Idiap Research Report

FUSING MATCHING AND BIOMETRIC SIMILARITY MEASURES FOR FACE DIARIZATION IN VIDEO

Elie Khoury Paul Gay Jean-Marc Odobez

Idiap-RR-31-2013

NOVEMBER 2013

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Fusing Matching and Biometric Similarity Measures for Face Diarization in Video

Elie Khoury*
*Idiap Research Institute
Rue Marconi 19
Martigny, Switzerland
elie.khoury@idiap.ch

Paul Gay*
†LIUM, University of Maine
Avenue Laennec
Le Mans, France
paul.gay@idiap.ch

Jean-Marc Odobez*
Idiap Research Institute
Rue Marconi 19
Martigny, Switzerland
odobez@idiap.ch

ABSTRACT

This paper addresses face diarization in videos, that is, deciding which face appears and when in the video. To achieve this face-track clustering task, we propose a hierarchical approach combining the strength of two complementary measures: (i) a pairwise matching similarity relying on local interest points allowing the accurate clustering of faces tracks captured in similar conditions, a situation typically found in temporally close shots of broadcast videos or in talk-shows; (ii) a biometric cross-likelihood ratio similarity measure relying on Gaussian Mixture Models (GMMs) modeling the distribution of densely sampled local features (Discrete Cosine Transform (DCT) coefficients), that better handle appearance variability. Experiments carried out on a public video dataset and on the data from the French REPERE challenge demonstrate the effectiveness of our approach in comparison with state-of-the-art methods.

Categories and Subject Descriptors

I.4.9 [**Image Processing and Computer Vision**]: Applications

Keywords

Face diarization; clustering; similarity measures

1. INTRODUCTION

We address the problem of face diarization within videos. That is, we aim to automatically answer the question “who (whose face) appears in the video, and when?”, as illustrated in Fig. 1. This task has direct applications in the structuring and indexing of video programs (and beyond, of personal photo or video collections) through the generation of metadata. It is useful as a preprocessing step for browsing or fast annotation of the person identities or form the basis, with audio diarization, for further analysis of people behaviors and of their interaction or relationships [12].

The face diarization process usually consists of four main steps as shown in Fig. 2: (i) shot boundary detection, that aims to split the video stream into homogenous video clips; (ii) face detection, generally consisting in detecting frontal and profile faces within each shot; (iii) face tracking, that temporally extends the face detections within each shot, and finally (iv) a face clustering step that groups all face tracks which belong to the same person. In this paper, we focus on the later step that is the most important and challenging due to potentially large within person variabilities (pose variation, lighting conditions, occlusion, make-up or accessories like glasses) as compared to between person variabilities.

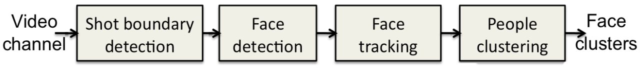

Figure 2: Face diarization process.

Most of the previous work on video face diarization have addressed the problem using matching-based face pairwise similarity measures [3, 20, 7, 10]. That is, they compare two clusters by directly computing the similarity between their corresponding face samples represented by a set of local descriptors like SIFT or Gabor filter outputs that are computed around detected interest points or landmarks. Such approaches are more appropriate for matching and comparing faces acquired within the same conditions. This often corresponds to faces appearing within a TV show episode, or during talk-shows or debates. However, when the face appearance variability increases due to pose, hair cut or illumination changes, the discrimination power of such matching similarity measures drops as the within person measure comparisons become closer to between-person ones. In short, such measures usually lack generalization capabilities.

To address this generalization issue, we propose to rely on biometric model-based approaches [4, 13, 21, 17], whose goals are specifically to handle face variations across conditions and time. In contrast to the matching case, their aim is to represent each face-cluster with a statistical model of the distribution of densely sampled local features (thus reducing alignment issues) whose parameter training leverages on statistics learned *a priori* from thousands of faces thanks to the use of Maximum a Posteriori (MAP) adap-

*This work was supported by the French Research Agency (ANR) under the Project SODA. The authors gratefully thank the ANR for their financial support. More information about ANR-SODA is available from the project web site http://anr-soda.univ-lemans.fr

†This work was done during the stay of the first two authors as visiting researchers at Idiap.

ICMR'13, April 16–20, 2013, Dallas, Texas, USA.
Copyright 2013 ACM 978-1-4503-2033-7/13/04 ...$15.00.

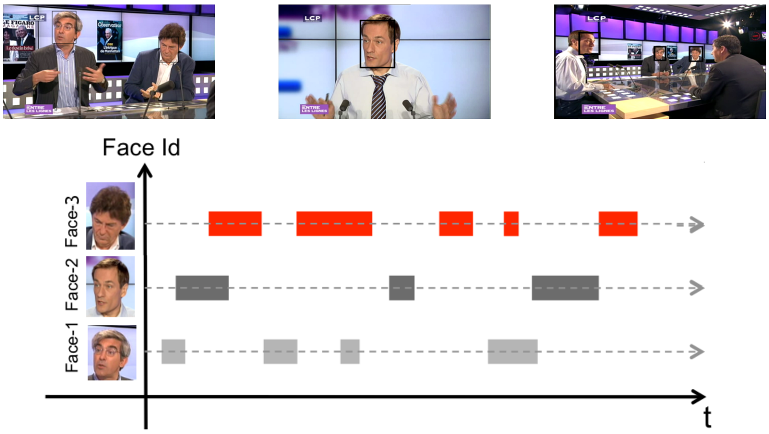

Figure 1: Face diarization task. Top row: Sample frames of a TV debate illustrating variability in face size, pose, background, number of people, etc. Bottom row: example of face diarization output. The system automatically retrieves 4 different faces that appear in the video. It provides all sequences of frames in which each face appears.

tation learning. The similarity between two face-clusters is then estimated by comparing indirectly their statistical models through their evaluation on the corresponding cluster data. Although the statistical modeling leads to less powerful plain matching capabilities, variabilities are better handled.

In this paper, we propose a novel hierarchical bottom-up clustering method that appropriately combines and takes advantage of a feature-based matching similarity measure and a model-based similarity measure. To the best of our knowledge, we are the first to investigate the model-based approach for face clustering, and hence also its combination with matching based methods. Indeed, a key assumption made by our bottom-up clustering method is that the optimal face representation and similarity comparison for clustering depends on the amount of data available and on the degree of face variability that are found in the data. On the one hand the feature-based matching similarity is optimal with relative small variabilities, i.e. when it is possible to perform direct comparison with high confidence, and is thus particularly robust for clustering faces acquired in similar conditions (close moment in time, same scene). In addition, as it has the advantage of being able to perfectly work with small-duration face tracks, it is more appropriate at an early stage of the clustering process. On the other hand the model-based method is optimal when more data is available in initial clusters and when more variabilities are present due to its robustness to several factors. Its exploitation is thus more appropriate at later stages of the clustering.

Experimental validations were conducted on two benchmarks datasets: the public dataset from the TV series “Buffy” provided by [6] exhibiting face variability across episodes, and 38 TV programs from the “REPERE” challenge [9] containing talks shows, news, and debates. They illustrate the behavior of the algorithm and the benefit of our approach, and demonstrate its state-of-the-art performance.

The remainder of this paper is organized as follows: in section 2, we review face clustering related work. Section 3 details the proposed matching-based and model-based measures as well as the hierarchical clustering proposed to combine them. Section 4 describes the datasets, metrics, and experimental results. Section 5 concludes the paper.

2. RELATED WORK

Face clustering requires the design of face representations and comparison approaches robust to intra-person variability (pose, lighting, partial occlusion,...). Below we first review methods that were specifically designed for face clustering, and then discuss biometric methods developed for the different but related face verification or recognition tasks.

In the work of [3, 5, 6, 7], researchers use local descriptors at facial landmarks as face representation. The first one [7] uses pixel-based descriptors while the others rely on points-of-interest descriptors like SIFT or SURF. If the data quality is not sufficient for reliable facial landmark detection, one alternative is to exploit those descriptors computed around automatically detected points-of-interest [3, 10, 19]. In this case, most of the time the spatial information is lost, although [19] keeps it by adding a spatial term depending on key-point positions to compare descriptors between two face images. In the paper, we will refer to these methods as relying on feature-based or matching similarity measures. They usually enable to perform pairwise comparison between images when small face variation are observed.

When clustering faces from videos, a useful preprocessing step can be done by tracking faces on consecutive frames within a shot. This produces face tracks, i.e. the images of a single character across multiple frames within a video shot. Comparing two face tracks offers the possibility to be more robust to pose variations, and this was used in most works cited above, e.g. by using as face-track distance the average of all face distances [10, 11].

Face representations have also been developed in the context of biometric tasks like face verification [4, 14, 17], and in the paper we will refer to them as model-based (or biometric) methods. Indeed, state of the art methods in this domain achieve high robustness by explicitly training a biometric model for each person they want to identify. The model is usually characterized by the parameters of the distribution (often a Gaussian Mixture Model, GMM) of densely sampled features. A first interesting property of these approaches is the use of a Universal Background Model (UBM) which is trained from a large number of subjects and aims at represent all the population. The UBM can be used as prior during fitting, preventing from over-fitting and allowing to handle small amount of data. In addition, the availability of the UBM model allows to compute a likelihood ratio of the test sample between the biometric model and the UBM. In other words, the ratio normalizes the likelihood of a test sample, allowing to detect for instance if a low likelihood of the sample for a biometric model is due to the inappropriateness of the biometric model or due to (potentially noisy) data itself. It is important to note that the UBM methodology has also been used with success for other modalities like in speaker diarization [21]. A second interest of these methods is the use of densely sampled features. For instance, a GMM (adapted from a UBM) representing the distribution of 2D DCT coefficients [14] of spatially neighboring blocks was shown in [4] to be more robust to alignment error than when using local descriptors, and the miss detection problem encountered with facial landmarks is also avoided. However, this is at the cost of the loss of the spatial information.

Despite their interesting properties, up to our knowledge these model-based techniques have never been used for a face clustering task. In our work, we aim at combining the feature-based method and the model-based method by making the following assumption. We state that when the clustering is performed over similar faces with little amount of data, the features-based method is more confident because the GMMs adapted from the UBM are too general to handle those little variations. On the other hand, when there is a high variability and relatively large amount of data, the model-based method becomes the optimal one. Indeed, it is more robust and there is enough data to learn it efficiently. We verified experimentally this assumption in section 4.

It should be noted that additional information can help the face diarization. To better compare people, [7] includes clothes information with a color histogram. [16] exploits uniqueness constraint of a face in a image and the fact that conversations in TV series induces a particular shot structure. Although promising, the inclusion of this additional information is beyond the scope of this paper.

3. PROPOSED FACE CLUSTERING

In this Section, we assume that a video has been first processed, using the different steps described in Fig. 2. At the end of this process, we end up with a set of face tracks $\{FT_i, i \in 1...N^{ft}\}$ that we would like to merge into clusters that contain only tracks of the same person.

Clustering overview. There is a large number of clustering methods. In the speaker diarization [21] and face clustering literature [20, 8], hierarchical bottom-up clustering approaches are dominantly used by state-of-the-art systems. Those approaches start with an over-segmentation of the data with high purity clusters (i.e. containing data of a single person) and then merge the more similar segments. If required, cluster representation is then updated and improved through model fitting as more and more data segments are included into each cluster. Hierarchical methods have no *a priori* knowledge about the desired number of clusters (the real number of persons), and leave the model selection issue to the choice of a threshold on a fitting criteria or distance. This threshold is often learned using a validation dataset.

In this paper, we follow this approach. Each face-track is initially considered as a cluster. Then, the cluster pairs that are most similar according to a similarity measure D_C are merged until a stopping criterion is verified.

In the following sections, we first describe our face representation, then the two cluster similarities involved in our algorithm, and finally the different clustering strategies we propose to appropriately combine them.

3.1 Face and face track representations

A face contains many discriminative features, like its shape, the eyes, the hair or the skin color. All those features can in principle be used jointly in order to recognize people. However, due to variations in illumination, scale, pose, or due to partial occlusions or un-aligned detections, representing faces (and more generally people) in an invariant yet discriminant fashion is difficult. In this paper, we adopted two types of features that contain complementary information regarding the clustering process.

SURF features. SIFT and SURF [2] features computed on regions around automatically detected points-of-interest are known to be robust to scale, rotation, and illumination variations. While initially developed for wide-baseline matching or image retrieval, they have also shown their interest in face clustering as well [20] due to their ability to provide a good matching measure of faces captured within a given context. For a given face F, the first face representation that we use is given by the set of associated SURF features: $\text{Surf}(F) = \{f_i^{surf}, i = 1, ..., N_F^{surf}\}$.

Discrete Cosine Transform (DCT) features. Detecting interest points is interesting for matching. Unfortunately the point locations do not carry any semantic information, which makes it difficult to build a single model from multiple faces of the same person. One alternative is to detect facial landmarks, and to rely on features extracted around them. However, detecting landmarks is not always trivial, and extracted features may be dependent on the precision of the localization. A solution is to extract features on a dense grid sampling of the face. Indeed, dense sampling has proven to be superior to interest points for many tasks of computer vision, including object, scene, and action recognition. In face biometry, dense feature sampling in combination with statistical models has also proved to be very competitive, even if the localization information is discarded, as shown for instance in [17].

We thus propose to adopt a similar strategy to extract the features of our second face representation: $\text{Dct}(F) = \{f_i^{dct}, i = 1, ..., N^{dct}\}$. More precisely, given the face image, we first estimate the eye locations and use them to register and normalize the image size. This results in an image of 80 × 64 pixels which is then pre-processed using the Tan and Triggs illumination normalization [15]. Then, the 2D DCT is applied on 8×8 densely sampled overlapping blocks (with a step of 1 pixel between block location), and only the subset of $D_{dim}^{dct} = 28$ low-frequency components of the DCT are kept using zig-zag pattern.

Face track representation. To avoid processing all images of a face track, we decided to only work with a limited number of images per track. More precisely, $N^{kf} = 9$ *keyfaces* are selected from each face-track by dividing the track in equal intervals.

3.2 Matching cluster similarity

To define the cluster similarity, we first have to define the similarity between individual faces.

Face feature similarity. As feature similarity between two faces F_1 and F_2, we use the "Average N-Minimal Pair Distance" (ANMPD) $d_s(F_1, F_2)$ between the two sets of SURF features that was proposed in [10]. As its name suggests, the ANMPD measure returns the average of the N (we used $N = 6$ in this work) smallest distances between the SURF feature vectors that match between the two faces. The measure d_s is thus small when the face similarity is high.

Matching cluster similarity. Considering two face-clusters C_i and C_j with their associated set of keyfaces $\{F_a^i, a = 1, \ldots, N_i\}$ and $\{F_a^j, a = 1, \ldots, N_j\}$, we compared them using the following cluster similarity:

$$D_f(C_i, C_j) = \frac{1}{N_i N_j} \sum_{a=1}^{N_i} \sum_{b=1}^{N_j} d_s(F_a^i, F_b^j) \qquad (1)$$

By definition, this measure favors the creation of compact clusters where all faces are compared to each other. In practice, we have found this to work as effectively as other approaches which for instance only seeks for the best subset of most similar faces between the two image sets. Note that after merging C_i and C_j, the similaritys between the new cluster $C_{i'}$ and any other cluster C_k can be computed recursively as:

$$D_f(C_{i'}, C_k) = \frac{N_i \times D_f(C_i, C_k) + N_j \times D_f(C_j, C_k)}{N_i + N_j} \qquad (2)$$

3.3 Model-based similarity

As motivated in the introduction, statistical models for representing the distribution of local descriptors are powerful tools to handle face variabilities while keeping information about the distinctive features. In this work, we use Gaussian Mixture Models (GMM) which have proved to be robust in this context, and allow the exploitation of reliable Maximum A Posteriori (MAP) parameter estimation schemes. Below, we describe the model, how to learn it, and then define the inter-cluster similarity measure that we use.

Model and parameter learning. We model the likelihood of any DCT feature vector f^{dct} within a face image as:

$$p(f^{dct}|\Lambda) = \sum_{i=1}^{N_g} \omega_i \mathcal{N}(f^{dct}; \mu_i, \Sigma_i) \qquad (3)$$

where $\Lambda = \{\omega_i, \mu_i, \Sigma_i, i = 1...N_g\}$ represent the GMM model parameters (we used $N_g = 200$ gaussians which is a good trade-off between effectiveness and efficiency). Thus, for a cluster C_i containing the faces $F_{ij}, j = 1, ..., N_i$, the set of local features is defined as the union of features over all faces: $\text{Dct}(C_i) = \bigcup_j \text{Dct}(F_{ij})$. Its log-likelihood L for a given model Λ is then given as:

$$L(\text{Dct}(C_i)|\Lambda) = \prod_{f^{dct} \in \text{Dct}(C_i)} p(f^{dct}|\Lambda) \qquad (4)$$

In practice, rather than using Maximum Likelihood, the GMM model parameters Λ_i of cluster i are learned through mean-only MAP adaptation from a prior universal background model (Λ_{ubm}) trained on an independent large dataset of images. This avoids the over-fitting problems that can occur given the large amount of parameters and the potentially low number of training faces.

Cross-likelihood ratio. To compare two face-clusters C_i and C_j with their corresponding model parameters, we rely on the Cross Likelihood Ratio (CLR) defined as:

$$CLR(C_i, C_j) = log \frac{L(\text{Dct}(C_i)|\Lambda_j)}{L(\text{Dct}(C_i)|\Lambda_{ubm})} + log \frac{L(\text{Dct}(C_j)|\Lambda_i)}{L(\text{Dct}(C_j)|\Lambda_{ubm})} \qquad (5)$$

The CLR is a symmetric similarity measure, which is positive when the clusters are similar, and negative in the other case. It captures how well the features from one cluster are likely according to the model of the other cluster, as compared to the likelihood given by the UBM model, and vice-versa. The UBM model thus serves as a reference. It allows to distinguish for instance whether a low data likelihood is due to an inadequacy with the tested model, or to the data itself.

Model-based similarity measure. The similarity measure that we used is defined as follows. Given an initial set of clusters $\{C_1, ..., C_{N_{init}}\}$, a model is trained for each cluster, and the CLR similarity between these clusters, denoted as $S_m(C_i, C_j) = CLR(C_i, C_j)$, is computed. Then, after each merge, we have a new cluster $C_{i'} = C_i \cup C_j$ with its size $N_{i'} = N_i + N_j$. We then define the new similarity of this cluster with any other cluster k as:

$$S_m(C_{i'}, C_k) = \frac{N_i \times S_m(C_i, C_k) + N_j \times S_m(C_j, C_k)}{N_i + N_j} \qquad (6)$$

Qualitatively, this means that rather than learning a single face model from all data belonging to a cluster -that could be corrupted in case of wrong merging of faces from different people- and exploiting it for likelihood evaluation, we prefer to rely on the original face models learned on purer clusters to evaluate the likelihood of a cluster k data, and defined this later one as a weighted average of its likelihood with respect to the models of all initial cluster belonging to the cluster i'.

3.4 Fusing matching-based and model-based methods

The matching similarity D_f and model-based similarity measure S_m satisfy different purposes. The first one is adequate to directly find matches between face tracks acquired in very similar conditions, while the second one, that may require to have sufficient data to adapt the GMM model, can better handle appearance variability at the cost of loosing some face representation accuracy. From a bottom-up clustering perspective, this means that the first one is more adapted at the beginning of the clustering process, whereas the second one can be applied later on. We therefore have adopted the following strategy:

- first, apply the clustering using only the feature-based similarity, ie define D_C as D_f;
- once a threshold is reached, i.e. $D_f(C_i, C_j) \geq T_f$ for any two cluster, use the current clusters as base cluster to learn GMM models (see previous section), and continue the clustering using a combination of the measures as cluster dissimilarity:

  $$D_C(C_i, C_j) = D_f(C_i, C_j) - \alpha S_m(C_i, C_j) \qquad (7)$$

  where $\alpha \geq 0$ denotes the contribution of the model-based similarity to the overall merging criterion.

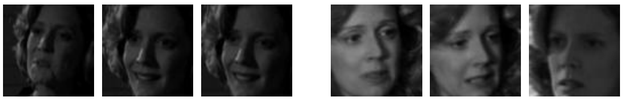

Figure 3: Detected face samples for the character "Joyce" in Buffy: first three samples extracted from Episode 1, last three from Episode 3. Notice the higher appearance variability between the inter episode samples than the intra episode ones.

4. EXPERIMENTAL EVALUATION

We first describe the datasets used in our experiments before presenting the evaluation metrics in section 4.2. Further evaluation protocols and results are detailed in section 4.3.

4.1 Datasets Description

Two datasets are used in order to evaluate the different contributions of our work.

Buffy dataset. The first dataset was used in [8, 6]. It contains 327 face-tracks selected from episodes 9, 21 and 45 of the TV series "Buffy the vampire slayer", where each episode belongs to a different season of the series to provide face variabilities. Face tracks were obtained using an automatic system, and false positive tracks and face-tracks that did not belong to the 8 main actors were manually discarded. This dataset shows generally a higher appearance variability between inter episode tracks with respect to intra episode ones tracks, as illustrated in Fig 3.

REPERE dataset. The second dataset contains 38 video files from the french evaluation challenge REPERE [9]. The videos feature news, debates and talk-shows recorded from two french information TV channels (LCP and BFM). The videos were manually but partially annotated by ELDA[1]. The total duration of the recordings is 21.5 hours, with 3 hours manually annotated. There are 1076 face tracks (initial clusters) that belong to 264 people. This dataset is challenging due to the large number of clusters, very unbalanced cluster sizes (ranging from one when a person appears in a single shot, to several tenth for anchor men or invited people) and view point changes. As with the first dataset, an automatic face track extraction algorithm was applied to the videos, that lead to different types of error, ans we similarly filtered out false-positive erroneous tracks to allow evaluation of the clustering task only.

For experimental and hyper-parameter setting purposes and following the REPERE experimental protocol, the dataset was further divided into a development set (DEV), and a test set (TEST) of 28, and 10 videos, respectively.

4.2 Evaluation metric

Buffy dataset. To allow comparison on this dataset, we used the clustering metric proposed in [8] and also used in [6]. It computes the number of clicks that would be needed to manually correct the automatic output and obtain the ideal result. More precisely, in this scheme, it is assumed that one click is needed to associate the correct name to a cluster (even containing wrong face-tracks), and that one click is needed to provide the correct name of a face-track whose identity is different than that assigned to the cluster it belongs to.

REPERE dataset. In this case, we primarily evaluated the methods in term of Diarization error rate (DER) that was previously proposed for speaker diarization in the NIST RT [2] competitions, and provided as well the results in terms of clicks for comparison with the Buffy dataset. The DER is computed from the ratio of three error rates divided by the total duration of the video: a miss detection error rate, that counts the number of times a face that exists in the ground truth is not detected by the automatic system; a false alarm error rate, that counts the number of times a face is detected by the system while no corresponding face is available in the ground truth[3]; and the confusion error rate, that counts the number of times for a given face, the ground truth identity associated with the cluster label automatically provided by system does not match the one in the ground truth for that face. In this dataset, methods are applied separately to each of the videos, and the final performance is obtained by measuring the DER from the aggregated error rates.

Compared methods. We compared three methods with our system. The two first ones rely on the individual (dis) similarity measure: D_f corresponds to using only the feature-based matching dissimilarity measure. The clustering is conducted until a threshold T_f is reached. S_m corresponds to using only the model-based distance (until a threshold T_m is reached). Finally, D_C denotes the combined method, as described in section 3.4. Note that all involved thresholds are learned on the development set for the REPERE data.

4.3 Experimental Results

Results on the Buffy dataset. Fig. 4 and Table 1 present the results obtained with the different clustering methods. In addition, it also shows the results from [8, 6] and reported in [6]. All three methods perform bottom-up clustering and rely on a face represented using a fixed set of descriptors computed around facial landmarks, but differs on the metric used to compare two faces: L2 uses a Euclidian distance; LFW uses a metric learned from the Labeled-Face-in-the-Wild dataset; and UML uses a metric learned in a discriminative and unsupervised fashion using faces within tracks as positive samples, and faces from different tracks in a given shot as negative samples.

Fig. 4 shows the evolution of the performance metric in function of the number of cluster (in logarithmic scale), where for clarity we have split the results into two figures. From the top one, we can see that (i) both our feature-based D_f and model-based S_m approached outperforms the landmark-based method relying on L2 and LFW metrics; (ii) the feature-based D_f outperforms S_m, probably due to the lack of training data at initialization for the model-based approach S_m; (iii) the feature-based D_f outperforms the UML approach [6] for a number of clusters higher than 60, but then perform worse as the number of clusters is reduced. This might be explained by the ability of our feature-based method to better match similar face images, and its higher difficulty when more variability is present.

The bottom plot of Fig. 4 further shows the result using our combined approach D_C (for $\alpha = 0.5$). As can be seen, it outperforms both the single measure based clustering D_f

[1] http://www.elda.org/

[2] http://nist.gov/itl/iad/mig/rt.cfm

[3] Note that as we removed these false alarms after the tracking process, this rate will be 0 in reported experiments.

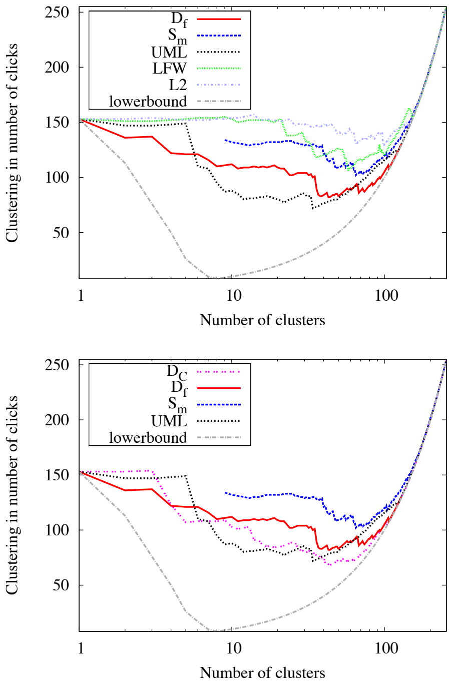

Figure 4: Results on the Buffy dataset. Number of clicks with regards to the number of clusters. Top: comparison between the proposed D_f and S_m measures and the L2, LFW and UML metrics [6]. Bottom: comparison of our proposed combination approach D_C with D_f, S_m and UML.

Table 1: Minimum number of clicks needed to correct the automatic on the Buffy dataset.

Method	Minimum Number of Clicks	Number of Clusters
L2	129	98
LFW	106	58
UML	72	34
D_f	82	43
S_m	102	65
D_C	**68**	44

and S_m, and is better than or equivalent to UML at the beginning of the clustering process (number of clusters higher than 38) and then is worse or better depending on the number of clusters. Table 1 details the values of the minimum number of clicks and its corresponding number of clusters for each of the methods. It shows that our combined approach D_C provides the best result with 68 clicks, showing its ability to achieve state-of-the-art results. In this case, 60 clicks come from the clustering errors, and the remaining 8 clicks come from annotating the clusters with their real names which makes the percentage of clicks due to the clustering errors equal to 23.3%.

In order to highlight the different behavior of the D_f and S_m measures, we computed their intra- and inter-episode values for the main character (Joyce) of the dataset. Fig. 5 plots the resulting corresponding cumulative histograms. Qualitatively, we can see from the top figure that D_f is suitable at comparing similar faces with high accuracy, as illustrated by the fact that 26% of intra-episode measures are lower than any inter-episode measure. As a comparison, for S_m, only 12% of intra-episodes have higher similarity than any inter-episode similarity.

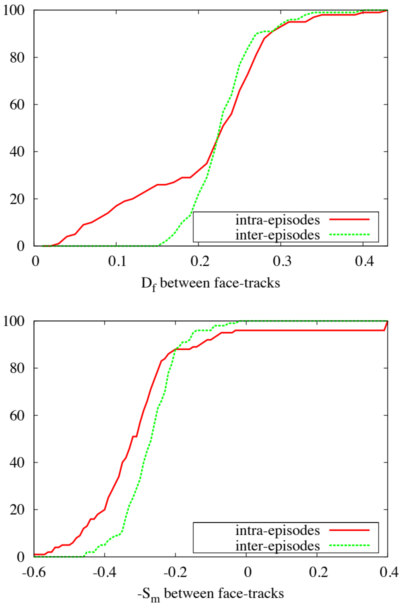

Figure 5: Cumulative percentage of tracks pairs from the character Joyce whose pairwise (dis)similarity measures are below a given threshold for the D_f (top) and (-S_m) (bottom) measures when considering track pairs within an episode or between an episode.

Results on the REPERE dataset. For this dataset, as we deal with multiple videos, we need to set the parameters involved in the different algorithms: for the D_f and $(-S_m)$ approaches, this mainly corresponds to the thresholds T_f and T_m respectively to be used to stop the clustering process. For the combined approach[4], there are the coefficient α controlling the contribution of the model-based measure to the overall dissimilarity measure (cf Eq. 7), and the stopping threshold T_C. These parameters are tuned to provide the best result for a given approach on the development set (DEV), and then further used on the TEST set. Results are presented in Table 2, where the minimum DER obtained by optimizing the parameters is reported as **min-DER**. Note that we also report on the DEV set the DER (denoted **cross-DER**) obtained through cross-validation: for each video, the parameters are tuned using the remaining files and used to the DER on that file. This is repeated for all files, and finally their average DER is computed. The cross-DER reduces the impact of files that have relatively long duration and provide an idea of the performance variance due to parameter setting.

According to the results in Table 2, we can see that the combined approach outperforms both the D_f and S_m methods on the DEV and TEST sets: we obtain at least 14% relative gain (in terms of cross-DER) on the DEV set and around 35% of relative gain on the TEST set. Interestingly, we can notice that the S_m approach provides the more stable results with respect to the stopping criterion, as there

[4] For this combined approach, the threshold T_f for which only the feature-based measure is used in the initial clustering stage was set to 0.09, see Section 3.4.

the cross-validation on the DEV provides the same results than when doing a direct optimization of the results, and the TEST results are closer to the DEV ones. This is probably due to the use of the UBM model and the inherent data normalization that it provides in the cross-likelihood ratio.

Fig. 7 illustrates qualitatively the output of the 3 clustering methods for 2 different persons, and where each row corresponds to an automatic output cluster. For person 1, the S_m based clustering suffers not only from sub-clustering, but also from confusion (mixing up 2 different persons in the same cluster). On the other hand, the D_f based clustering suffers from sub-clustering. However, the clustering output for that person is perfect for the combined measure. Similar observation is found for person 2.

Impact of different α values. Fig. 6 shows the results obtained by varying the contribution of the model-based measure to the overall cluster dissimilarity measure. The behavior is quite similar on the two datasets, with optimal values found in the 0.3 to 0.5 range. It is also worthy to notice that the clustering error rate of the combination measure D_C in this range is lower than the error rates obtained by D_f and S_m alone (the curve keeps increasing beyond 0.7).

Difficulty of the databases. We also evaluated the clustering on REPERE database using the metric used on Buffy database. In this case, the percentage of clicks due to the clustering errors on both DEV and TEST sets are equal to 7.6% and 4.5%, respectively less than the one obtained on Buffy (23.3%). This shows that, the clustering task is more challenging on series and movies than on news and debates. Note however that in this domain, there is some discrepancies between programs, where lively talk-shows actually generate more errors than political debates.

5. CONCLUSIONS

We proposed a face diarization method which combines a feature-based and a model-based (dis)similarity measures. We show that each measure is the most efficient in different cases depending on the variability of the faces and the sizes of the clusters. The two approaches are combined appropriately, and this results in a decrease of the diarization error rate. As a future work, the automatic extraction of head pose [1] could be used to generate pose dependent face models and improve comparison between faces, while additional person detector [18] would allow a better tracking of people. We also plan to extend our work to a person diarization method, by integrating more visual features derived from clothes [16], and integrate audio information. The similarity between our face diarization method and the models from speaker diarization [21] should facilitate this integration.

6. REFERENCES

[1] S. O. Ba and J. M. Odobez. A rao-blackwellized mixed state particle filter for head pose tracking. In *ACM-ICMI Worksh. on Multi-modal Multi-party Meeting Processing(MMMP)*, pages 9–16, 2005.

[2] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *ECCV*, pages 404–417, 2006.

[3] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli. On the use of sift features for face authentication. In *CVPRW*, pages 35–35. IEEE, 2006.

[4] F. Cardinaux, C. Sanderson, and S. Bengio. User authentication via adapted statistical models of face images. *IEEE Trans. on Signal Processing*, pages 361–373, 2006.

[5] W. Chu, Y. Lee, and J. Yu. Visual language model for face clustering in consumer photos. In *ACM Int. Conf. on Multimedia*, pages 625–628, 2009.

[6] R. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in tv video. In *IEEE ICCV*, pages 1559–1566, 2011.

[7] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automated naming of characters in tv video. *Image and Vision Computing*, pages 545–559, 2009.

[8] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *IEEE ICCV*, pages 498–505, 2009.

[9] J. Kahn, O. Galibert, M. Carré, A. Giraudel, P. Joly, and L. Quintard. The repere challenge: Finding people in a multimodal context. In *Odyssey The Speaker and Language Recognition Workshop*, 2012.

[10] E. Khoury, C. Senac, and P. Joly. Face-and-clothing based people clustering in video content. In *ACM MIR*, pages 295–304, 2010.

[11] E. Khoury, C. Senac, and P. Joly. Audiovisual diarization of people in video content. *Multimedia Tools and Applications*, 2012.

[12] S. Kim, F. Valente, and A. Vinciarelli. Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates. In *IEEE ICASSP*, 2012.

[13] S. Lucey and T. Chen. A gmm parts based face representation for improved verification through relevance adaptation. In *CVPR*, pages II–855, 2004.

[14] C. Sanderson and K. Paliwal. Fast features for face authentication under illumination direction changes. *Pattern Recognition Letters*, (14):2409–2419, 2003.

[15] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing*, (6):1635–1650, 2010.

[16] M. Tapaswi, M. Bauml, and R. Stiefelhagen. "knock! knock! who is it?" probabilistic person identification in tv-series. In *IEEE CVPR*, pages 2658–2665, 2012.

[17] R. Wallace, M. McLaren, C. McCool, and S. Marcel. Cross-pollination of normalization techniques from speaker to face authentication using gaussian mixture models. *IEEE Transactions on Information Forensics and Security*, 7(2):553–562, 2012.

[18] J. Yao and J.-M. Odobez. Fast human detection from joint appearance and foreground feature subset covariances. *Computer Vision and Image Understanding (CVIU)*, 115(10):1414–1426, 2011.

[19] S. Zhao, F. Precioso, and M. Cord. Spatio-temporal tube kernel for actor retrieval. In *IEEE ICIP*, pages 1885–1888, 2009.

[20] S. Zhao, F. Precioso, M. Cord, and S. Philipp-Foliguet. Actor retrieval system based on kernels on bags of bags. In *EUSIPCO*, pages 234–778, 2008.

[21] X. Zhu, C. Barras, S. Meignier, and J. Gauvain. Combining speaker identification and bic for speaker diarization. In *Europ. Conf. on Speech Communication and Technology*, pages 2441–2444, 2005.

Table 2: REPERE dataset. Diarization error rate (DER) on the DEV and TEST sets. The hyper-parameters tuned on the DEV set (corresponding to the min-DER results) were: $T_f = 0.15$ for D_f; $T_m = 0.15$ for S_m measure; and $(\alpha, T_C) = (0.3, 0.13)$ for the combined approach. These parameters were used on the TEST set. Cross-DER are the results obtained through parameter cross-validation on the DEV set.

Method	cross-DER (DEV)	min-DER (DEV)	DER (TEST)
D_f	6.41%	5.13%	8.33%
S_m	6.68%	6.68%	8.21%
D_C	**5.49%**	**4.37%**	**5.28%**

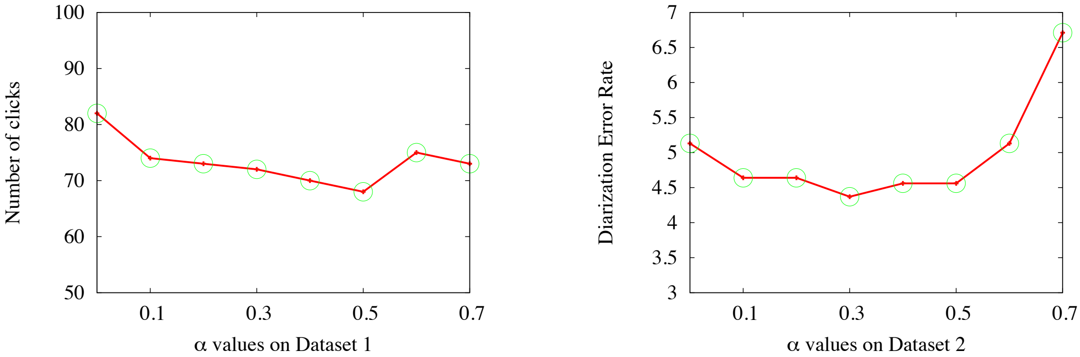

Figure 6: Impact of α on the performance on the two datasets. For the Buffy dataset the minimum number of clicks is used as measure. For the second dataset, we report the min-DER on the DEV set.

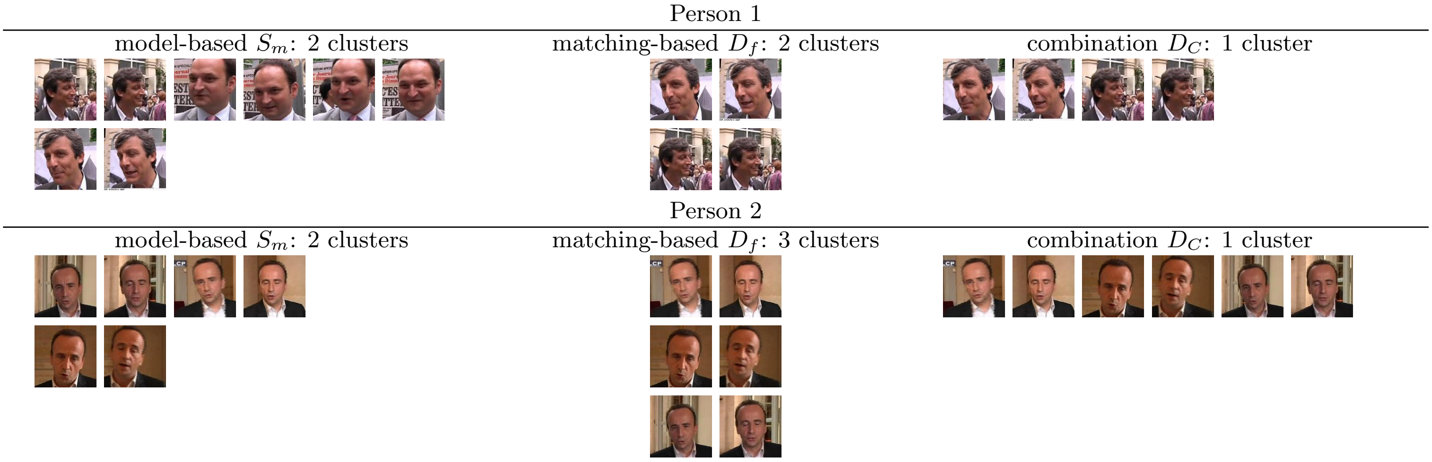
Person 1
model-based S_m: 2 clusters
matching-based D_f: 2 clusters
combination D_C: 1 cluster
Person 2
model-based S_m: 2 clusters
matching-based D_f: 3 clusters
combination D_C: 1 cluster

Figure 7: Illustration of the clustering results for the 3 methods discussed here. For each person and each method, all the clusters containing that person are represented by 2 images per face track. Each row corresponds to an automatic output cluster.