Shot Classification and Keyframe Detection for Vision based Speakers Diarization in Parliamentary Debates*

Pedro A. Marín-Reyes[1], Javier Lorenzo-Navarro[1], Modesto Castrillón-Santana[1], and Elena Sánchez-Nielsen[2]

[1] Instituto Universitario SIANI, 35017 Las Palmas, Spain,
Universidad de Las Palmas de Gran Canaria
[2] Departamento de Ingeniería Informática y de Sistemas, 38271 Santa Cruz de Tenerife, Spain
Universidad de la Laguna
pedro.marin102@alu.ulpgc.es

Abstract. Automatic labelling of speakers is an essential task for speakers diarization in parliamentary debates given the huge amount of video data to annotate. In this paper, we address the speaker diarization problem as a visual speaker re-identification issue with a special emphasis on the analysis of different shot types. We propose two approaches that makes use of convolutional neural networks (CNN) and biometric traits for keyframe extraction. Experimental results have been evaluated with challenging real-world datasets from the Canary Islands Parliament, and contrasted with a similar approach that does not analyze the shot type. Results show that the use of CNN for shot classification and biometric traits help to improve the performance of the re-identification outcomes in an average rate of 9.8%.

Keywords: visual diarization, re-identification, CNN classification, biometric traits.

1 Introduction

Speaker diarization is a common topic for the speech research community. The aim is to identify the number of participants and creation of the list of time intervals of each participant speech, e.g. "who spoke when" [1,15]. Although this problem has received the interest of the speech processing community, just recently the use of visual features has been considered to strength the performance of audio-only diarization systems [5,6,9,12,16].

In [5] the scenario is restricted to a meeting. The diarization is done using an agglomerative clustering method using Mel Frequency Cepstral Coefficients (MFCCs), head pose and motion intensity. Vallet et al. [16] also employ an

* This work has been partially supported by the Spanish Government under the projects TIN2011-24598 and TIN2015-64395-R.

agglomerative clustering in a talk-show scenario. They use as visual features HSV color components cumulative histograms of the clothes for shots presenting lip movement. In a recent work, Sarafianos et al. [12] implement a semi-supervised variant of the Fisher Linear Discriminant Analysis named FLsD. Gabor based features are extracted after a face detection and normalization stage. Feature reduction is applied in FLsD followed by a C-means clustering process. In another recent work [6], also a fusion of audio and visual features is employed. In this case, the visual features are based on Local Binary Patterns (LBP) and two variants Center-Symmetric LBP (CS-LBP) and Thresholded CS-LBP (tCS-LBP).

Unlike previous works that rely on the combination of audio and video features to perform the diarization process, in this work we focus only on the use of visual information. Debates in the Canary Islands Parliament is the chosen scenario, that although is a well defined scenario, it poses some challenging situations that the system must cope with. The contribution of the paper is twofold. First, the identification of different shot types with the use of a Convolutional Neural Network that allows to implement the proper strategy to identify the speaker without the sound cue. Second, the proposal of a measure based on anthropometric relations of facial elements to discard non frontal faces that introduce noise in the diarization process. To validate the proposals, they have been tested on 31 videos that add up more than 100 hours.

1.1 Parliamentary sessions scenario

The parliamentary sessions scenario is challenging given that the recording does not provide a single field of vision focused on each speakers intervention. Instead of it, several and different views of the Parliament are captured by a camera network including different individuals and changes in pan, tilt and zoom. This scenario is also characterized by clothing similarities among speakers, changing lighting conditions and automatic color adjustment during speakers speeches, viewpoint variations across camera views when a speaker is giving the speech, cluttered background and occlusions. Given our aim is to roughly label in each time interval speaker apparence, the developed system must know the different types of shots in order to process only valid shots and avoid redundant computation.

Recent computer vision literature is rich in people detection approaches [14]. There are different visual patterns that have been taken into account for that purpose: the face/head, the upper body, the entire body, or just the legs. For our scenario, even if the speaker will be looking at the audience instead of the camera, his/her pose will be typically frontal. Hence, the speaker could be standing surrounded by the audience while they are sitting near him/her. Therefore, face and upper body detectors fit the problem restrictions depending upon the shot type (Fig. 1).

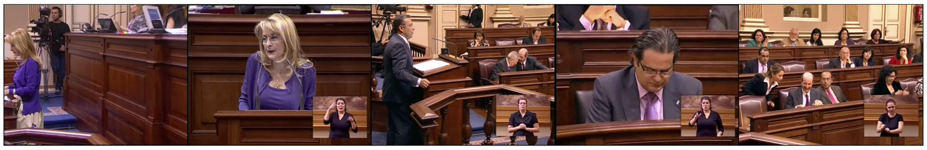

Fig. 1. Different camera views during a parliamentary session.

2 Methodology

The proposed system is composed by six modules, see Fig. 2. Input frames feed a shot detector which determinates if the frame is a new shot. Shots are classified into four types (Fig. 3), considering only the two leftmost of interest. The image is processed by a upper body detector if the shot satisfies some conditions, returning a new cropped image. After that, a face detector is used to identify the area of the speaker and the position of eyes and mouth. These metrics are computed in order to verify that the area of the frame corresponds to a real face. This area is modelled by visual features and then the label of the most similar speaker is given if it is similar enough, or a new label is created.

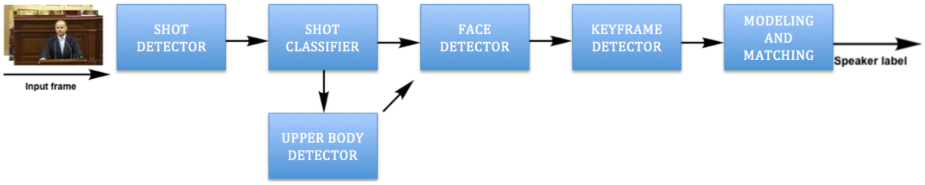

Fig. 2. System pipeline overview.

2.1 Shot boundary detector

Shot boundary detection is a required task for automatic video indexing. Their detection provides semantics about the video stream processed. Different techniques have been described in the literature mainly based on statistics computation and the definition of a threshold between frames. In this work, the shot boundary detection method presented in [11] is used. The method is based on the comparison of consecutive frames using the Kullback-Leibler (KL) divergence [4] between the HSV histograms of the frames.

2.2 Shot classification

In a parliamentary debate, deputies can participate from their own seat or from the platform. We call *medium close up* a shot where only the speaker appears

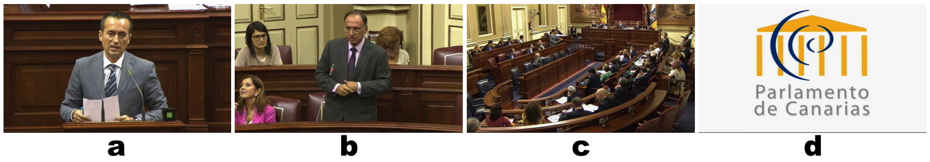

Fig. 3. Shot types. a) Medium close up. b) Mid shot. c) Long shot. d) Others.

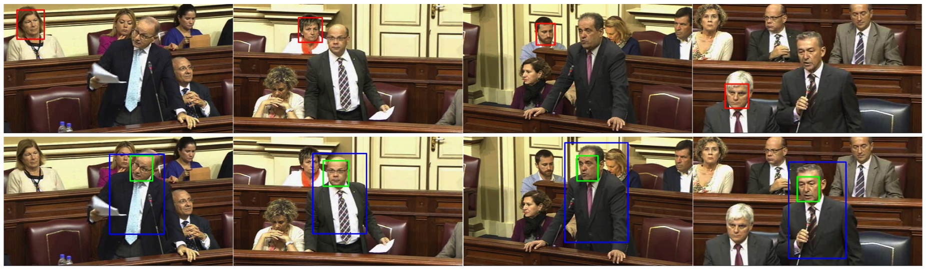

Fig. 4. Face detection samples using face and upper body detectors.

and the face pose is mainly frontal. The second type is called *mid shot*, where a speaker is the main subject of the frame but it can be surrounded by other deputies. The other two types corresponds to *long shot* that are general views of the parliament and *others* which are ratings and titles.

In order to differentiate among the four types of shots under consideration a convolutional neural network (CNN) [8] has been trained. The architecture of the CNN is as follows. The input to the network is a 227×227 RGB image. Then, three convolutional layers each one with a ReLU activation function, followed by a maximum pooling and a local response normalization stage. The sizes are are 96, 256 and 384 respectively. The two next layers correspond to fully connected both with size of 512, each of these levels has ReLU activation function and a dropout phase. The last layer is a fully connected layer with 4 outputs.

2.3 Speaker detection

As mentioned above, the scenario configuration allows the system to introduce some restrictions in the kind of shots that belong to the speaker exposition. Long shots are excluded from the detection process because normally they correspond to a general view. Face detector is used for medium close up shots due this kind of shots are close and only the speaker appears. Mid shots introduce a limitation to face detecton based systems, e.g. [10] because face detectors localize all the faces of the image and the biggest detected one could not correspond to the speaker due to the angle of the camera. Some samples of this problem are shown in Fig. 4. As commented before, in this kind of shots the speaker is standing. This fact can be detected with an upper body detector obtained as region of interest only the standing person area. Fig. 4 shows some frames where the introduction of the standing person detection (in blue) has removed the false speaker detection obtained with only a face detector. After upper body detection, a face detector is used on this region. See Algorithm 1 for a brief description of the method.

Algorithm 1 Speaker face detector algorithm

```
$shotType \leftarrow$ classifyShot($frame$)
$faceRect \leftarrow []$
if $shotType =$ MediumCloseUp then
    $faceRect \leftarrow$ faceDetector($frame$)
if $shotType =$ MidShot then
    $upperBodyRect \leftarrow$ detectorUpperBody($frame$)
    $frameRegion \leftarrow$ crop($frame, upperBodyRect$)
    $faceRect \leftarrow$ faceDetector($frameRegion$)
return $faceRect$
```

The upper body detector [2] is used to detect standing speakers in *mid shots*. On the other hand, to detect faces we have made use of Viola-Jones face detector [17]. The areas of the images detected as faces are validated by means of the detection of both eyes and mouth. Additionally, it is checked that distance between eyes and distance between the middle point of eyes and mouth correspond to a real face.

2.4 Biometric keyframe extraction

In [10] all faces detected in a shot are considered to label the speaker because a majority voting approach was used. However, some non-frontal faces can be detected and this introduces noise in the process. To alleviate this fact, the detection of keyframes is considered in this work.

Keyframe is the frame that represents the relevant content of the shot. It reduces the amount of images that the system has to process and it deletes possible noise errors. There are diferent methods to extract the keyframe [13] such as visual frame descriptors, motion attention model or camera motion and object motion. We propose a biometric keyframe based on the facial element interdistances, distance of eyes and distance between the middle point of eyes and mouth. Statistically we analyzed the influence of these metrics to define a coeficient, eq. (1), that represent a non-dimensional measure.

$$c = \frac{D_{eyes}}{D_{eyes/mouth}} \tag{1}$$

where c represent dimensionless relational coefficient between eyes distance (D_{eyes}) and the distance from the middle point of the eyes respect with the mouth

($D_{eyes/mouth}$). The following decision rule is implemented according to eq. (2)

$$frame_i \text{ is keyframe} \begin{cases} \text{true} & \text{if } shotType \text{ is MediumCloseUp and } 0.76 \leq c \leq 0.82 \\ \text{true} & \text{if } shotType \text{ is MidShot and } 0.83 \leq c \leq 0.89 \\ \text{false} & \text{otherwise} \end{cases} \tag{2}$$

where $frame_i$ corresponds to each video frame and $shotType$ is the type of shot of the frame.

2.5 Speaker modeling

Once the face of the speaker is detected, three areas of interest are considered to model her/him. One of those areas is the face where Histograms of Oriented Gradients (HOG) are computed using a 3×3 grid to obtain nine HOG cell histograms. Another area of interest is the one surrounding the head that carries out information about hair styles and can introduce a discriminant element between speakers with similar faces. In this area, also a HOG is computed but as the information is coarser than in the face a 2×2 grid is defined. Finally, the color of the clothes is also used to model the speaker given the fact that during a debate session the deputies wear the same outfits. This is done with the YCbCr color components histogram of the region just under the face because it is always visible both in medium close up and mid shots.

The matching process between speakers is done using the previous described visual features. As the nature of the visual features extracted from the individual are different, two similarity measures are used in the matching. The comparison of the HOG features is done with the cosine distance and the comparison between color histogram is calculated with the KL divergence.

The approach proposed by Sánchez et al. [10] for parliamentary debate scenarios is not based on clustering the different detected speakers after recording. Instead they realize an on-the-fly assignment to previously seen speakers, or create a new label for different enough individuals. This is done by combining the three above mentioned matching measures into a decision rule to create a new label or assigning to an existing one. Re-identification techniques are considered based exclusively on visual features extracted from the upper body. Most computer vision identity modeling approaches are based on the face pattern, working with identity models that are previously pre-computed based on the image [18] or facial descriptors [7].

3 Experiments

This section presents the experimental evaluation of the approaches proposed in this paper. The video dataset consists of 31 videos extracted from `http://www.parcan.es/video/canales.py`. Table 1 summarizes the main details for each specific video. In our experimental evaluation, we compared three approaches for speaker detection. The first one, taken as baseline, is the method described in

Video				Measures results per method					
Features				Baseline		DSC		DSCK	
Id	Frames	Shots	Speakers	TRR	TDR	TRR	TDR	TRR	TDR
2770	314050	660	8	100.0	100.0	100.0	100.0	93.8	100.0
2785	242850	325	32	40.0	99.6	50.0	99.4	100.0	100.0
2786	265500	396	17	75.0	100.0	76.9	100.0	80.0	100.0
2787	464000	738	24	80.8	99.2	88.5	99.2	79.4	97.8
2789	232350	334	26	92.3	99.8	100.0	100.0	50.0	100.0
2790	243450	451	13	94.1	96.9	93.3	96.7	100.0	100.0
2791	442625	636	25	82.8	99.6	80.7	98.8	83.8	97.3
2792	162000	318	11	100.0	100.0	100.0	100.0	83.3	96.2
2799	241925	269	33	0.0	99.9	0.0	100.0	100.0	100.0
2800	273300	255	19	66.7	98.8	70.6	99.3	57.1	97.2
2817	299450	281	18	73.9	98.7	72.0	98.7	85.7	97.9
2818	540350	713	14	73.3	100.0	92.3	99.4	47.1	98.9
2904	247725	389	30	100.0	100.0	100.0	100.0	100.0	100.0
2905	293400	325	20	73.9	97.8	66.7	97.8	62.5	95.1
2907	210500	257	15	87.5	97.4	87.5	97.4	100.0	100.0
2908	350025	503	24	90.5	99.1	95.0	99.1	89.3	97.6
2918	122075	143	7	90.0	94.7	90.0	94.7	83.3	90.5
2940	297250	402	17	71.4	97.9	71.4	97.9	66.7	96.4
2959	217925	265	24	28.6	99.4	100.0	100.0	60.0	98.4
2960	317850	340	22	66.7	99.3	60.0	99.1	66.7	97.8
2977	247575	447	32	0.0	99.8	0.0	100.0	100.0	100.0
2978	323175	371	20	80.0	99.8	80.0	99.8	69.0	95.5
2992	192900	149	2	100.0	100.0	100.0	100.0	100.0	100.0
2995	265475	580	9	76.5	99.0	76.5	99.0	52.2	98.1
3011	182550	315	25	0.0	98.7	0.0	98.3	100.0	100.0
3012	325750	365	24	71.4	98.7	63.2	98.1	75.0	98.4
3013	382900	501	19	66.7	97.8	72.2	98.5	84.2	98.5
3014	251050	270	20	33.3	98.5	50.0	99.1	62.5	96.7
3015	274100	252	13	80.0	97.3	83.3	97.3	80.0	99.3
3017	278400	291	18	100.0	100.0	100.0	100.0	92.9	98.2
3020	332950	390	14	72.2	98.0	91.7	99.5	66.7	97.6
Mean	**277672**	**376**	**19**	**69.9**	**98.9**	**74.6**	**98.9**	**79.7**	**98.2**
Median	**273300**	**340**	**19**	**75.0**	**99.2**	**80.7**	**99.2**	**83.3**	**98.4**

Table 1. Features and re-identification measures results in percentage of the whole set of evaluated videos.

Sánchez et al. [10] where only a single face detector is used to localize speakers. The second one, is Diarization Shot Classification (DSC) method that combines a CNN shot classifier and a people standing person detector previous to detecting the face in mid shots. The third, is our complete method, called Diarization Shot Classification Keyframe (DSCK) that uses DSC method and biometric verification of the face.

Four videos have been used for training the CNN shot classifier. Shots were manually labelled into four classes: general shot (4,740 samples), mid shot (1,395

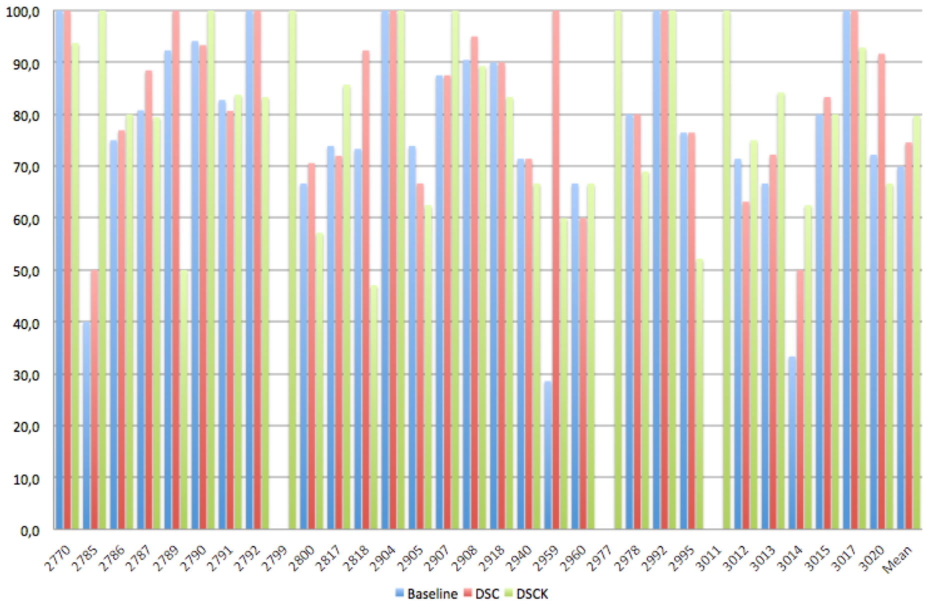

Fig. 5. Comparison methods per video using TRRs (y axis) achieved per video (x axis).

samples), medium close up (4,309 samples) and others contains 143 items that represent title and rating shots.

The assignment of the thresholds for the coeficient of eq. 2 has been calculated using four videos. Biometric measures are calculated for each frame and shot type to obtain the corresponding coeficients. The median value of these coeficients is used as the center value of the thresholds, with an interval of ±3.

For each video a coarse-grained annonation is provided by the Canary Islands Parliament Media Service. As our approach is not based on a clustering technique but in a matching process similarly to a re-identification task, to evaluate the performance of the proposal, the measures described in [3] are used:

- True Re-identification Rate (TRR): the system declares two speakers as the same speaker and they are the same person.
- False Re-identification Rate (FRR): the system declares two speakers as the same speaker but they are different person.
- True Distinction Rate (TDR): the system declares two speakers as different speaker and they are different.
- False Distinction Rate (FDR): the system declares two speakers as different speaker and they are the same person.

4 Results

This section presents the speakers labelling results in the parliamentary sessions scenario. In Fig. 5, ”Baseline” is the baseline strategy [10], ”DSC” is our initial proposed strategy and ”DSCK” is our completed proposed strategy. The summarized rates are presented in Table 1.

In general, the proposed methods give better results than the baseline. This fact can be explained because the baseline method fails in the detection of the speaker in mid shots since deputies that are next to the speaker act as distractor for the face detector. A clear example appears in video 2959, the most part of the video are mid shots with a few medium close up shots. The TRR improvement in this video is 71.4%. On the other hand, the use of biometric traits reduce the error rate of the face detector. We obtain in three videos an improvement of 100%.

On the contrary, in videos 2905 and 2960, the most of the shots are medium close up, the shot classification errors affect negatively to the performance. When there is a shot misclassification, the shot is not processed, or the system tries to find an upper body where there is not, resulting an unpreprocessed shot. Also, at the time to identify a keyframe, the misclassification can introduce an error in the evaluation of the threshold.

Summarizing, the mean TRR with our proposals are better than the baseline, and only just a reduced number of videos reported slightly worse results. Thus, observing Table 1, we can remark that in 80.6% and 61.3% of the videos the DSC and DSCK approaches increases or equals the performance. DSC obtains an average of 4.7% improvement of all processed videos for the TRR, while keeping a similar TDR. Compared to DSC, DSCK obtains a mean of 5.1% TRR improvement. Finally, the TRR increment of DSCK with respect to the baseline is 9.8%.

5 Conclusions

In this paper, we have proposed two new strategies, DSC and DSCK, for labelling speakers in the visual context of diarization system in the Canary Islands Parliament. The focus has been put on the analysis of the shot type, with the purpose of implement a shot classifier for taking a decision if the system has to detect something or not. In the case of mid shots where more deputies apart from the speaker can appear, we take into account upper body and face detection or only face detection. Also, for avoiding false faces biometric traits are used for keyframe extraction. An average improvement in term of TRR of 4.7% and 9.8% for DSC and DSCK respectively is achieved.

References

1. R. Barra-Chicote, J. M. Pardo, J. Ferreiros, and J. M. Montero. Speaker diarization based on intensity channel contribution. *IEEE Transactions on Audio, Speech & Language Processing*, 19(4):754–761, 2011.

2. M. Castrillón, O. Déniz, D. Hernández, and J. Lorenzo. A comparison of face and facial feature detectors based on the violajones general object detection framework. *Machine Vision and Applications*, 22(3):481–494, 2011.
3. D.-N. T. Cong, L. Khoudour, C. Achard, C. Meurie, and O. Lezoray. People re-identification by spectral classification of silhouettes. *Signal Processing*, 90(8):2362 – 2374, 2010. Special Section on Processing and Analysis of High-Dimensional Masses of Image and Signal Data.
4. T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley Interscience, 2006.
5. G. Garau and H. Bourlard. Using audio and visual cues for speaker diarisation initialisation. In *2010 IEEE international conference on acoustics speech and signal processing (ICASSP)*, page 49424945, 2010.
6. I. Kapsouras, A. Tefas, N. Nikolaidis, G. Peeters, L. Benaroya, and I. Pitas. Multimodal speaker clustering in full length movies. *Multimedia Tools and Applications*, 2016. `http://dx.doi.org/10.1007/s11042-015-3181-5`.
7. N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1962–1977, October 2011.
8. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278 – 2324, 1998.
9. A. Noulas, G. Englebienne, and B. J. A. Krose. Multimodal speaker diarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):79–93, Jan 2012.
10. E. Sánchez-Nielsen, F. Chávez-Gutiérrez, J. Lorenzo-Navarro, and M. Castrillón-Santana. A multimedia system to produce and deliver video fragments on demand on parliamentary websites. *Multimedia Tools and Applications*, 2016. `http://dx.doi.org/10.1007/s11042-016-3306-5`.
11. N. Sao and R. Mishra. A survey based on video shot boundary detection techniques. *nternational Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)*, 3(4), 2014.
12. N. Sarafianos, T. Giannakopoulos, and S. Petridis. Audio-visual speaker diarization using fisher linear semi-discriminant analysis. *Multimedia Tools and Applications*, 75(1):115–130, 2016.
13. C. Sujatha and U. Mudenagudi. A study on keyframe extraction methods for video summary. In *Computational Intelligence and Communication Networks (CICN), 2011 International Conference on*, pages 73–77, Oct 2011.
14. T. Teixeira, G. Dublon, and A. Savvides. A survey of human-sensing: Methods for detecting presence, count, location, track, and identity. *ACM Computing Surveys*, 5:1–77, 2010.
15. S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565, Sept 2006.
16. F. Vallet, S. Essid, and J. Carrive. A multimodal approach to speaker diarization on TV talk-shows. *IEEE Trans. Multimedia*, 15(3):509–520, 2013.
17. P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):151–173, May 2004.
18. W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *Association for Computing Machinery*, 35(4):399–458, 2003. `http://doi.acm.org/10.1145/954339.954342`.