Deep Multimodal Speaker Naming

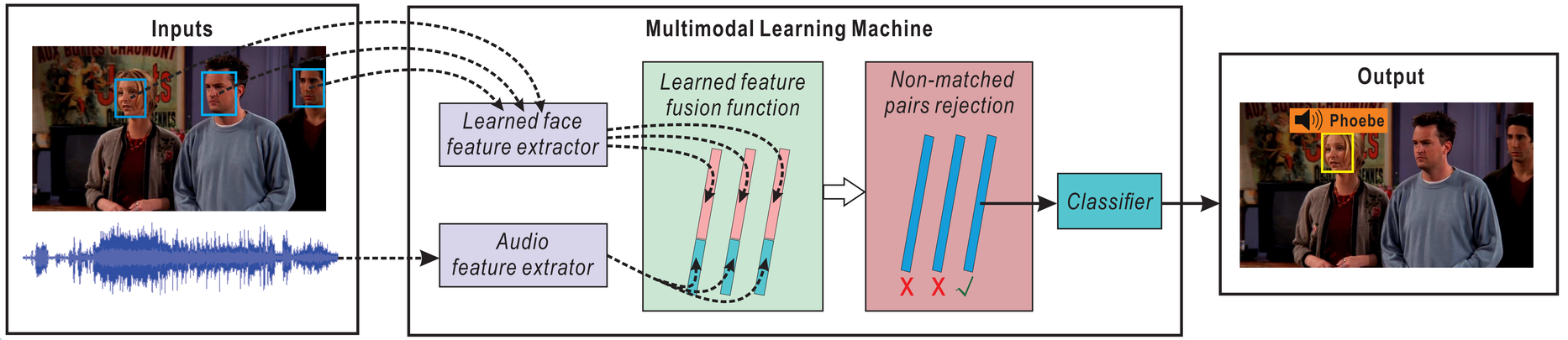
Yongtao Hu, Jimmy Ren, Jingwen Dai, Chang Yuan, Li Xu, and Wenping Wang

Project page, code and dataset: http://herohuyongtao.github.io/projects/deepsn/

Overview

Automatic speaker naming is the problem of localizing as well as identifying each speaking character in a video. Previous multimodal approaches to this problem usually process the data of different modalities individually and merge them using handcrafted heuristics. In this work, we propose a novel CNN based learning framework to automatically learn the fusion function of both face and audio cues. Without using face tracking, facial landmark localization or subtitle/transcript, our system with robust multimodal feature extraction is able to achieve state-of-the-art speaker naming performance evaluated on 2 diverse TV series.

Multimodal learning framework



Experimental setup

Dataset. We evaluate on over three hours videos of nine episodes from two TV series, i.e. *"Friends"* and *"The Big Bang Theory"* (*"BBT"*).

- *"Friends"*: *S01E03* (Season 01, Episode 03), *S04E04*, *S07E07* and *S10E15* for training and *S05E05* for testing.
- *"BBT"*: as in [3], *S01E04*, *S01E05* and *S01E06* for training and *S01E03* for testing.

Features.

- Face: all face images are resized to 50×40 and raw RGB pixels are sent to the learning framework.
- Audio: a window size of 20ms and a frame shift of 10ms are used. We then select mean and standard deviation of 25D MFCCs, and standard deviation of 2-ΔMFCCs, resulting in a total of 75 features per audio sample.

Results

Face model.

- Previous work: Eigenface 60.7%, Fisherface 64.5%, LBP 65.6% and OpenBR/4SF 66.1%.
- Ours: face-alone 86.7% and face-audio 88.5%.

Identifying non-matched pairs. Note that, in the above experiments of face models, all the face-audio samples for training and evaluation are matched pairs, i.e. belong to the same person. Instead of using the final output label of our face models, we explore the effectiveness of the features returned from the model in the last CNN layer. We evaluated three SVM models on three different features:

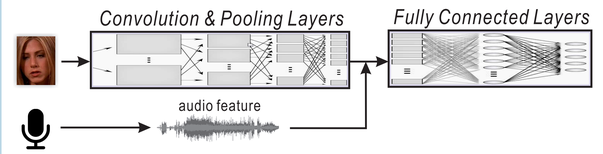
- fused feature: 82.2%.
- face feature+MFCC: 82.9%.
- fused feature+MFCC: 84.1%.

Speaker naming. It can be viewed as an extension of previous experiments.

Video \ Voting window size N	1	10	30	50	60	70
Friends	75.3	79.6	84.4	88.2	90.0	90.5
The Big Bang Theory	73.6	74.7	78.2	82.5	83.0	83.4

Compared with previous works. Previous works [1, 3] achieved speaker naming accuracy of 77.8% and 80.8% respectively (evaluated on *"BBT S01E03"*) by incorporating face, facial landmarks, cloth features, character tracking and associated subtitles/transcripts (equal to our method when $N = 58$). In comparison, we can achieve SN accuracy of 82.9% without introducing face/person tracking, facial landmark localization or subtitle/transcript.

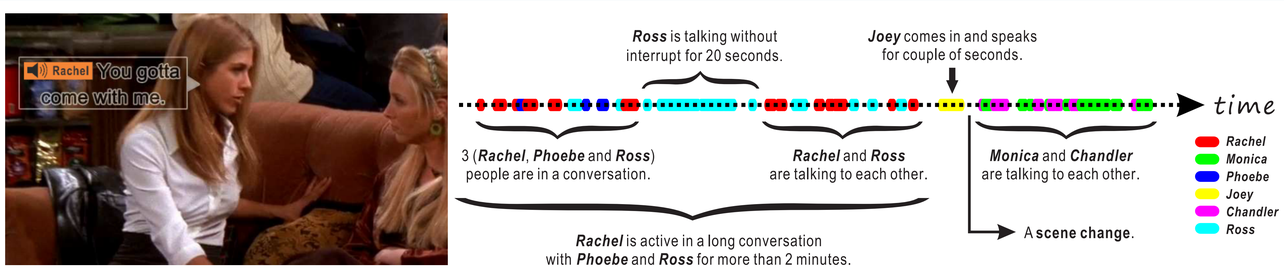
Multimodal CNN



Demo

Applications

1. **Content accessibility enhancement.** Using speaker-following subtitles [2] to enhance overall viewing experience for videos and automatic speech balloons for games or other VR/AR applications.
2. **Video retrieval and summarization.** With detailed speaking activities, many high-level tasks can be achieved.



Reference

[1] M. Bauml, M. Tapaswi, and R. Stiefelhagen. Semi-supervised learning with constraints for person identification in multimedia data. In *CVPR*, 2013.

[2] Y. Hu, J. Kautz, Y. Yu, and W. Wang. Speaker-following video subtitles. *TOMM*, 2014.

[3] M. Tapaswi, M. Bauml, and R. Stie-felhagen. "Knock! Knock! Who is it?" probabilistic person identification in TV-series. In *CVPR*, 2012.