Descriptors Comparisons for Vision based Speaker Diarization Approaches 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

1 Extended abstract

Speaker Diarization deals with annotating who and when a speaker is talking, it represents a challenge for scientific community [1]. This problem can be tackled from a point of view of a re-identification process, detecting a speaker and checking whether he/she appears again.

Several approaches have been proposed to solve the diarization problem. They have been based on audio, video and a combination of both. Intensity Channel Contribution (ICC) [1], bottom-up hierarchical agglomerative clustering of Mel Frequency Cepstral Coefficients (MFCC) are used for audio. In video diarization, lips movement or face centered is used. Finally, as dual solution, audio and video descriptors are used, both descriptors have to keep the coherence.

In this paper, we are interested in diarization of parliamentary debates sessions based only on video. The scenario is composed of a presidential table, platform and seats, from deputies could intervene. Theses interventions are recorded by a network of cameras distributed in the Parliament, which can do pan, tilt and zoom.

To extract features, first the face is obtained as region of interest (ROI) and then local descriptors are extracted because they have demonstrated good performance in facial analysis [2]. After the speaker face is modeled, a matching stage is carried out by comparing the extracted features against the database models.

The experiments have been realized using 29 videos[3]. Those videos variate in number of frames, shots and speakers. As features we have considered in the comparison the following ones: Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), LBP Uniform (LBPu2), Intesity based LBP (NI-LBP), Local Gradient Patterns (LGP), Local Phase Quantization (LPQ), Local Salient Patterns (LSP0), Local Ternary Patterns (LTP), LTP high (LTPh),

[3] Videos available at http://www.parcan.es

LTP low (LTPl), Weber Local Descriptor (WLD) and Local Oriented Statistics Information Booster (LOSIB). They are calculated in the ROI using a 5×5 grid. At the time, the comparison of the models were calculated with Canberra, Chebyshev, Cosine, Euclidean and kullback–Leibler (KL) divergence histogram measures. Getting an idea of the number of experiments executed, 29 videos with 12 local descriptors with five different measures were processed, a total of 1740 experiments. Moreover, those experiments were validated by five diarization approximations, obtaining a total of 8700 experiments. Those diarization approximations are methods to evaluate the performance of the local descriptors under consideration, True Re-identification Rate (TRR) and True Distinction Rate (TDR) are used as measures [3]. Besides, for a specific speaker audio annotation, audio fragment, could appear different deputies shots in the video sequences. For this reason, we propose the following four approaches:

- First Appearance (FA): The person of the first shot detected by the system in the audio fragment is taken as representative speaker shot.
- Most Frequent (MF): The person that the system detect as greater number of occurrences in the audio fragment is taken as representative speaker shot.
- Greater Length (GL): The person that the system detect as higher duration shot in the audio fragment is taken as representative speaker shot.
- Greater Total Length (GTL): The person that the system detect as higher duration in the audio fragment is taken as representative speaker shot.

Considering the mean TRR and TDR, the best local descriptors are WLD (45.04%, 78,91%) and NILBP (44.97%, 77.15%), that behave better in this scenario. Then, if we focus on the descriptors comparison measures, the two higher are KL (49.47%, 70.98%) and Cosine distance (43.21%, 79.81%). Ultimately, comparing the employed approaches to evaluate the performance, MF and GTL achieve highest values.

To summarize this paper, four approaches to measure the performance of diarization problems have been proposed. Moreover, different local descriptors were compared, obtaining a general idea of their behavior. Finally, multiple histograms measures to matching have been compared, allowing us to know what configuration give us greater results for upcoming test.

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. Castrillón-Santana, M., Lorenzo-Navarro, J., Ramón-Balmaseda, E.: Multi-scale score level fusion of local descriptors for gender classification in the wild. Multimedia Tools and Applications (in press) (2016), `http://dx.doi.org/10.1007/s11042-016-3653-2`
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.