MediaEval SEM task and used the provided datasets and ground truth annotations to compare directly with the works of [31], [33], [34], [35], [37]. To compare with [24], we re-implemented the method and tested it under the same experimental setup. For methods [31], [34], [35] we report the published results; thus, for [31], [34] we report the results for the *Vancouver* and *London* datasets, while for [35], we report the results for the *NAMM15* and *TDF14* datasets. The best results for each column in this table are shown in bold.

According to these results, the proposed method achieves the best precision in almost all datasets and outperforms on average all other approaches in the literature ([24], [31], [33], [34], [35], [37]). Regarding accuracy, [34] scored the best accuracy in the *London* dataset but managed to synchronize a very small portion of the galleries. The same holds for the method of [35] and the *TDF14* dataset. Jointly considering precision and accuracy, by calculating the harmonic mean measure across all datasets (last column of Table V), reveals that the proposed method performs the best in comparison to all literature works, by a large margin.

Concerning [36], the proposed method significantly outperforms it, primarily due to [36] managing to synchronize only a small number of galleries (precision ranging from 15.6%

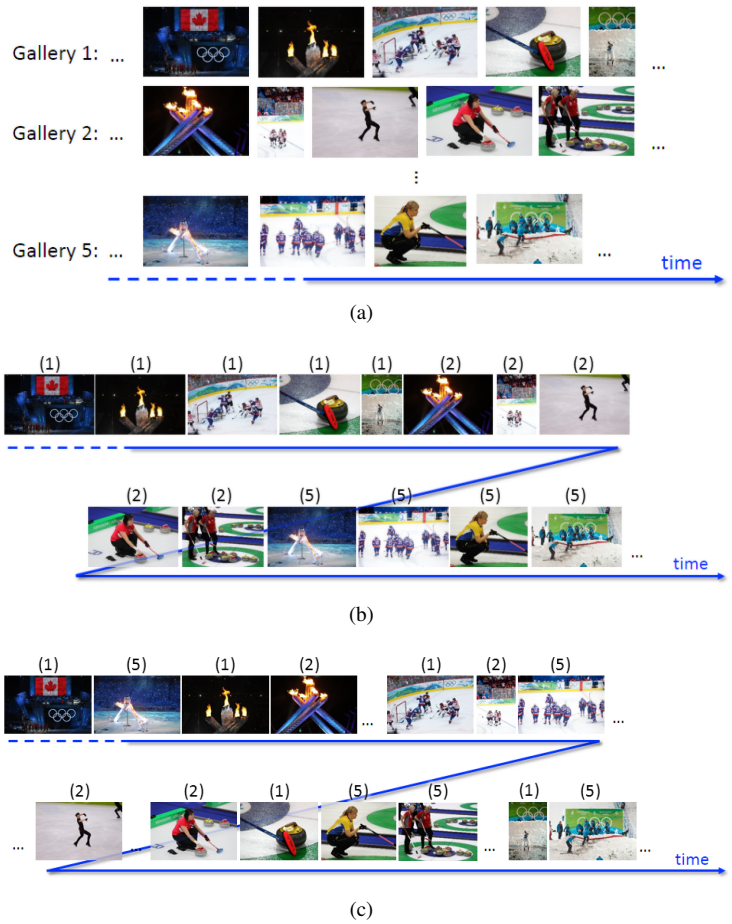

Fig. 9: Qualitative example of synchronization for a small subset of photos from the *Vancouver* dataset. (a) Original photo galleries. (b) Temporal order of the photos according to the original (noisy) timestamps. (c) Temporal order of the same photos using the timestamps estimated by the proposed method; these corrected timestamps make straightforward the meaningful ordering of the photos (opening ceremony, first ice hockey game, etc.) based on time information. In (b) and (c), the number above each photo indicates its gallery membership.

to 44.5% for the different datasets). This continues to be the case when [36] is modified so as to use the DCNN-based features discussed in Section III-C instead of the original ones; following this modification, the accuracy improves in two out of the four datasets but precision drops considerably for three datasets. The reason for this is that in [36] no pre-filtering is foreseen and the features are directly used in the graphical model together with location and time. Instead, the adoption of the new features leads to considerably improved results when combined with the graph construction and navigation approach proposed in the present work. Similarly [38] also performs worse than the proposed method (with average $H = 64.5\%$ versus $H = 81.7\%$ for the proposed). The modified version of [38], that also uses the powerful DCNN-based features instead of the original ones, performs slightly worse on two datasets and significantly better on the third and fourth ones (NAMM15 and TDF14), compared to the original [38]. Overall, considering the average harmonic mean across all datasets, the performance of the modified [38] is better ($H = 64.5\%$ versus $H = 72.7\%$). The fluctuations in performance across datasets are attributed to the reliance of [38] and modified [38] methods to a multitude of thresholds, which cannot be tuned to a single set of values that is optimal for all datasets. These two comparisons highlight the fact that simply introducing the DCNN features in a synchronization method is not enough to obtain satisfactory results. The excellent performance of the proposed method is due in part to these features (with the harmonic mean value increasing from 64.5% for the original method of [38], to 72.7% for the variant of it using our new features), and to an even greater extent to the MRF-based temporal offset estimation (proposed in Section III-E) which exploits the information extracted from previous stages (with the average harmonic mean further increasing from 72.7% to 81.7%).

The straightforward combination of our preliminary methods [38] and [36] performs poorly, especially in terms of precision, which means that only few galleries are correctly synchronized. This is mainly due to the fact that [36] relies on the assumption that all galleries must have a considerable overlap with the reference gallery (i.e., there should be several links from the reference to all other galleries). It is clear that when straightforwardly combining [38] and [36], there is no

guarantee that a reasonably high number of links (or even at least one link) will be obtained for each pair of galleries. Due to this lack of links, the offset error propagated along the graph is higher than using our proposed approach, and the number of galleries correctly synchronized is very low.

A qualitative example of the temporal synchronization achieved by the proposed method is shown, for a small subset of photos from the *Vancouver* dataset, in Fig. 9.

D. Computational Complexity Concerns

In this subsection, we derive the time and storage complexity required by our algorithm. To simplify the analysis without loosing generality, we assume that all galleries have the same number of photos, namely $\bar{N}$.[3] Therefore, the total number of photos is $n = k\bar{N}$.

For simplicity, in the subsequent tables and discussion we use the following notation for the different stages of the proposed method's pipeline:

1) **FE**: Feature extraction (Section III-C)
2) **SIM**: Construction of the visual similarity matrix of all images (Section III-C)
3) **GL**: Identification of the most similar photos between different galleries (Section III-C)
4) **GC**: Construction of the galleries graph (Section III-D)
5) **MST**: Finding of the MST in the galleries graph (Section III-D)
6) **MRF**: Calculation of the temporal offsets using MRF (Section III-E)

In Table VI we report the time needed (in seconds) for each of the aforementioned stages, on a Windows 7 64-bit machine with Intel i7 processor and 16GB of RAM.

TABLE VI: Empirical time measurements (in seconds) of all stages of the proposed method.

	Vancouver	*London*	*NAMM15*	*TDF14*
FE	1330.8	1890.2	387.3	2162.5
SIM	2.4	5.7	0.2	7.9
GL	0.2	0.2	0.1	0.4
GC	< 0.1	< 0.1	< 0.1	< 0.1
MST	< 0.1	< 0.1	< 0.1	< 0.1
MRF	6.1	9.3	2.5	6.2

Concerning time and space complexity (Table VII), it is easy to derive the worst case for the **FE**, **SIM**, **GL** and **GC** stages. Regarding the computation of the MST of the galleries graph (**MST** stage), we consider that: if E is the number of edges and V is the number of vertices in a graph, the Kruskal algorithm, which we employ is known to run in $O(E \log V)$ time. Since, in our case $V = k$ and $E \leq k^2$ the time complexity is $O(\frac{k^2}{\log k})$. The storage complexity of this stage is $O(k)$, since in the worst case one needs to hold all vertices (galleries) in the queue. In the **MRF** stage, inference is performed through the max-sum algorithm described in Section III-E3. Also in this case it is worth noting that the number of possible states/offsets for $\mathbf{x}$ is at most $\eta = \bar{N}^2$, but this number tends to be smaller when good matches across galleries are found. Similarly to [51], one can easily derive the time and storage complexities of the max-sum algorithm for our graphical model, namely $O(\eta(\bar{N} - 1))$ and $O(\eta)$, respectively.

[3] $\bar{N}$ can be regarded as the average number of photos per gallery, if one wants to relax this assumption.

TABLE VII: Time and storage complexity of all stages of the proposed method with respect to the number of galleries (k), the number of photos in each gallery ($\bar{N}$), and the total number of photos in the collection (n).

	Time Complexity	Storage Complexity
FE	$O(n)$	$O(n)$
SIM	$O(n^2)$	$O(\alpha n^2)$
GL	$O(k^2)$	$O(k^2)$
GC	$O(k^2)$	$O(k^2)$
MST	$O(\frac{k^2}{\log k})$	$O(k)$
MRF	$O(k\bar{N}^3)$	$O(k\bar{N}^2)$

We notice from the data reported in Table VII that the **SIM** stage exhibits the highest complexity if the number of users is large. Especially if the number of photos increases significantly the pairwise photo comparisons stage could consume considerable computational resource. However, the complexity of this stage can be reduced by using known data indexing schemes, such as Locality Sensitive Hashing and KD-trees, to avoid computing the similarities for all n^2 possible pairs of photos. The **MRF** stage is more computationally expensive in case of very large photo collections. Nevertheless, it is clear from our empirical time measurements (Table VI), that even stages with relatively high complexity have very short running times in practice.

V. Conclusions

In this paper we presented a method addressing the problem of synchronizing multiple photo galleries. The proposed approach exploits Deep Convolutional Neural Network features to measure photo similarity; subsequently, similar photos are used to define a graph of galleries, and finally each pair of galleries in the graph is synchronized using a probabilistic graphical model. Extensive experiments on four benchmark datasets documented the merit of the proposed method, which achieves very good synchronization precision and accuracy and outperforms the state of the art.

Acknowledgments

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement H2020-687786 InVID.

References

[1] C. Tzelepis, Z. Ma, V. Mezaris, B. Ionescu, I. Kompatsiaris, G. Boato, N. Sebe, and S. Yan, "Event-based media processing and analysis: A survey of the literature," *Image and Vision Computing Journal*, vol. 53, pp. 3–19, September 2016.

[2] J.-T. Tsai, Y.-Y. Lin, and H.-Y. M. Liao, "Per-cluster ensemble kernel learning for multi-modal image clustering with group-dependent feature selection," *IEEE Trans. on Multimedia*, vol. 16, no. 8, pp. 2229–2241, 2014.

[3] M. L. Kherfi and D. Ziou, "Image collection organization and its application to indexing, browsing, summarization, and semantic retrieval," *IEEE Trans. on Multimedia*, vol. 9, no. 4, pp. 893–900, 2007.

[4] S. Rudinac, M. Larson, and A. Hanjalic, "Learning crowd sourced user preferences for visual summarization of image collections," *IEEE Trans. on Multimedia*, vol. 15, no. 6, pp. 1231–1243, 2013.

[5] R. Wang, S. J. McKenna, J. Han, A. Ward, et al., "Visualizing image collections using high-entropy layout distributions," *IEEE Trans. on Multimedia*, vol. 12, no. 8, pp. 803–813, 2010.

[6] J. Yuan, J. Luo, and Y. Wu, "Mining compositional features from gps and visual cues for event recognition in photo collections," *IEEE Trans. on Multimedia*, vol. 12, no. 7, pp. 705–716, 2010.

[7] M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, and T.-S. Chua, "Event driven web video summarization by tag localization and key-shot identification," *IEEE Trans. on Multimedia*, vol. 14, no. 4, pp. 975–985, 2012.

[8] A. Rosani, G. Boato, and F. G.B. De Natale, "Eventmask: A game-based framework for event-saliency identification in images," *IEEE Trans. on Multimedia*, vol. 17, no. 8, pp. 1359–1371, Aug. 2015.

[9] N. O'Hare and A. F. Smeaton, "Context-aware person identification in personal photo collections," *IEEE Trans. on Multimedia*, vol. 11, no. 2, pp. 220–228, 2009.

[10] J. Y. Choi, W. De Neve, K. N. Plataniotis, and Y. M. Ro, "Collaborative face recognition for improved face annotation in personal photo collections shared on online social networks," *IEEE Trans. on Multimedia*, vol. 13, no. 1, pp. 14–28, 2011.

[11] J. Sang, C. Xu, and J. Liu, "User-aware image tag refinement via ternary semantic analysis," *IEEE Trans. on Multimedia*, vol. 14, no. 3, pp. 883–895, 2012.

[12] S. Liu, S. Yan, T. Zhang, C. Xu, J. Liu, and H. Lu, "Weakly supervised graph propagation towards collective image parsing," *IEEE Trans. on Multimedia*, vol. 14, no. 2, pp. 361–373, 2012.

[13] A. Graham, H. Garcia-Molina, A. Paepcke, and T. Winograd, "Time as essence for photo browsing through personal digital libraries," in *Proc. of the 2nd ACM/IEEE-CS joint Conf. on Digital libraries*. ACM, 2002, pp. 326–335.

[14] P. Mulhem and J.-H. Lim, "Home photo retrieval: time matters," in *Image and Video Retrieval*, pp. 321–330. Springer, 2003.

[15] P. Sinha, H. Pirsiavash, and R. Jain, "Personal photo album summarization," in *Proc. of the 17th Int. Conf. on Multimedia*. ACM, 2009, pp. 1131–1132.

[16] M. Rabbath, P. Sandhaus, and S. Boll, "Automatic creation of photo books from stories in social media," *Trans. on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 7, no. 1, pp. 27, 2011.

[17] A. Pigeau and M. Gelgon, "Spatio-temporal organization of one's personal image collection with model-based icl-clustering," in *3rd Int. Workshop on Content-Based Multimedia Indexing (CBMI)*, 2003, pp. 111–118.

[18] G. Kim and E. Xing, "Reconstructing storyline graphs for image recommendation from web community photos," in *Proc. of the Conf. on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 3882–3889.

[19] G. Kim, L. Sigal, and E. Xing, "Joint summarization of large-scale collections of web images and videos for storyline reconstruction," in *Proc. of the Conf. on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 4225–4232.

[20] G. Kim, S. Moon, and L. Sigal, "Joint photo stream and blog post summarization and exploration," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 3081–3089.

[21] Y. Moses, S. Avidan, et al., "Space-time tradeoffs in photo sequencing," in *Proc. of the Int. Conf. on Computer Vision*. IEEE, 2013, pp. 977–984.

[22] N. Suematsu and A. Hayashi, "Time series alignment with gaussian processes," in *21st Int. Conf. on Pattern Recognition (ICPR)*. IEEE, 2012, pp. 2355–2358.

[23] G. Kim and E. Xing, "Jointly aligning and segmenting multiple web photo streams for the inference of collective photo storylines," in *Proc. of the Conf. on Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 620–627.

[24] M. Broilo, G. Boato, and F. G.B. De Natale, "Content-Based Synchronization for Multiple Photos Galleries," in *19th Int. Conf. on Image Processing (ICIP)*. IEEE, Sept. 2012, pp. 1945–1948.

[25] J. Yang, J. Luo, J. Yu, and T. S. Huang, "Photo stream alignment and summarization for collaborative photo collection and sharing," *IEEE Trans. on Multimedia*, vol. 14, no. 6, pp. 1642–1651, 2012.

[26] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. Journ. of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

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

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

[29] N. Conci, F. G.B. De Natale, and V. Mezaris, "Synchronization of Multi-User Event Media (SEM) at MediaEval 2014: Task Description, Datasets, and Evaluation," in *Proc. MediaEval 2014*, Barcelona, Spain, Oct. 2014.

[30] N. Conci, F. G.B. 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 2015*, Wurzen, Germany, Sept. 2015.

[31] M. Zaharieva, M. Riegler, and M. Del Fabro, "Multimodal synchronization of image galleries," in *Proc. MediaEval 2014*, Oct. 2014, vol. 1263.

[32] D. S. Messing, P. Van Beek, and J. H. Errico, "The MPEG-7 colour structure descriptor: Image description using colour and local spatial information," in *Proc. of Int. Conf. on Image Processing*. IEEE, 2001, vol. 1, pp. 670–673.

[33] K. Apostolidis, C. Papagiannopoulou, and V. Mezaris, "CERTH at mediaeval 2014 synchronization of multi-user event media task," in *Proc. MediaEval 2014*, Barcelona, Spain, Oct. 2014, vol. 1263.

[34] P. Nowak, M. Thaler, H. Stiegler, and W. Bailer, "JRS at event synchronization task," in *Proc. MediaEval 2014*, Barcelona, Spain, Oct. 2014, vol. 1263.

[35] H. Fassold, H. Stiegler, F. Lee, and W. Bailer, "Jrs at synchronization of multi-user event media task," in *Proc. MediaEval 2015*, Wurzen, Germany, Sept. 2015.

[36] E. Sansone, G. Boato, and M.-S. Dao, "Synchronizing multi-user photo galleries with MRF," in *Proc. MediaEval 2014*, Barcelona, Spain, Oct. 2014, vol. 1263.

[37] K. Apostolidis and V. Mezaris, "CERTH at mediaeval 2015 synchronization of multi-user event media task," in *Proc. MediaEval 2015*, Wurzen, Germany, Sept. 2015.

[38] 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.

[39] X. Duan, L. Lin, and H. Chao, "Discovering video shot categories by unsupervised stochastic graph partition," *IEEE Trans. on Multimedia*, vol. 15, no. 1, pp. 167–180, Jan 2013.

[40] L. Lin, R. Zhang, and X. Duan, "Adaptive scene category discovery with generative learning and compositional sampling," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 25, no. 2, pp. 251–260, Feb 2015.

[41] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015.

[42] J. Ng, F. Yang, and L. Davis, "Exploiting local features from deep networks for image retrieval," in *Proc. of the Conf. on Computer Vision and Pattern Recognition Workshops*. IEEE, 2015, pp. 53–61.

[43] L. Xie, R. Hong, B. Zhang, and Q. Tian, "Image classification and retrieval are one," in *Proc. of the 5th Int. Conf. on Multimedia Retrieval*. ACM, 2015, pp. 3–10.

[44] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. on Image Processing*, vol. 24, no. 12, pp. 4766–4779, Dec 2015.

[45] D. Song, W. Liu, D. A. Meyer, D. Tao, and R. Ji, "Rank preserving hashing for rapid image search," in *2015 Data Compression Conference*. IEEE, 2015, pp. 353–362.

[46] D. Song, W. Liu, R. Ji, D. A. Meyer, and J. R. Smith, "Top rank supervised binary coding for visual search," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1922–1930.

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

[48] 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.
[49] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3304–3311.
[50] R. Arandjelovic and A. Zisserman, "All about VLAD," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2013, pp. 1578–1585.
[51] C. M. Bishop, *Pattern recognition and machine learning*, vol. 1, Springer New York, 2006.

Emanuele Sansone received the BSc, MSc in Telecommunications Engineering from University of Trento, Italy, in 2010 and 2013, respectively. He joined the Department of Information Engineering and Computer Science at University of Trento as a PhD student in 2013, and he is currently pursuing the PhD degree. His research interests are in machine learning, especially in semi-supervised learning, positive-unlabeled learning and deep learning.

Konstantinos Apostolidis received the BSc in Applied Informatics from the University of Macedonia in 2010, and the MSc in Digital Media from the Aristotle University of Thessaloniki in 2012. He is a Researcher Assistant at the Information Technologies Institute (ITI) of the Centre for Research of Technology Hellas (CERTH). His research interests include content-based and semantic image retrieval and machine learning for multimedia analysis.

Nicola Conci (S'04, M'08) received the bachelor and master degree in telecommunication engineering from the University of Trento (Italy) in 2002 and 2004 respectively. From the same University he received the Ph.D in 2007. He was visiting student at the Image Processing Lab. at University of California Santa Barbara and post-doc researcher in the Multimedia and Vision research group at Queen Mary University of London. Dr. Conci has authored and co-authored more than 70 scientific papers in international journals and conferences. His current research interests are in the area of video analysis for human behavior analysis with particular application to environmental monitoring, surveillance, and assisted living. He is co-founder of Xtensa srl.

Giulia Boato is Assistant Professor at the Department of Information Engineering and Computer Science (DISI) of the University of Trento (Italy). She was in the Project staff of many projects (FP7 FET-IP LIVINGKNOWLEDGE, FP7 IP GLOCAL, FP7 CA ETERNALS, ICT-COST 3D-ConTourNet). She was co-chair of the International Workshop Living Web: making diversity a true asset (Washington DC, October 2009) within the International Semantic Web Conference 2009 and of the workshop on Event-based Media Integration and Processing co-located with ACM Multimedia conference 2013. She is reviewer for many international journals, e.g., IEEE Transactions on Information Forensics and Security, IEEE Transactions on Signal Processing, IEEE Transactions on Multimedia, IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology. She is in the editorial board of the Elsevier Image and Vision Computing Special Issue on Event-based Media Processing and Analysis. Her research interests are focused on image and signal processing, with particular attention to multimedia data protection, data hiding and digital forensics, but also intelligent multidimensional data management and analysis. She is author of 85 papers in international conferences and journals. She is member of the IEEE Multimedia Signal Processing Technical Committee (MMSP TC).

Vasileios Mezaris received the BSc and PhD in Electrical and Computer Engineering from the Aristotle University of Thessaloniki in 2001 and 2005, respectively. He is a Senior Researcher (Researcher B) at the Information Technologies Institute (ITI) of the Centre for Research of Technology Hellas (CERTH). His research interests include image and video analysis, event detection in multimedia, machine learning for multimedia analysis, content-based and semantic image and video retrieval, applications of image and video analysis in specific domains (TV broadcasting and News, medical images, ecological data, educational and cultural applications). He has co-authored more than 30 papers in refereed journals, 10 book chapters, 130 papers in international conferences, and 3 patents. He served as Associate Editor for the IEEE Transactions on Multimedia (2012-2015), serves as Associate Editor for the IEEE Signal Processing Letters, and is a Senior Member of the IEEE.

Francesco De Natale (M.Sc. '90, Ph.D. '94) is a Professor of Telecommunications at the University of Trento, Italy, where he leads the Multimedia Lab (mmlab.disi.unitn.it). His research interests are focused on multimedia communications, with particular attention to multimedia signal processing, analysis, and retrieval. He was Program Co-Chair of the IEEE Intl. Conf. on Image Processing (ICIP-2005) and General Chair of the ACM Intl. Conf. on Multimedia Retrieval (ICMR-2011). He has been Associate Editor of the IEEE Trans on Multimedia and of the IEEE Trans. on Circuits and Systems for Video Technologies, as well as a member of the IEEE Signal Proc. Society Technical Committee on Multimedia Signal Processing (MMSP), chairing the Technical Directions Subcommittee. Currently, he is member of the Board of Directors of the Italian Consortium for Telecommunications (CNIT), and member of the management of the Italian Group of Telecommunications and Information Technologies (GTTI). He is also co-founder of the university startup Xtensa, a company that develops tools in the area of user interaction and ambient intelligence. He published more than 150 papers in international journals and conferences, mostly in the area of multimedia signal processing and communications, and has been scientific coordinator of many large-scale research and development projects, both at the national and international level. Prof. De Natale was appointed evaluator for several international bodies, including the European Commission and the NSFs of US and Ireland. Prof. De Natale is a Senior Member of IEEE and a member of ACM and GIRPR.