Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

Multimodal speaker clustering in full length movies

**I. Kapsouras · A. Tefas · N. Nikolaidis ·
G. Peeters · L. Benaroya · I. Pitas**

Received: date / Accepted: date

Abstract Multimodal clustering/diarization tries to answer the question ”who spoke when” by using audio and visual information. Diarization consists of two steps, at first segmentation of the audio information and detection of the speech segments and then clustering of the speech segments to group the speakers. This task has been mainly studied on audiovisual data from meetings, news broadcasts or talk shows. In this paper, we use visual information to aid speaker clustering and we introduce a new video-based feature, called actor presence that can be used to enhance audio-based speaker clustering. We tested the proposed method in three full length stereoscopic movies, i.e. a scenario much more difficult than the ones used so far, where there is no certainty that speech segments and video appearances of actors will always overlap. The results proved that the visual information can improve the speaker clustering accuracy and hence the diarization process.

Keywords Multimodal · Diarization · Clustering · Movies · Actor Presence

1 Introduction

Speaker diarization/clustering tries to detect speech segments and then cluster them in order to group together segments of the same speaker. Diarization can answer the question ”who spoke when” when used together with speaker recognition systems. This is because clustering does not determines the true identity of the speakers. It assigns labels in the detected audio segments, of the form Speaker 1, Speaker 2 etc. The true identity of the speakers can be decided by speaker recognition systems or by combining diarization results with manual labelling of the speaker clusters with their true identity, i.e. by replacing labels Speaker 1, Speaker 2 with ”John Smith”, ”Jack Brown” etc.

Department of Informatics, Aristotle University of Thessaloniki
54124 Thessaloniki, Greece
E-mail: jkapsouras@aiia.csd.auth.gr

Automatic extraction of human related semantic information from audiovisual data (such as those provided by speaker diarization) is a very important task. Methods that extract semantic information from movies, such as where and when an actor appears or speaks are important both for the general public and (even more) for the film production industry. Such information can help in indexing, organizing and searching the vast audiovisual data that exist.

Usually, video and audio are considered as different modalities and are analysed separately. In this paper, the combination of the two modalities (audio and video) for the task of speaker clustering/diarization in movies is investigated. Multimodal speaker diarization in movies content has certain inherent difficulties since unlike meetings or talk shows (targeted by most of the proposed methods), audio (speech) and video are often not coherent in movies. For example the person depicted in the video might not be the one that is speaking. Indeed the task of diarization is much easier, when the input data are from meetings or talk shows. In such setups, the visual appearance of a speaker (i.e., its clothing or facial appearance) does not change within the duration of the meeting/show. The composition of the group of participating persons typically does not change either. For talk shows, one can further assume that the speaker is in a close-up view. Moreover, the possibility that the speaker is the person that is shown (actor) is very high. These observations do not apply in the case of 3D films or films in general. In this case, the speaker/actor visual appearance may change over the duration of the movie. Furthermore, the group of people may change over time. Finally, the coherence between visual and audio scene is not guaranteed, since, for example, 3D movies video and audio scenes often capture only a part of the real scene (there may be people speaking that are not displayed or displayed people may speak but one may hear the voice of somebody else). Due to the above, the situation is much less constrained in the case of movie content and speaker diarization is more difficult than in the previously discussed setups.

The intuition behind the modalities fusion is that one can perform a similar to speaker diarization analysis upon the visual data: face clustering. In more detail, assume that faces are detected in the frames of a movie and then the detected faces are tracked over time, resulting in a number of video facial trajectories [32], [3], [8]. A representative face is selected to represent a facial trajectory. The selected faces can then be clustered into clusters, each ideally corresponding to a single actor/person. The face clustering results can be used in the audio based diarization process in order to improve the speaker clustering accuracy. In the proposed method, information from both left and

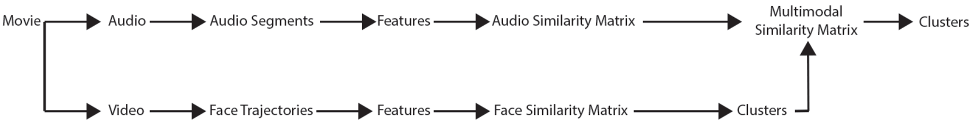


Fig. 1 A general flowchart of the proposed method.

right channels of a stereoscopic (3D) film, was exploited in order to improve the face clustering.

Three ways were investigated in order to exploit video information in the speaker diarization process. In the first approach, the similarity of two speech segments was increased when these segments have overlap with visual appearances of the same actor and was decreased otherwise. Second, visual speech detection was used. In this case, the same approach as before is followed to enhance speaker clustering but the similarity between two speech segments is altered only when visual speech is detected in the involved facial trajectories. Third, we introduce a new feature, calculated for each speech segment, which consists of the durations of visual actor appearances within this segment. The new features, which characterize each speech segment, were used to enhance speech clustering.

The main contributions of the proposed method are:

- the introduction of the actor presence features. These features incorporate the actors' visual presence in an audio segment.
- the fact that it can be successfully applied in movies (where diarization is considerably more difficult than in meeting or talk shows content) and is, as far as we know, the only method that is applicable in 3D movies and exploits information from both the left and the right channel.

The proposed multimodal approaches were tested in stereoscopic feature length movies with very good results.

2 Previous Work

Multimodal speaker diarization, which is closely related to multimodal person clustering has already been studied in the literature, but mainly on audiovisual data from meetings or talk shows, which impose far less difficulties than movie content. The video information can enhance the audio information during the speaker diarization, hence a multimodal approach to diarization (audio + video) can improve performance. In [7], Khoury et al. proposed a framework for audio-visual diarization. The authors combined audiovisual information using co-occurrence matrices. Moreover, they used information, such as face size and lip activity rates to improve the audiovisual association. The authors evaluated their method in a number of news videos, meetings videos and movies. The method proposed in [7] does not perform speaker diarization as the proposed method but correlates the audio clusters with the video clusters and performs voice, face and voice-and-face clustering. Moreover, the method proposed in [7] use different features (SIFTS instead of LBPs) and different clustering (Hierarchical bottom-up clustering instead of Normalized Cuts) in video processing. In [17], Noulas et al. proposed a probabilistic framework to perform multimodal speaker diarization. The proposed method uses a Dynamic Bayesian Network (DBN) to model the people as multimodal entities that are involved in audio and video streams and also in audiovisual space.

The model is generated by using the Expectation Maximization algorithm. The proposed DBN, also called factorial Hiden Markov Model (HMM), can be treated as an audiovisual framework. The factorial HMM arises by forming a dynamic Bayesian belief network composed of several layers. Each of the layers has independent dynamics, but the final observation vector depends upon the state in each layer. Their method was tested in meetings and news videos. Both the proposed method and the method presented in [17] try to improve speaker clustering by utilizing video information. The two methods differ in the way video information is used and also by the fact that [17] applies face recognition whereas the proposed method relies on the results of face clustering. Multimodal speaker diarization is also addressed by Friedland et al. in [10]. The method combines audio and video low level features (unlike the high level features used in the proposed method), by using agglomerative clustering, where Gaussian Mixture Models (GMMs) are used to model the clusters. Garau and Bourlard [12] proposed the use of psychology-inspired visual features to improve the initialization of the agglomerative speaker clustering, in order to perform speaker diarization. Features derived from visual focus of attention and motion intensity features were used. There are two main differences in this method and the proposed one. First, the method proposed in [12] uses a different speaker clustering algorithm and secondly the visual information in this method is used for better initialization of the speaker clustering while the proposed method uses video information to improve the audio similarity matrix. The methods proposed in [10] and [12] were tested on meetings video. Vallet et al. [28] proposed a multimodal diarization system that operates on talk shows data. Their system can be divided in three distinct steps, starting with feature extraction both in audio and video domains, followed by creation of a model for each speaker and, finally, by classification of the speech parts of the talk show. As for feature extraction, the authors used MFCCs to represent the audio signal and features that characterize the clothing of the TV-show participants for the video data. In order to collect training examples, the authors perform shot and lip activity detection and the collected data are clustered to model the different speakers of a show. Finally, an SVM is used to classify the remaining parts of the show. Compared to the proposed method, the method proposed in [28] derives information from both faces and clothing and also uses different features in order to describe this information. Moreover, the method in [28] performes speaker classification in order to label the audio clusters derived by the diarization procedure. It should be mentioned that only the proposed method and the method presented in [7] perform speaker clustering in movies, while all other methods are applied in videos from meetings and TV-shows.

The use of audiovisual information for person recognition/verification (a problem somewhat similar to person clustering) has also been proposed in the literature. Sargin et al. in [23] proposed a method for celebrity recognition in web videos. At first, face tracking and face recognition is performed on video data. Then, speaker segmentation is performed on audio data. Audio and video results are then combined to find consistent one-to-one associations between the face trajectories and the speech segments. Finally, for each recognized face, the biometric characteristics of voice from the corresponding speaker segment are used for speaker verification. Feng et al. proposed a semi-supervised audio-visual method for human recognition [9]. The method includes a new spectral learning algorithm for face recognition. A speaker identification agent is used to improve the proposed human recognition system. This agent is based on Hidden Markov Models (HMM) and models the statistical variations of speech in both spectral and temporal domains.

Another somewhat related field of research where multimodal information can be used is speaker localization, which tries to solve the problem of localizing the current speaker in a scene that contains several people engaged in a multiple-speaker conversation. Friedland et al. [11] proposed a method that performs both speaker diarization and localization that takes into account multimodal data. The proposed method extracts MFCC features from the audio signal in order to detect, segment and cluster the audio signal into speaker clusters. The authors extract block motion vector magnitude features from video. The extracted video features are used to both enhance speaker clustering and also estimate the position of each speaker. In [15], Khalidov et al. proposed a method for speaker localization in 3D space that is applicable to meetings. The authors try to cluster audio-visual observations from a binaural microphone and a stereo camera into coherent groups. Relations between audio and visual observations are found by using a probabilistic generative model. The proposed model uses GMMs to represent the data in the 3D space. The Expectation Maximization algorithm is used to estimate the activity and the 3D position of the speakers.

The use of different data modalities has been also studied in research areas other than diarization and clustering. Many methods consider multimodality as a combination of at least two information channels. These channels are usually visual, auditory or textual. Multimodal approaches that have been used for video retrieval can be found at [4], while Snoek and Worring in [24] present a survey on multimodal methods for video indexing. Moreover, multimodal methods for human-computer interaction are reviewd in [14]. However, multimodality refers not only to the use of visual, auditory and textual information channels but also to the use of different types of data from the same medium. Subramanian et al. in [26] proposed a method that uses different modalities in order to perform personality classification, i.e. prediction of the Extraversion and Neuroticism personality traits in an unstructured and dynamic cocktail-party scenario. They used behavioral features, i.e. proxemic features such as the minimum distance between the target and the rest of the group and the variation of the target's position and social attention features such as the time that the target is focused in the rest of the group alongside with head pose estimation. Pineda et al. exploited multimodal information in [1] in order to classify formations of free-standing conversational groups. They combined video information for head and pose estimation alongside with data from distributed and wearable sensors. They tried to detect F-formations and social attention attractors in free standing conversational group. Two different modalities, video and text were exploited by Yan et al. in [31]. The author used text descriptors, high level concept features and low level video features to perform event detection in a video. Addressing the problem of action recognition, Ohn-Bar et Trivedi [18] used both skeleton data and depth maps derived from Kinect. In general, the use of different modalities increases the available information thus multimodal methods can usually achieve better results than the single modality ones. Additionally, information derived from one modality can improve the analysis of another modality. Indeed, in the proposed method, information derived from face clustering (video modality) is used to improve the audio clustering (audio modality).

3 Method Description

The proposed methods use information derived from video to improve the speaker clustering. In order to combine video and audio information, video facial trajectories and speech segments were used. Video trajectories are series of facial images in consecutive frames (usually depicting the same person) and speech segments are segments where speech has been detected in the audio channel of a movie. The use of visible speech and a new feature that captures the patterns of actors appearing during speech segments were also investigated.

3.1 Audio processing

The first step in speaker diarization is speech detection. Speech segments are detected and subsequently segmented in the audio channel of a video. Then, speaker clustering is performed in order to group together speech segments in clusters that are homogeneous. Each cluster should ideally correspond to a single speaker. The three steps (Fig. 2) of the speaker diarization approach used in this paper are:

- **Speech detection**: using Mel Frequency Cepstral Coefficients (MFCC) features and SVM classifiers
- **Change point detection**: in order to further segment the speech segments to homogeneous parts.
- **Normalized Cut (NCut) clustering**: to group speech segments that belong to the same speaker.

Features are extracted from the segmented speech segments. The audio features used were the Mel Frequency Cepstral Coefficients (MFCC) and the Spectral Flatness Measures (SFM). The feature extraction was done over frames of 40 ms duration with a 20 ms hop size. In speaker diarization, the standard score is based on the Bayesian Information Criterion (BIC) using Gaussian models [6]. The BIC criterion is used to measure how well a model fits to a specific dataset. It is composed of two terms, a log-likelihood term and a negative penalty term that corresponds to the model complexity. Here, the

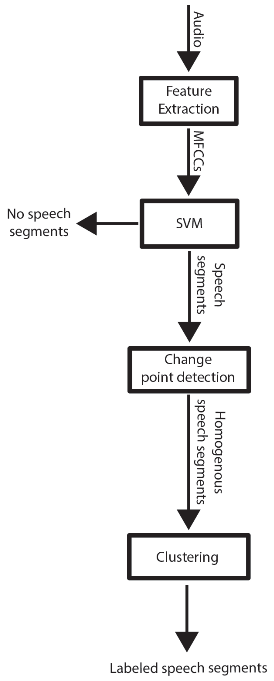


Fig. 2 Flowchart of the audio-based speaker diarization system.

BIC criterion is computed using two Gaussian models, one for each segment in a segment pair (BIC_2). The BIC criterion with one Gaussian model for the union of both segments (BIC_1) is also computed. The distance between two pairs is defined as $\Delta BIC = BIC_2 - BIC_1$. BIC_1 and BIC_2 are defined as:

$$BIC_1 = -log \mid \Sigma \mid -\frac{K}{2}log(N) \tag{1}$$

$$BIC_2 = -\frac{1}{2}log \mid \Sigma_1 \mid = \frac{1}{2}log \mid \Sigma_2 \mid -2\frac{K}{2}log(N) \tag{2}$$

where K is the number of parameters. If the Gaussian model has diagonal covariance, $K = 2p$, where p is the dimension of the data and $K = p + \frac{p(p+1)}{2}$ in the full covariance case. Σ_1 is the covariance matrix on segment 1, Σ_2 is the covariance matrix on segment 2, Σ is the covariance matrix on the union of two segments and N is the total number of segments. A distance matrix $\mathbf{D}$ of dimensions $N \times N$ is derived using the MFCC features and ΔBIC (where N is the number of the audio segments). Both NCut clustering [16] and a novel variant [20] of the spectral clustering proposed in [16] were used for clustering.

3.2 Video processing

The first step in video processing is face detection and tracking. Faces are detected using [25] and tracked in the video channel or channels (in the case of 3D videos) of a movie using the algorithm proposed in [33]. In more detail, each detected face is tracked for K frames. A series of tracked images of a detected face form a facial trajectory (Fig. 3). Each facial trajectory is represented by one of the images included in it and these trajectories are clustered by using their representative images. It is obvious that all faces included in a trajectory belong to the same actor unless tracking error occurs. Faces, rather than full bodies, were selected to represent the actors in video data for two reasons. At first, in dialogues or in general when a person speaks in a movie, the entire body is often not visible since directors prefer close ups or medium close ups in such cases. At second, the face of an actor in a movie is more characteristic than its entire body since actors may change their outfit in different shots while their faces are in general the same. Thus body-based visual clustering would most probably provide worse results than face clustering.

Local Binary Patterns (LBP) [19] were used as features to represent the facial images, for two reasons. First, LBPs can be calculated very fast. Secondly, LBPs have been proven to achieve good performance in face clustering ([21]).

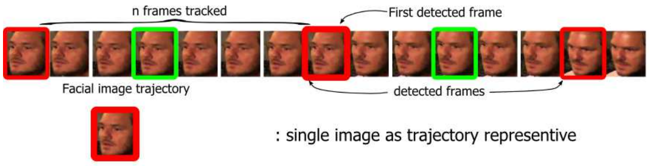


Fig. 3 Face trajectory computed by using face detection and tracking.

The LBP for a pixel (x_c, y_c) with grayscale intensity g_c is defined as:

$$LBP_{P,R}(x_c, y_c) = \sum_{i=1}^{P} s(g_i - g_c)2^{i-1}, \quad (3)$$

where R is the radius of the neighbourhood where the LBP is calculated, P is the number of the pixels of the neighbourhood, $s(z)$ is 1, if $z \geq 0$ and 0 otherwise. LBPs are computed for a number of pixels of a facial image and then a histogram of these LBP values is calculated. This histogram is used as a feature to represent the facial image. Two variations of the LBPs were also used, the CS-LBPs and the tCS-LBPs. The CS-LBP is defined as:

$$CS - LBP_{P,R}(x_c, y_c) = \sum_{i=1}^{\frac{P}{2}} s(g_i - g_{P/2+i})2^{i-1}, \quad (4)$$

where $s(z)$ is defined as in the standard LBP. The tCS-LBP is defined as the CS-LBP but in this case $s(z) = 1$ if $z \geq tm$ and 0 otherwise where $m = \frac{1}{P+1}(g_c + \sum_{i=1}^{P} g_i)$ and $t \in [0.01, 0.1]$.

Calculating LBPs for all pixels of an image is not the best solution neither in terms of effectiveness nor in terms of calculation time. In our case we have chosen to calculate LBPs only in pixels that carry important information (i.e. mouth, eyes, etc.), thus two passes of fiducial points detectors were used. The first one is for the calculation of 66 fiducial points, such as outline of eyes, eyebrows, mouth etc, [2] and the second one [27] for better localization of these points. Moreover, these fiducial points are used in order to scale and align the detected images. LBPs are calculated upon patches around these 66 aligned points. Final, a histogram with K bins is calculated for each of these features. By this way a descriptor of dimension $66 \times K$ is calculated for each image.

In order to perform face clustering, similarities between each pair of images (each image representing a facial trajectory) have to be computed. The χ^2 distance was used to calculate the distances between two corresponding LBP histograms on a pair of images i, j and the final d_{ij} distance value was computed as the sum of the 66 distances (one per histogram). The similarity between the two images was calculated as $1/d_{ij}$ and a similarity matrix $\mathbf{V}$ between facial images (or more precisely facial image trajectories) was computed. Finally the clustering method in [20] is used to perform face clustering by utilizing $\mathbf{V}$. The result of face clustering are used to improve speaker clustering (Section 3.3) and to compute Actor Presence features (Section 3.5).

3.3 Multimodal approach

As can be seen in Sections 3.1 and 3.2, speaker clustering and face clustering group the speakers and the actors in the audio and visual data of a movie respectively. The speakers and the (visible) actors of a movie are in general the same people (people that speak in a movie, usually appear in it also), thus face clustering can improve the speaker clustering i.e. the diarization process.

The input of the algorithm used for multimodal speaker clustering is the similarity matrix of the audio segments. The main idea is to a) increase the similarity of two speech segments, when these segments overlap with visual appearances of the same actor or b) decrease the similarity value, if no such overlap exists. The matrix derived by audio features (Section 3.1) is actually a speech segments distance/dissimilarity matrix, i.e., has small values when two speech segments are similar and high values otherwise. Therefore, the first step towards combining audio and video information was to transform this matrix to a similarity matrix $\mathbf{S}$. This was done by using a sigmoid function:

$$S_{i,j} = \frac{1}{1 + exp(4 * (D_{i,j} - \bar{D})/\sigma)}, \tag{5}$$

where $D_{i,j}$ is an element of the distance matrix $\mathbf{D}$, $\bar{D}$ the mean value of $\mathbf{D}$ and σ the standard deviation of $\mathbf{D}$. To combine information from video and audio, in order to enhance speaker clustering using video, a new matrix $\mathbf{Q}$ is created with dimensions equal to those of the speech similarity matrix $\mathbf{S}$. The next step is to find, for each element (i, j) of the matrix $\mathbf{Q}$, the video trajectories that overlap in time with the speech segments that correspond to this element. Then, if the same actor appears in the corresponding video trajectories, the corresponding element of $\mathbf{Q}$ is increased, otherwise it is decreased. The final similarity matrix $\mathbf{F}$ is formed by combining the speech similarity matrix $\mathbf{S}$ and matrix $\mathbf{Q}$:

$$\mathbf{F} = \mathbf{S} + \alpha\mathbf{Q},\ 0 \leq \alpha \leq 1. \tag{6}$$

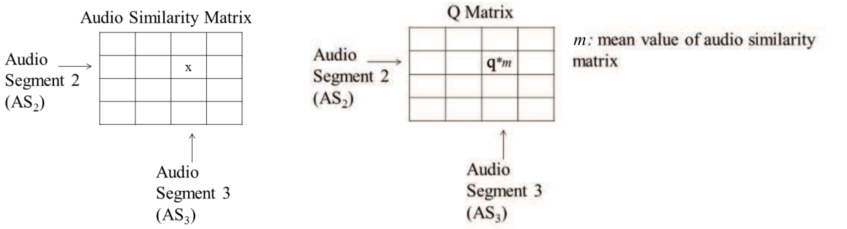


Fig. 4 Example of the creation of matrix $\mathbf{Q}$. Assume that audio segment AS_2 coincides with video trajectory 5 and that AS_3 coincides with video trajectory 9. If the same actor appears in the above pair $(5, 9)$, increase the corresponding value in $\mathbf{Q}$, otherwise decrease.

Two different approaches were implemented and tested (see Section 4), in order to create the matrix $\mathbf{Q}$, i.e., to change the elements of $\mathbf{Q}$ that correspond to speech segments which overlap with video trajectories. In the first approach, the ground truth for the actors depicted in the video trajectories was used, in order to check performance when the face clustering is perfect, i.e., it contains no errors. In more detail, for each pair of audio segments, the overlapping facial trajectories are found and, if the same actor appears in these trajectories according to the ground truth information, then the value in the corresponding element in matrix $\mathbf{Q}$ is multiplied with q where $q > 1$, otherwise it is multiplied with $1/q$. In the second more realistic approach, the same procedure is used, but instead of using the ground truth for the actors, the results of the face clustering algorithm are used. In other words, the results of face clustering described in Section 3.2 are used to check if the same actor appears in the overlapping facial trajectories.

Finally, after the calculation of matrix $\mathbf{F}$ using (6), the clustering algorithm in [16] or in [20] are used for speaker clustering.

3.4 Use of visual speech

A variant of the method described in Section 3.3 that utilizes the results of visual speech detection was also investigated. Visual speech detection (detection of persons that speak, using only visual information) was used in order to find out in which facial trajectories the depicted person is speaking. In such trajectories, it is likely that the person detected with visual speech is the speaker in the overlapping audio segment. The intuition for the use of visible speech is that the information of facial trajectories that overlap with a speech segment should be taken into account in the method described in Section 3.3 only when visible speech is detected in these trajectories. Thus only the facial trajectories where visible speech has been detected were taken into account in the algorithm of Section 3.3. The remaining facial trajectories were ignored.

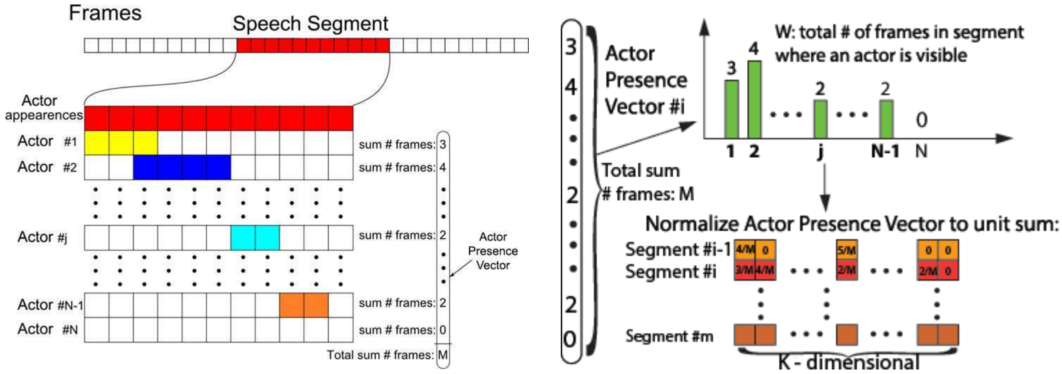
The problem of visual speech detection was addressed using the approach in [22], i.e. as an action recognition problem by utilizing relevant video descriptors and classification techniques. More specifically, video description via dense trajectories was adopted and classification was performed using Kernel ELMs. In more detail, video representation adopted the state-of-the-art Dense Trajectories [29] approach applied only on facial regions in videos. Features extracted using Dense Trajectories were processed with a Bag of Features [30] approach and transformed into a histogram of standard size per video. Classification was performed using a Kernel Extreme Learning Machine (KELM) with χ^2 kernel [13]. The result of visual speech detection is a label for each facial trajectory, that shows whether the actor appearing in this trajectory talks ($Label = 1$) or not ($Label = 0$). As already mentioned only, the trajectories with $Label = 1$ were taken into account in the multimodal clustering approach described in Section 3.3.

3.5 Speaker clustering using actor appearance features

In an effort to use features that utilize visual actor/person information so as to aid speaker clustering, a new type of video-based features was introduced. These features are computed for each speech segment. The idea behind them is that each speech segment which consists of N frames has potentially one or more visual/face trajectories occurring in the same temporal interval. For each actor in the video, the number of frames of the speech segment under examination, where he/she appears is evaluated and divided (normalized) by the total number of speech segment frames in which at least one actors appears in the video. This results in one value per actor for each speech segment. Thus, a feature (one per speech segment) called Actor Presence (AP) feature which is a vector whose dimensionality is equal to the number of actors K in the film is evaluated. In more detail, the AP feature for an audio segment x can be defined as:

$$\mathbf{p}_x = \left[p_1, p_2, \ldots, p_K\right], p_i = \frac{k_i}{W}, i = 1, \ldots, K \tag{7}$$

where k_i the number of frames that actor i appears during the audio segment x and W the number of audio segment frames where at least one actor appears in the video. The method used to calculate and normalize those features is depicted in Fig. 5. A similarity matrix A for the AP features is computed using the formula:



(a) (b)

Fig. 5 a) AP feature creation. b) Normalization of AP features.

$$\mathbf{A} = \mathbf{P}\mathbf{P}^T \tag{8}$$

where $\mathbf{P} \in \Re^{N \times K}$ is a matrix containing the AP features of the N audio segments of a movie. The final similarity matrix that is used for speaker clustering is a linear combination of the audio (speech segment) similarity matrix $\mathbf{S}$ and $\mathbf{A}$ as in Section 3.3:

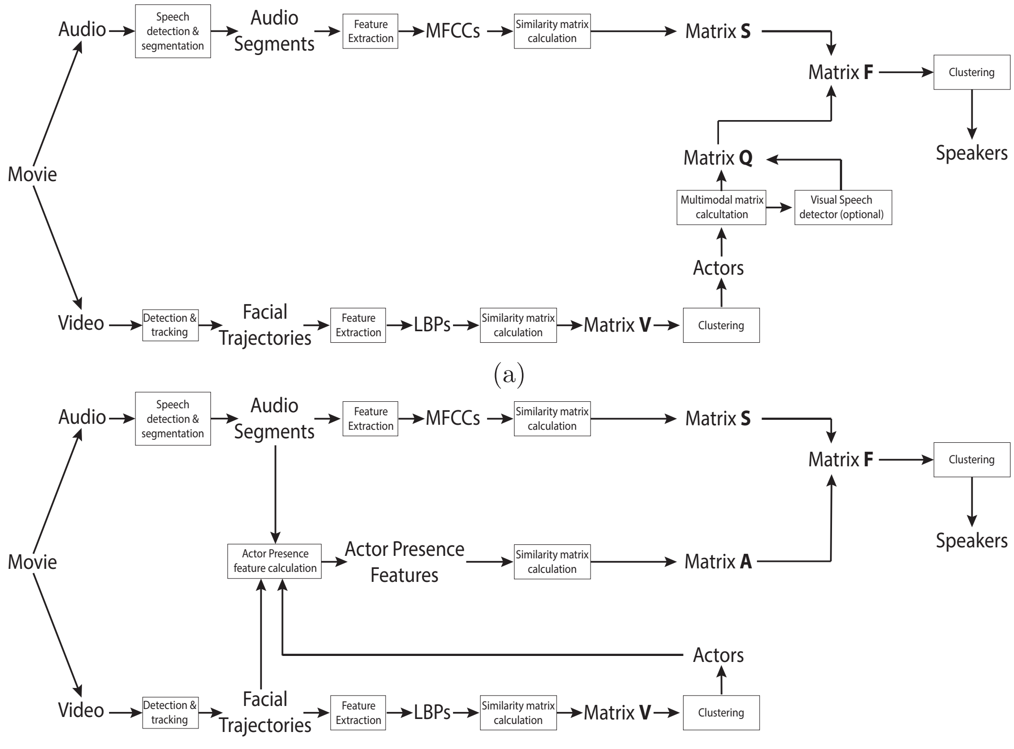
$$\mathbf{F} = \mathbf{S} + \alpha\mathbf{A}, 0 \leq a \leq 1 \tag{9}$$

It should be noted that, in a significant number of speech segments, there is no visual appearance of actors at all. For these segments, the AP feature will be an all-zeros vector and, thus, will bear no useful information. For these segments their similarity with any other segment (contained in matrix $\mathbf{A}$) will be zero. Finally, matrix F is used for speaker clustering by applying again one of the algorithms in [20] or in [16].

Two block diagrams that show in detail the proposed methods for multimodal speaker diarization are shown in Fig. 6.

4 Experimental Results

The evaluation of the proposed multimodal speaker clustering approaches was made by using a modified F-measure. This is because F-measure punishes the erroneous split of a class into 2 parts quite strictly. In the modified version of F-measure used in this paper, overclustering is performed by creating more than the needed clusters and then the clusters that correspond to the same speaker


Audio
Speech detection & segmentation
Audio Segments
Feature Extraction
MFCCs
Similarity matrix calculation
Matrix S
Matrix F
Clustering
Speakers
Matrix Q
Multimodal matrix calcultation
Visual Speech detector (optional)
Actors
Movie
Video
Detection & tracking
Facial Trajectories
Feature Extraction
LBPs
Similarity matrix calculation
Matrix V
Clustering
Actor Presence feature calculation
Actor Presence Features
Similarity matrix calculation
Matrix A

(a)

(b)

Fig. 6 Flowchart of the proposed methods. Matrix **F** is evaluated by summing either (a) matrices **S** and **A** (Section 3.5) or (b) matrices **S** and **Q** (Sections 3.3 and 3.4)

are merged. The final F-measure is evaluated upon this merged clustering result. By this way, F-measure becomes less strict in the evaluation of splitted classes and evaluates more the purity of clusters.

The proposed approaches were tested in three full length 3D feature films of different duration, size of cast and genre. These movies were selected in order to test the proposed approaches in a difficult and realistic scenario. Stereo information of the video channels was exploited in two ways. Face detection [25] was applied on both channels (left and right), mismatches between the two channels were rejected and a stereo tracking algorithm [33] was applied in both channels. By using the above approaches we end up with a number of facial trajectories, namely series of consecutive facial images. As previously stated, each of these stereo trajectories is represented by a single facial image for each stereo channel (Left-Right). The similarity between the two stereo facial trajectories T_n, T_k is evaluated by evaluating the LBP-based similarity for all pairs of facial images, namely $(\mathbf{X}_n^R, \mathbf{X}_k^L)(\mathbf{X}_n^L, \mathbf{X}_k^L)(\mathbf{X}_n^L, \mathbf{X}_k^R)(\mathbf{X}_n^R, \mathbf{X}_k^R)$ and keeping the biggest one, where L, R denote the Left and Right channel.

The number of speech segments and video trajectories for each movie alongside with the number of speech segments, that have overlapping facial trajectories is presented in Table 1. As can be seen, only a relatively small number of speech segments overlapping with facial trajectories, which is a usual phenomenon in movies and makes multimodal diarization difficult in such content.

Table 1 Number of speech segments and facial trajectories alongside with the number of the speech segments that overlap with facial trajectories.

	Speech segments	**Face trajectories**	**Segments that have overlap with trajectories**	**%**
Movie 1	979	1171	587	50.13
Movie 2	890	1146	619	54.01
Movie 3	1318	1541	657	42.63

Experiments have been conducted to verify the performance of the 3 multimodal speaker clustering methods that are described in Section 3. The results for method in Section 3.3 can be seen in Table 2 alongside with the performance of the clustering when only audio modality was taken into account. Results are presented for both the clustering methods in [16] and [20]. As can be seen in this Table, the use of video ground truth information for the actors depicted in each facial trajectory (Multimodal 1 column) improves the clustering performance by approximately 8% in every movie, in terms of the modified F-measure compared to audio only diarization. Since the ground truth was used, it can be deducted that this is the best possible improvement for speaker clustering by using the video information with the approach described in Section 3.3. When using information derived from actual facial image clustering in video the improvement (column Multimodal 2), is approximately 5%. Face clustering results (i.e., when only video information was taken into account) [20] are show in Table 3. As can be seen in Tables 1, 2 and 3, movie 1, that has the highest overlap between speech segments and face trajectories and a

Table 2 Speaker Clustering F-measure, when video information is incorporated.

	Audio only		**Multimodal 1**		**Multimodal 2**	
Clustering	[20]	[16]	[20]	[16]	[20]	[16]
Movie 1	0.51	0.51	0.59	0.60	0.56	0.57
Movie 2	0.48	0.48	0.57	0.57	0.51	0.52
Movie 3	0.45	0.46	0.53	0.55	0.51	0.50

Table 3 Face clustering results in the 3 movies.

	F-measure
Movie 1	67.96
Movie 2	56.73
Movie 3	54.47

high face clustering accuracy, achieves the highest increase in speaker clustering accuracy when video information is taken into account. On the other hand, Movie 2, that has slightly smaller overlap between speech segments and trajectories and a low face clustering accuracy, achieves the lowest increase in speaker clustering accuracy when video information is taken into account.

Experiments of the variant that exploits visible speech detection (Section 3.4) were also conducted. F-measure for the 3 movies when visible speech was taken into account are shown in Table 4. As can be seen in this Table, even if the use of visual speech increases the F-measure with respect to the results of the single modality speaker clustering, the increase of the F-measure is not as high as when video information is used without visible speech detection (compare Table 2). These results can be explained by the fact that the accuracy of visible speech detection algorithm used was good but not excellent due to the fact that it was applied in challenging data (faces from feature movies). Indeed in Table 5, visible speech detection accuracy for the three movies is shown. One can assume by looking at this table that visible speech detection will probably help more the multimodal clustering when its detection accuracy gets higher.

Table 4 Speaker Clustering F-measure, when video information and visible speech are incorporated.

	Audio only		**Multimodal 1**		**Multimodal 2**	
Clustering	[20]	[16]	[20]	[16]	[20]	[16]
Movie 1	0.51	0.51	0.56	0.59	0.53	0.55
Movie 2	0.48	0.48	0.55	0.58	0.49	0.53
Movie 3	0.45	0.46	0.53	0.53	0.49	0.48

Table 5 Visible speech detection accuracy.

	Accuracy
Movie 1	74.2
Movie 2	72.1
Movie 3	73.7

In the last of the proposed methods for multimodal speaker clustering (Section 3.5) the Actor Presence features were used. Similar to the method in Section 3.3, two different approaches for calculating the AP features were used. In the first approach (Multimodal 1) the ground truth of the actors depicted in the facial trajectories was used for the evaluation of the AP features whereas in the second approach the actual face clustering results obtained by the method in [20] were used (Multimodal 2). As can be seen in this table, the use of AP features for multimodal speaker clustering achieves better results that the single modality clustering in all three movies. Compared to the method in Section 3.3, the use of AP features leads to better results when using ground

Table 6 Speaker Clustering F-measure, when Actor Presence features are used.

	Audio only		**Multimodal 1**		**Multimodal 2**	
Clustering	[20]	[16]	[20]	[16]	[20]	[16]
Movie 1	0.51	0.51	0.59	0.61	0.54	0.57
Movie 2	0.48	0.48	0.60	0.60	0.51	0.49
Movie 3	0.45	0.46	0.57	0.57	0.51	0.50

truth for their computation. This is because AP features capture information for all the actors that appear during a speech segment, thus they can provide additional information to the speaker clustering and improve the results. When results from face clustering are used, the use of AP features achieve slightly better results in movies 3 and 1 and the same results in movie 2. It can be deducted that the AP features are more sensitive in the errors of the face clustering results.

As can be seen from the experimental results, information derived from video data can help the audio-based speaker diarization. The increase in performance is higher when ground truth information is used, which leads to the obvious conclusion that the better the face clustering in the video, the better the effect of multimodal speaker clustering in the speaker diarization. It should be noted that, face clustering is not the only way to cluster the actors facial images in a video. Face recognition or label propagation can also be used to cluster the actors to groups and use this information for multimodal speaker clustering.

Experiments were also conducted in order to display the effect of the parameters used in the proposed method. Two parameters are used in the proposed method, α in Eq. 6 and Eq. 9 and q in the creation of $\mathbf{Q}$ matrix. As can be seen in Eq. 6 and Eq. 9, parameter α controls the impact of the face clustering results in the creation of $\mathbf{F}$ matrix, for $\alpha = 0$ the method is using only audio information. The impact of this parameter in the speaker diarization of movie 1 can be seen in Fig. 7. The best result for this movie is achieved when $\alpha = 0.9$. This is because face clustering results are good in movie 1, and thus the best results of multimodal clustering are achieved with a high value in α. However the effect of α in the method performance is quite limited since the F measure for α values from 0.1 to 1 ranges from 0.55 to 0.57.

Parameter q has impact in the creation of $\mathbf{Q}$ matrix. In more detail, q determines the increase/decrease of the similarity between two audio segments. The impact of this parameter on the diarization of movie 1 can be seen in Fig. 8. It is obvious that different values of q lead to similar results, in more detail the F-measure for q values from 1.5 to 6 ranges from 0.54 to 0.57. This is an expected result, since all similarities between audio segments that have overlaps with face trajectories increase or decrease by the same amount, determined by q. In conclusion, α, q have little impact and thus their selection is not critical. What affects most the proposed method is the quality of the face clustering. The better the face clustering, the better the multimodal diarization of a movie.

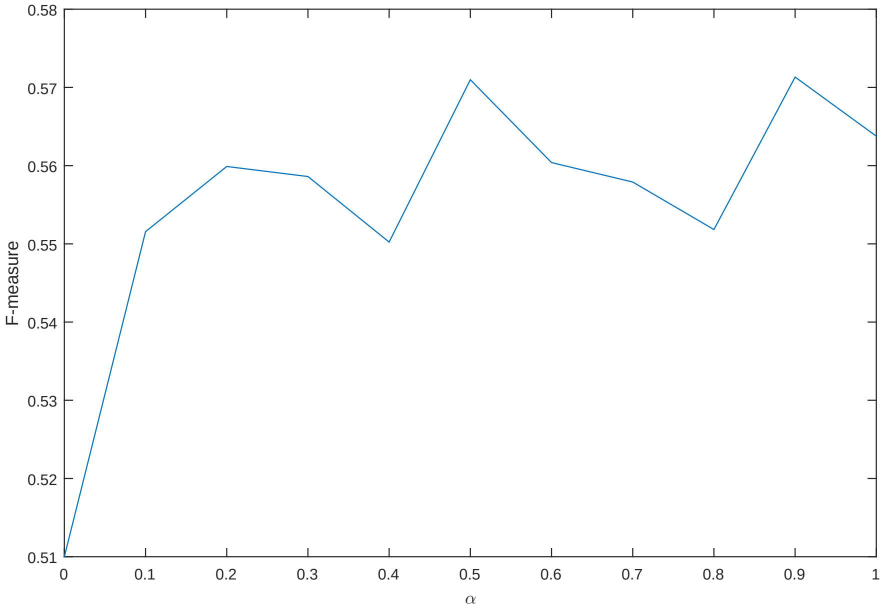


Fig. 7 Impact of the α parameter in the speaker clustering in movie 1.

The proposed method was developed in order to perform speaker diarization in full length 3D movies, however in order to be able to compare its performance with other methods, it was applied on a meetings video. In more detail, in order to asses its performance the method was applied in an IDIAP meeting of the Augmented Multimodal Interaction (AMI) data set [5] and compared with the method proposed in [17]. It should be noted that, as mentioned in Section 1, the proposed method is designed for movies and thus is based on the assumption that the person that speaks in a movie is usually the visible actor whereas other actors that participate in the dialoque are not visible. However this assumption does not apply in videos from meetings where there is a camera for every participant or all participants are visible. The basic assumption of the proposed method mentioned above is indeed not true in the IDIAP meeting video since there is a camera for each one of the four participants and all participants are visible all the time. Therefore, there is only a slight improvement of less than 2% of the clustering results when video information is taken into account. The diarization results alongside with a comparison with [17] can be seen in Table 7. The clustering results in this table are evaluated with the metric used in [17], i.e. each frame of the video is annotated with a label and the overall accuracy is computed. As can be seen in this table the proposed method achieves better results when audio only information is taken into account, but fails to increase the clustering results as much as the method in [17] when multimodal information is exploited.

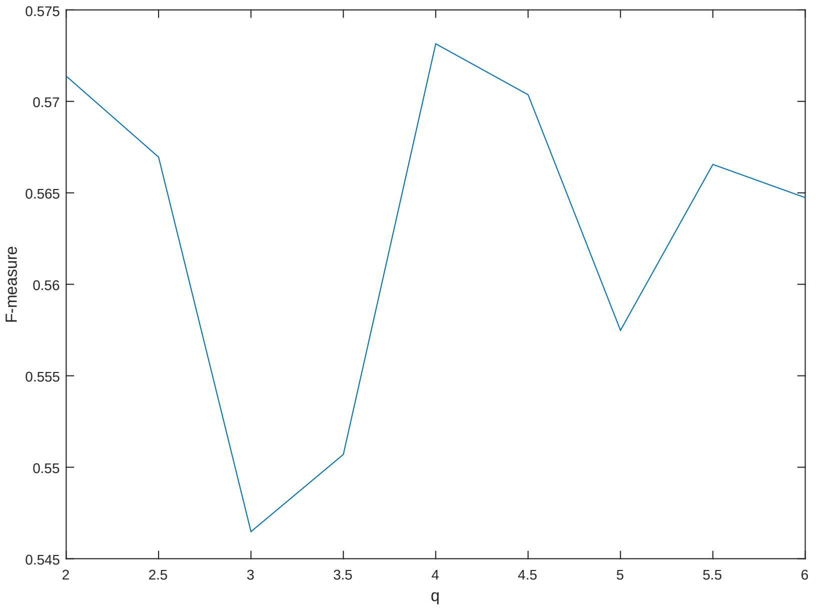


Fig. 8 Impact of the q parameter in the speaker clustering in movie 1.

Table 7 Speaker Clustering overall accuracy in the IDIAP meeting.

	Audio only	**Multimodal**
Proposed method	**72.54**%	74.02%
[17]	67%	**84**

The variant presented in this table is the one that uses multimodal clustering (Multimodal 2) along with Visible Speech since it achieved the best result in the case of IDIAP meeting. The two methods achieve different levels of improvement in the multimodal case for another reason, besides the fact that the proposed method is not suitable for meeting data. The method proposed in [17] uses face recognition to group the visible persons instead of face clustering. Face recognition achieves in general better results than face clustering, but it requires training data.

5 Conclusion

In this paper we proposed methods to improve speaker diarization through a multimodal approach. The improvement of speaker clustering can be done by using video information derived from video data through face clustering. Experiments in three full stereo movies have shown that multimodal speaker clustering achieves better results that single modality speaker clustering. However, when tested on meetings video the method fails to considerable increase the performance, since it is not designed for meeting like scenarios

Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287674 (3DTVS). This publication reflects only the authors views. The European Union is not liable for any use that may be made of the information contained therein.

References

1. Alameda-Pineda, X., Yan, Y., Ricci, E., Lanz, O., Sebe, N.: Analyzing free-standing conversational groups: A multimodal approach. In: Proceedings of the 23rd ACM International Conference on Multimedia, MM '15, pp. 5–14. ACM, New York, NY, USA (2015)
2. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3444–3451 (2013)
3. Baltzakis, H., Argyros, A., Lourakis, M., Trahanias, P.: Tracking of human hands and faces through probabilistic fusion of multiple visual cues. In: Proceedings of the 6th International Conference on Computer Vision Systems, ICVS'08, pp. 33–42. Springer-Verlag, Berlin, Heidelberg (2008)
4. Calic, J., Campbell, N., Dasiopoulou, S., Kompatsiaris, Y.: A survey on multimodal video representation for semantic retrieval. In: The International Conference on Computer as a Tool, 2005. EUROCON 2005., vol. 1, pp. 135–138 (2005)
5. Carletta, J.: Announcing the ami meeting corpus. The ELRA newsletter **1**(1), 3–5 (2006)
6. Chen, S., Gopalakrishnan, P.: Speaker, environment and channel change detection and clustering via the bayesian information criterion. In: Proceedings of DARPA Broadcast News Transcription and Understanding Workshop (1998)
7. El Khoury, E., Snac, C., Joly, P.: Audiovisual diarization of people in video content. Multimedia Tools and Applications **68**(3), 747–775 (2014)
8. Elmansori, M.M., Omar, K.: An enhanced face detection method using skin color and back-propagation neural network. European Journal of Scientific Research **55**(1), 80 (2011)
9. Feng, W., Xie, L., Zeng, J., Liu, Z.Q.: Audio-visual human recognition using semi-supervised spectral learning and hidden markov models. Journal of Visual Languages and Computing **20**(3), 188–195 (2009)
10. Friedland, G., Hung, H., Yeo, C.: Multi-modal speaker diarization of real-world meetings using compressed-domain video features. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009., pp. 4069–4072 (2009)
11. Friedland, G., Yeo, C., Hung, H.: Visual speaker localization aided by acoustic models. In: Proceedings of the 17th ACM International Conference on Multimedia, MM '09, pp. 195–202. ACM, New York, NY, USA (2009)
12. Garau, G., Bourlard, H.: Using audio and visual cues for speaker diarisation initialisation. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 4942–4945 (2010)
13. Iosifidis, A., Tefas, A., Pitas, I.: On the kernel extreme learning machine classifier. Pattern Recognition Letters **54**(0), 11 – 17 (2015)
14. Jaimes, A., Sebe, N.: Multimodal human computer interaction: A survey. In: Computer Vision in Human-Computer Interaction, *Lecture Notes in Computer Science*, vol. 3766, pp. 1–15. Springer Berlin Heidelberg (2005)
15. Khalidov, V., Forbes, F., Hansard, M., Arnaud, E., Horaud, R.: Audio-visual clustering for 3d speaker localization. In: Proceedings of the 5th International Workshop on Machine Learning for Multimodal Interaction, MLMI '08, pp. 86–97. Springer-Verlag, Berlin, Heidelberg (2008)
16. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Proceedings of NIPS, pp. 849–856. MIT Press (2001)
17. Noulas, A., Englebienne, G., Krose, B.: Multimodal speaker diarization. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(1), 79–93 (2012)
18. Ohn-Bar, E., Trivedi, M.M.: Joint angles similiarities and HOG^2 for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops: Human Activity Understanding from 3D Data, CVPR '13. IEEE PRESS (2013)
19. Ojala, T., Pietikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: Proceedings of

the 12th IAPR International Conference on Pattern Recognition, vol. 1, pp. 582–585 vol.1 (1994)

20. Orfanidis, G., Tefas, A., Nikolaidis, N., Pitas, I.: Facial image clustering in stereo videos using local binary patterns and double spectral analysis. In: IEEE Symposium Series on Computational Intelligence (SSCI) (2014)
21. Orfanidis, G., Tefas, A., Nikolaidis, N., Pitas, I.: Facial image clustering in stereoscopic videos using double spectral analysis. Signal Processing: Image Communication **33**, 86 – 105 (2015)
22. Patrona, F., Iosifidis, A., Tefas, A., Nikolaidis, N., Pitas, I.: Visual voice activity detection based on spatiotemporal information and bag of words. In: IEEE International Conference on Image Processing, ICIP 2015 (2015)
23. Sargin, M., Aradhye, H., Moreno, P., Zhao, M.: Audiovisual celebrity recognition in unconstrained web videos. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009., pp. 1977–1980 (2009)
24. Snoek, C.G.M., Worring, M.: Multimodal video indexing: A review of the state-of-the-art. Multimedia Tools and Applications **25**(1), 5–35 (2005)
25. Stamou, G., Krinidis, M., Nikolaidis, N., Pitas, I.: A monocular system for person tracking: Implementation and testing. Journal on Multimodal User Interfaces **1**(2), 31–47 (2007)
26. Subramanian, R., Yan, Y., Staiano, J., Lanz, O., Sebe, N.: On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, ICMI '13, pp. 3–10. ACM, New York, NY, USA (2013)
27. Uricar, M., Franc, V., Hlavc, V.: Detector of facial landmarks learned by the structured output svm. In: Proceedings of VISAPP 2012, pp. 547–556 (2012)
28. Vallet, F., Essid, S., Carrive, J.: A multimodal approach to speaker diarization on tv talk-shows. IEEE Transactions on Multimedia. **15**(3), 509–520 (2013)
29. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011. CVPR 2011., pp. 3169–3176. IEEE (2011)
30. Wang, H., Ullah, M., Kläserr, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC 2009-British Machine Vision Conference (2009)
31. Yan, Y., Yang, Y., Meng, D., Liu, G., Tong, W., Hauptmann, A., Sebe, N.: Event oriented dictionary learning for complex event detection. IEEE Transactions on Image Processing **24**(6), 1867–1878 (2015)
32. Zoidi, O., Nikolaidis N.and Tefas, A., Pitas, I.: Stereo object tracking with fusion of texture, color and disparity information. Signal Processing: Image Communication **29**(5), 573 – 589 (2014)
33. Zoidi, O., Nikolaidis, N., Pitas, I.: Appearance based object tracking in stereo sequences. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)., pp. 2434–2438 (2013)