MULTIPLE-VIEW CONSTRAINED CLUSTERING FOR UNSUPERVISED FACE IDENTIFICATION IN TV-BROADCAST

Meriem Bendris[1], *Benoit Favre*[1], *Delphine Charlet*[2], *Géraldine Damnati*[2], *Rémi Auguste*[3]

[1]Aix Marseille Université, [2]OrangeLabs, [3]Université Lille 1
{firstname.lastname}@{[1]lif.univ-mrs.fr,[2]orange.com,[3]lifl.fr}

ABSTRACT

Our goal is to automatically identify faces in TV broadcast without a pre-defined dictionary of identities. Most methods are based on identity detection (from OCR and ASR) and require a propagation strategy based on visual clustering. In TV content, people appear with many variations making the clustering difficult. In this case, speaker clustering can be a reliable link for face clustering. We propose in this paper to build automatically an incomplete speaker-face mapping based on local evidence of OCR and Lip activity links. Then, we propose schemes of speaker constraints propagation to the face constrained-clustering problem. Experiments performed on the REPERE corpus show an improvement of face identification by propagating names to face clusters ($+3.7\%$ F-measure compared to the baseline).

1. INTRODUCTION

The proliferation of multimedia content has changed users consumption making them choosing the information they want to visualize. It became necessary to develop technologies that facilitate the navigation through these multimedia data. One key to efficient browsing is to locate sequences of a specific person. The TV-context introduces many ambiguities making biometric models unreliable. In addition, maintaining up-to-date large dictionaries of face models is prohibitively expensive. In this paper, we are interested in identifying people in TV broadcast without biometric models. To this end, most methods are based on identity detection (ASR [1], OCR [2], subtitles [3]) and propagation to face clusters [2, 3]. Unfortunately, when processing unconstrained videos, the appearance variability of faces make the clustering very difficult. Speaker clustering can be good indicator for face identification. For instance, if two faces with lip activity (presumably speaking) are associated to the same speaker, they can be associated in the visual modality as well. In this work, we want to use Multi-modal sources to improve visual clustering.

We would like to thank PERCOL consortium participants for providing their subsystem outputs. This work is funded by ANR under project PERCOL 2010-CORD-102-01.

Multi-modal clustering has received much attention during the past decade [4, 5]. However, most methods assume a complete bipartite mapping between modalities. In our study, faces and speakers can be directly bound on only a subset of frames and there are instances for which only the speaker or only the face is available. In [6], authors proposed a semi-supervised multiple-view clustering on which only a subset of objects has a multiple-view representation. This incomplete mapping is considered as a hard must-link constraint between views allowing only symmetric constraint propagation. In our work, we propose to widen the problem allowing cannot-link constraints between views in addition to the must-link ones. We present a new face clustering algorithm based on speaker constraint propagation: first, the Face⇔Speaker mapping is obtained automatically using lip activity and OCR propagation. Then, following this mapping, speaker constraints obtained from a clustering process are propagated to the face constrained-clustering. Finally, faces are identified by propagating names to the resulting face clusters.

Fig. 1. Multiple sources of identification - *REPERE* corpus.

This paper is organized as follows: Section 2 introduces the constrained clustering problem; Section 3 describes the automatic Face⇔Speaker mapping based on OCR input and Lip activity; Section 4 describes our method of face clustering based on speaker constraint propagation. Finally, Section 5 presents results on the *REPERE* corpus.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. CONSTRAINED CLUSTERING

Clustering consists in grouping similar objects into clusters. Given a collection of objects $X = \{x_1, x_2, ..., x_n\}$ and a distance function $d()$, [7] proposed a clustering formulation that jointly minimizes the intra-cluster distance and the number of clusters. This approach can be expressed as the following integer linear program (ILP):

$$Min \sum_i l_{j,j} + \frac{1}{F} \sum_{i,j} d(x_i, x_j) l_{i,j} \quad (1)$$

$$S.t. \sum_{i \neq j} l_{i,j} - l_{j,j} \geq 0 \quad \forall j \quad (2)$$

$$l_{j,j} - l_{i,j} \geq 0 \quad \forall i, j \quad (3)$$

$$l_{i,j} \in \{0, 1\} \quad \forall i, j \quad (4)$$

in which $l_{i,j}$ is a binary variable representing the fact that x_i is a member of cluster j, and $l_{j,j}$ is 1 when cluster j exists and 0 when it's empty. $\frac{1}{F}$ is a mixing factor for comparing intra-cluster distances with the number of clusters ($\sum_j l_{j,j}$). Constraint (2) ensures that a cluster is not active when it does not contain any member and constraint (3) activates a cluster when one of its member is active.

Constrained clustering [8] (or semi-supervised clustering) aims to cluster objects given constraints specifying the pairs of objects that need to be in the same cluster (must-link) or not (cannot-link). In the aforementioned ILP, a must-link constraint between x_i and x_j can be expressed as $l_{i,k} = l_{j,k} \forall k$ while a cannot-link would be $l_{i,k} + l_{j,k} \leq 1 \forall k$ (at most one of them can be 1). Let $C_=$ (resp. $C_\neq$) be the set of pairs of elements (x_i, x_j) subject to must-link constraints (resp. cannot-link constraints), then the following constraints can be added to the ILP:

$$l_{i,k} - l_{j,k} = 0 \quad \forall (i, j) \in C_= \quad (5)$$

$$l_{i,k} + l_{j,k} \leq 1 \quad \forall (i, j) \in C_\neq \quad (6)$$

Since cross-modality constraints might be spurious because of speaker clustering, OCR or lip activity detection errors, we choose to enforce soft constraints (can be violated with a penalty) instead of hard constraints. Soft constraints are implemented by introducing them in the objective function with Lagrange Multipliers. The ILP formulation can be extended as follows:

$$\begin{aligned} Min \sum_i l_{j,j} + \frac{1}{F} \sum_{i,j} d(x_i, x_j) l_{i,j} \\ - \lambda_1 \sum_{(i,j) \in C_=} \sum_k (l_{i,k} - l_{j,k}) \\ - \lambda_2 \sum_{(i,j) \in C_\neq} \sum_k (l_{i,k} + l_{j,k} - 1) \end{aligned} \quad (7)$$

$S.t.$ constraints (2), (3), (4)

where λ_1 and λ_2 are the costs associated to violating must-link and cannot-link constraints. Note that $C_=, C_\neq$ are disjoint. In our experiments, $F = \sum_{i \neq j}^{n} d(x_i, x_j)$; a maximum distance criterion $d(x_i, x_j) < \sigma$ reduces the size of the problem; and $\lambda_1 = \lambda_2 = 100$ (trained from the development set).

3. AUTOMATIC FACE ⇔ SPEAKER MAPPING

In this section, we describe how cross-modal link evidence is gathered using lip-activity and Overlaid Person Name detection. In the rest of the paper, $M^{F \Leftrightarrow S}$ refers to the Face ⇔ Speaker mapping.

3.1. Lip Activity Detection

Measuring the lip activity between a face track and speaker turn detected at the same time can be a good indicator of a cross-modal mapping. We measure lip activity as follows: the lower region of consecutive face detections is aligned, we then measure the entropy of the pixel movement direction (optical flow) on that region. Detailed experiments described in [9] show that the lip-activity detector is able to classify talking-faces and non talking-faces shots in TV-shows with an error rate of 20%. The average lip activity over the face track is used to generate must-link mapping when motion is highly disordered ($> \theta_1$) and cannot-link mapping when motion is organized ($< \theta_2$). The generated mapping is denoted $M^{F \Leftrightarrow S}_{Lip}$.

3.2. Overlaid Person Names

In TV broadcast, most of Overlaid Person Names (OPN) occur while the corresponding face appears talking. Statistics on the *REPERE* corpus presented in Table 1 corroborate this idea, showing that 98.5% of the annotated OPNs appear with the corresponding face. In the same way, 80.4% of the annotated OPNs appear with the corresponding speaker. Consequently in unambiguous shots where only one face and speaker is detected, we locally propagate the OPN to the face track and speaker. The generated mapping is denoted by $M^{F \Leftrightarrow S}_{OPN}$.

The OPN detection is performed as follows: first, text detection is achieved on each frame in a predefined area of interest (at the bottom) using a convolutional neural network [10]. Then, each text region is tracked on consecutive frames using bounding box overlap. We used two OCR systems: GOCR[1] and Tesseract[2]. Their resulting character sequence hypotheses for a given track are merged to form confusion networks from which the most probable sequence is extracted. A rule-based classifier (the number of words, vicinity of another box, etc) distinguishes Overlaid Person Names (OPN) from any

[1] http://www.jocr.sourceforge.net

[2] http://code.google.com/p/tesseract-ocr/

other text. Finally, the character sequence hypothesis associated to a detected OPN is submitted to a normalization module which consists in finding the most similar name in a large dictionary of person names.

4. MAPPING-BASED FACE CLUSTERING

4.1. Constraints propagation schemes

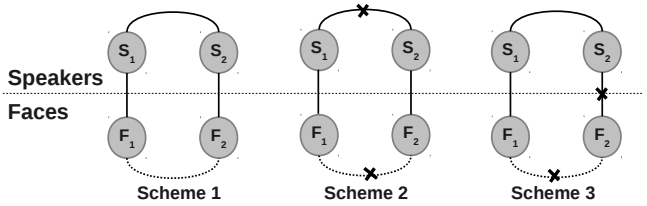

Fig. 2. multi-modal constraints propagation schemes in asynchronous data. Dashed lines represent generated constraints, solid lines represent observed evidence.

Our goal is to propagate a set of constraints from a modality to affect clustering in another modality considering multimodal objects as independent (because of asynchrony between modalities). In [6], the authors developed a method for clustering images and text appearing in the same document using a subset of hard must-link mapping between the views. Let F_1 and F_2 be two face tracks, and S_1 and S_2 be two speaker segments. The following symmetric propagation constraints can be established (see Figure 2):

- Scheme 1: must-link constraint $F_1 \leftrightarrow F_2$ if must-link evidence $S_1 \leftrightarrow F_1$, $S_2 \leftrightarrow F_2$ and $S_1 \leftrightarrow S_2$
- Scheme 2: cannot-link constraint $F_1 \not\leftrightarrow F_2$ if must-link evidence $S_1 \leftrightarrow F_1$ and $S_2 \leftrightarrow F_2$ and cannot-link evidence $S_1 \not\leftrightarrow S_2$

In our experiments, it is possible to detect cannot-link cross-modal mapping. Thus, a third propagation scheme appears to be possible:

- Scheme 3: cannot-link constraint $F_1 \not\leftrightarrow F_2$ if must-link evidence $S_1 \leftrightarrow F_1$ and $S_1 \leftrightarrow S_2$ and cannot-link evidence $S_2 \not\leftrightarrow F_2$.

4.2. Speaker Clustering

We used a speaker clustering system that follows the principles in [11]: First, agglomerative clustering of speech segments is performed based on Bayesian Information criterion. Then, that initial set of clusters is modelled with a GMMs in order to more accurately compare voices using a Cross-Likelihood Criterion (CLR) for another pass of agglomerative clustering. At each iteration, Viterbi decoding is performed to re-segment the speech data into speaker turns given the new clusters. The obtained clustering is called $\tilde{X}_{spk}$ and generate must-link constraints $S_1 \leftrightarrow S_2$ if S_1 and S_2 are in the same speaker cluster, and cannot-link constraints $S_1 \not\leftrightarrow S_2$ if S_1 and S_2 are in different clusters.

4.3. Face Clustering

Faces are detected using OpenCV's cascade classifier [12] for frontal and profile faces. The resulting detections are tracked until shot boundaries using bounding box overlap. Then, the upper body is detected using a background subtraction algorithm based on Grabcut [13], initialized with detected faces. The background subtraction algorithm yields a very accurate silhouette of the person, even in the presence of a dynamic background. Each extracted person is then modelled using a space-time color histogram. A distance matrix between face tracks is obtained using a combination of Bhattacharyya coefficient and Mahalanobis distance [14].

Algorithm 1 Clustering with multi-modal constraints propagation in asynchronous data

1. Generate unimodal evidence $\tilde{X}_{spk}$.
2. Generate constraints $C_=$ and $C_\neq$ from evidence according to Scheme 1, 2 and 3.
3. Generate face clustering $\tilde{X}_{face}$ using Eq. (7).

Algorithm 1 describes the proposed face clustering method based on speaker constraint propagation. Let X_S and X_F denote respectively sets of speakers and faces. First, the Speaker-Face mapping $M^{F \Leftrightarrow S}$ is generated following the method described in Section 3. The speakers partition $\tilde{X}_S$ results from the Speaker Clustering method described in Section 4.2. Then, new face clustering constraints $C_=$ and $C_\neq$ are obtained from cross-modal evidence using the proposed propagation schemes. Finally, face clustering $\tilde{X}_F$ is performed by solving the ILP problem described in Equation 7 with the generated soft constraints.

5. RESULTS AND DISCUSSION

5.1. Corpus

We used the TV recording corpus "*phase1_train*" from the *REPERE* Challenge [15]: 135 videos from LCP and BFMTV channels. It consists of 8624 annotated keyframes (about 1 every 10s). For each keyframe, annotations cover three modalities: text (overlaid text, person names in the text), speech (speaker identity, speech transcript and names in the transcript) and video (face outline, person name, occlusions and attributes). The annotated keyframes give a total amount of 9748 faces to be identified.

Table 1 shows the potential of inter-modality propagations. A face is visible in 98.5% of keyframes containing their cor-

Modality A	**Modality** B	$A \Rightarrow B$	$B \Rightarrow A$
Overlaid name	Face	98.5	10.0
Face	Speaker	42.1	63.4
Overlaid name	Speaker	80.4	12.3

Table 1. Co-occurrence statistics on reference keyframes in the Repere Corpus in % of keyframes.

responding name and 80.4% of OPNs identify the current speaker. In addition, 42.1% of the time a face is visible the person is also speaking, while 64.4% of the time the speaker is also visible on screen. This justifies our intuition that cross-modal information should be used to help face clustering.

5.2. Evaluation protocol

To evaluate the capacity of our method to identify faces, we propagate OPNs to the face clusters that maximize the total overlap duration with face tracks in the cluster. The usual metrics precision (P), recall (R) and F-score (F) are used in addition to *EGER* (Estimated Global Error Rate), the official metric of the *REPERE* challenge, defined as follows:

$$\text{EGER} = \frac{\#\text{inserted} + \#\text{missed} + \#\text{confused}}{\#\text{references}} \quad (8)$$

where #references is the number of named people in the reference keyframes, #inserted, #missed and #confused are the number of errors in each category. A lower EGER means better performance.

5.3. Experiments

We performed experiments on the *REPERE* corpus using the cross-modal constraint generation described in Algorithm 1. Each face cluster is named with the overlaid name cooccurring most often. In addition to cross-modal constraints, we add cannot-link constraints between face tracks which overlap in time or space. The size of the problem was controlled by setting $\sigma = 0.025$ and all thresholds are determined on a disjoint development set. The following variants are evaluated:

- Baseline: faces named from clustering without cross-modal constraints.
- Cross-modal: faces named from clustering with constraints from $M^{F \Leftrightarrow S}$.
- OPN-only: $M_{OPN}^{F \Leftrightarrow S}$ and additional cannot-link constraints between overlapping face tracks.
- Lip-only: $M_{Lip}^{F \Leftrightarrow S}$ and additional cannot-link constraints between overlapping face tracks.
- Full: $M^{F \Leftrightarrow S}$ and additional cannot-link constraints between overlapping face tracks.

System	**P**	**R**	**F**	**EGER**
Baseline	62.9	48.1	54.5	62.2
Cross-modal	68.2	50.0	57.7	58.9
OPN-only	63.7	49.9	56.0	61.5
Lip-only	67.3	50.4	57.6	59.0
Full	**68.3**	**50.7**	**58.2**	**58.0**

Table 2. Performance of unsupervised face identification systems for all shows.

5.4. Results and discussion

Table 2 summarizes the performance of the unsupervised face identification system using multi-modal constraints. All cross-modal variants obtain an improvement in term of precision, recall and EGER compared to the baseline. These variants result in more faces correctly identified and less confusion. The improvement is explained by the fact that constraints generate from the speaker modality allow to obtain purer face clusters.

Comparing OPN-only and Lip-only shows that the Face⇔Speaker mapping from lip activity only achieved better results than the one based only on OPN propagation. This can be explained by the disproportion between the number of constraints generated by Lip and OPNs evidence (Lip activity is measured for all frontal faces while OPNs are detected in maximum 10% of keyframes). The use of both mappings resulted in significant improvement (+5.3 Precision). Finally, adding cannot-link constraints between overlapping face tracks in addition to cross-modal constraints based on both OPN and Lip activity allowed to avoid grouping similar but overlapping tracks.

6. CONCLUSION AND FUTURE WORK

In this paper, we introduced a general framework of constrained multiple view clustering in asynchronous data. Its application to unsupervised face identification based on constrained clustering obtained promising results (a gain of 3.7 absolute in F-measure and 4.2 in EGER). The proposed constrained-clustering based on the generation of cross-modal constraints from speaker clustering has limits since the face-speaker mapping is automatic and potentially conflicting constraints might be generated. A way of reducing that problem is to train an adaptive penalty weight on the propagated constraints.

In our application, cross-modal evidence propagation was applied from the speaker to the face views but that idea can be applied iteratively on both views. In future work, we want to introduce a variety of constraints based on structural cues of the shows being processed. For instance, interviews in TV studio can add a prior on the number of clusters. Also, constraints generated from shot-clustering can help infer the presence of a person even if the associated face is not detected.

7. REFERENCES

[1] Feifan Liu and Yang Liu. Identification of soundbite and its speaker name using transcripts of broadcast news speech. *ACM*, 2010.

[2] J. Poignant, H. Bredin, V. Le, L. Besacier, C. Barras, and G. Quénot. Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast. *Interspeech*, 2012.

[3] Ma. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automated naming of characters in tv video. *Image Vision Comput.*, 2009.

[4] Ron Bekkerman and Jiwoon Jeon. Multi-modal clustering for multimedia collections. *CVPR*, 2007.

[5] Matthew B Blaschko and Christoph H Lampert. Correlational spectral clustering. *CVPR*, 2008.

[6] E. Eaton, M. desJardins, and S. Jacob. Multi-view clustering with constraint propagation for learning with an incomplete mapping between views. In *ACM on Information and knowledge management*, 2010.

[7] M. Rouvier and S. Meignier. A global optimization framework for speaker diarization. In *Speaker Odyssey*, 2012.

[8] S. Sugato Basu, M. Bilenko, A. Banerjee, and R. Mooney. Probabilistic semi-supervised clustering with constraints. *SIAM*, 2004.

[9] M. Bendris, D. Charlet, and G. Chollet. Lip activity detection for talking faces classification in tv-content. *CBMI*, 2010.

[10] M. Delakis and C. Garcia. Text detection with convolutional neural networks. *VISAPP*, 2008.

[11] C. Barras, X. Zhu, S. Meignier, and J-L. Gauvain. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech and Language Processing*, 2006.

[12] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 2002.

[13] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 2004.

[14] R. Auguste, A. Aissaoui, J. Martinet, and C. Djeraba. Les histogrammes spatio-temporels pour la réidentification de personnes dans les journaux télévisés. *CORESA*, 2012.

[15] J. Kahn, O. Galibert, L. Quintard, M. Carré, A. Giraudel, and P. Joly. A presentation of the repere challenge. In *CBMI*, 2012.