Synchronizing Multi-User Photo Galleries with MRF

Emanuele Sansone, Giulia Boato
DISI, University of Trento
Trento, Italy
e.sansone@unitn.it, boato@disi.unitn.it

Minh-Son Dao
UIT-HCM
HCMC, Viet-Nam
sondm@uit.edu.vn

ABSTRACT

We present a novel solution to the MediaEval 2014 Event Synchronization Task: Synchronization of Multi-User Event Media (SEM). The framework is based on a probabilistic graphical model. Thanks to the simple topology of the graph, the estimation of the true temporal displacement among multiple photo collections can be performed efficiently through exact inference. The underlying fitness function is defined in a flexible way, for which it is possible to integrate easily new information (e.g., text tags or social network data). The flexibility makes the framework suitable and adaptable to cope with many real situations. The method is evaluated on two datasets obtaining an overall accuracy of more than 85% in both cases.

1. INTRODUCTION

The problem of photo stream synchronization has been investigated in little work in the current literature. Nevertheless, it represents an open and attractive research topic, especially if one considers its potential applicability to the context of online photo sharing communities.

Indeed, a novel task on this issue has been introduced in MediaEval 2014 [4], where the scenario considered is represented by a number of users attending the same event and taking photos and videos with different non-synchronized devices. The goal of the task is twofold.

The *synchronization* consists of finding the correct temporal offset of each photo collection, denoted as $P^1 = (p_1^1, p_2^1, \ldots, p_N^1)$, with respect to a reference gallery, namely $P^2 = (p_1^2, p_2^2, \ldots, p_M^2)$, where N and M correspond to the lengths of the two streams.

Once the sequential chronological order of all pictures is restored, the *clustering* phase is evaluated based on the number of sub-events detected as well as on the quality of the obtained groupings.

2. PROPOSED APPROACH

The proposed framework for synchronization is based on a probabilistic graphical model. Each temporal displacement can be uniquely identified by a set of nearest-neighbor picture pairs across the two photo sets. Fig. 1 shows an example of this concept, where each image in the collection to be synchronized can be compared with the nearest image in the reference gallery (notice that N will be often smaller than M but this is not always true).

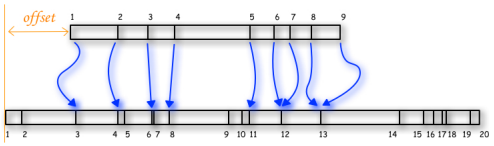

Figure 1: Example of correspondences between two photo collections. The two stripes represent the two photo collections, where the lower is the reference and the upper is the gallery to be synchronized; vertical lines correspond to pictures and arrows depict correspondences.

Given the set of all possible temporal displacements, namely $\{\Delta T_i : i = 1, \ldots, Q\}$, and the sequence of all correspondences between pictures given the offset ΔT_i, namely $\mathbf{x}^{\Delta T_i} = (x_1^{\Delta T_i}, \ldots, x_j^{\Delta T_i}, \ldots, x_N^{\Delta T_i})$, where $x_j^{\Delta T_i}$ identifies the picture in the reference P^2 associated with image p_j^1 given ΔT_i, the synchronization task can be cast into an optimization problem for finding the best offset ΔT^*. In other words,

$$\Delta T^* = \arg\max_{\Delta T_i} f(\mathbf{x}^{\Delta T_i}) \tag{1}$$

where $f : X \to \mathbb{R}$ is the function that associates a similarity score to each sequence of associations and $X = \{\mathbf{x}^{\Delta T_i} : i = 1, \ldots, Q\}$.

Now it is possible to define an undirected graphical model through a sequence of observed nodes $\mathbf{y} = (y_1, \ldots, y_j, \ldots, y_N)$, where y_j refers to the image p_j^1 in P^1, and a sequence of latent variables $\mathbf{x} = (x_1, \ldots, x_j, \ldots, x_N)$, whose admissible values are defined over the set X. The edges of the model are of two kinds: links between nodes in $\mathbf{x}$ and $\mathbf{y}$, that compare the similarity across photos of the two galleries, and links between nodes in the same sequence $\mathbf{x}$, which take into account the temporal structure of the collections. Fig. 2 summarizes the graphical model described so far.

The joint distribution associated with the graph can therefore be factorized in the following form:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_j \psi(x_j, x_{j+1}) \prod_k \phi(x_k, y_k) \tag{2}$$

where $\psi(x_j, x_{j+1})$ is the potential associated with the link that connects x_j and x_{j+1} and $\phi(x_k, y_k)$ corresponds to the potential of the edge between x_k and y_k [3]. The distribution $p(\mathbf{x})$ can be interpreted as a function measuring the quality of the alignment given an offset, and can be exploited as

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

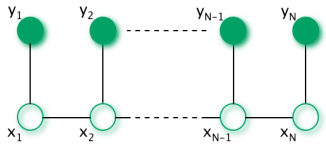

Figure 2: Markov network with observed and hidden nodes.

the objective f in Eq. (1). For the sake of computational simplicity, we define the potential functions belonging to the expontential family, namely:

$$\phi(x_k, y_k) = \exp\Big\{ -\Big[\alpha \frac{D_H(x_k, y_k)}{D_H^{max}(k)} + \\ + \beta \frac{D_S(x_k, y_k)}{D_S^{max}(k)} + \gamma \frac{D_G(x_k, y_k)}{D_G^{max}(k)}\Big]\Big\} \quad (3)$$

$$\psi(x_j, x_{j+1}) = \exp\Big\{ -\delta \frac{D_T(x_j, x_{j+1})}{D_T^{max}(j)}\Big\} \quad (4)$$

where D_H, D_S and D_G represent distance metrics between images computed on HSV color histograms, SURF descriptors [2] and GPS coordinates, respectively. In particular, D_H is obtained by first dividing each image into 9 blocks, by computing the Hellinger distance between color histograms extracted from their respective blocks and by combining the distances linearly assigning a higher weight to the central component. D_S corresponds to the average of Euclidean distances evaluated over all pairs of matched salient points. Finally, D_G is computed by approximating the Earth surface as a sphere. D_T is evaluated on timestamp information according to the following relation:

$$D_T(x_j, x_{j+1}) = |t_{y_{j+1}} - t_{x_{j+1}}| + |t_{y_j} - t_{x_j}| \quad (5)$$

Every distance measure is therefore normalized by its respective maximum value, and is finally combined linearly as shown in Eqs. (3) and (4).

As far as the clustering challenge is concerned, the k-means algorithm is used to find the natural grouping of images, see Section 3.

3. RESULTS AND DISCUSSION

The learning of the parameters described in the previous section is carried out by performing an optimization of the joint distribution over the parameter solution space and using the training dataset made available by [4]. The estimated values for the parameters are $\alpha = 2.4249, \beta = 0.9509, \gamma = 0.9594$ and $\delta = 3.8597$. Once the training phase has completed, one can start to synchronize any pair of galleries.

Results obtained in the SEM task for synchronization are summarized in Table 1. In both datasets, the accuracy of synchronization is greater than 85%, which proves the effectiveness of the proposed algorithm. Nevertheless, the precision obtained is quite poor since on average only one fourth of the photo collections are correctly synchronized. The low performance is mainly due to forcing associations between images of the two galleries. In many cases, there are pictures that have no correspondence in the reference set.

Once the galleries are synchronized, the corrected temporal information can be exploited to perform clustering. At this purpose, the k-means algorithm is used over three different combinations of features. The first configuration consists of the concatenation of Global Color Structure Descriptors (CSD) [1] with Local Binary Patterns [5]. The second configuration consists only of 6 CSD values obtained by performing PCA reduction on the original CSD descriptor, while the last set up is obtained by adding the temporal information to the second configuration. Table 2 shows the results associated with the three different runs. In general, the inclusion of the temporal feature doesn't significantly increase the performance. But it's evident that if one is able to carry out a precise synchronization, then temporal information becomes a very reliable feature to perform clustering. This is confirmed by the fact that the results are slightly more than 10% in terms of F-measure and low-level visual features are therefore not sufficient.

Future work will be devoted to increasing the synchronization precision in order to allow the exploitation of the temporal component. One possible approach consists of modifying the structure of the graphical model, such that new binary latent variables take into account the possibility of having no associations between photos.

Table 1: Synchronization performance on the two datasets in terms of precision and accuracy.

Dataset	Precision	Accuracy
Vancouver	0.35	0.86
London	0.25	0.89

Table 2: Clustering performance on the two datasets in terms of Rand Index, Jaccard Index and F-measure.

Dataset	Run	RI	JI	F1
Vancouver	1	0.9749	0.1673	0.1433
	2	0.9737	0.1382	0.1214
	3	0.9730	0.1315	0.1162
London	1	0.9852	0.1287	0.1140
	2	0.9836	0.0742	0.0691
	3	0.9841	0.0885	0.0813

4. REFERENCES

[1] M. Baştan, H. Çam, U. Güdükbay, and Özgür Ulusoy. BilVideo-7: An MPEG-7-Compatible Video Indexing and Retrieval System. *IEEE MultiMedia*, 17(3):62–73, 2009.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded Up Robust Features. In *Computer Vision - ECCV 2006*.

[3] C. M. Bishop et al. *Pattern Recognition and Machine Learning*. Springer New York, 2006.

[4] N. Conci, F. D. Natale, and V. Mezaris. Synchronization of Multi-User Event Media (SEM) at MediaEval 2014: Task Description, Datasets, and Evaluation. In *Proceedings of MediaEval 2014*, Barcelona, Spain, October 2014.

[5] T. Ojala, M. Pietikäinen, and T. Mäenpää. Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns. In *Computer Vision - ECCV 2000*, volume 1842 of *Lecture Notes in Computer Science*, pages 404–420. Springer Berlin Heidelberg, 2000.