

Detecting speaking persons in video

Hannes Fassold
JOANNEUM RESEARCH - DIGITAL
Graz, Austria
hannes.fassold@joanneum.at

Abstract—We present a novel method for detecting speaking persons in video, by extracting facial landmarks with a neural network and analysing these landmarks statistically over time.

Index Terms—Speaking person detection, facial landmark extraction

I. INTRODUCTION

Knowing which persons are actually speaking in a video at a certain time is an important semantic information which is useful in several application fields. We therefore present a novel robust method for detecting speaking persons in video which relies only on the visual information. It first extracts for all persons visible in the video their facial landmarks with a high-quality deep learning approach (see section II). Afterwards, a statistical analysis of the landmark trajectories over time is employed to detect the temporal segments in which a certain person is speaking (see section III).

II. FACIAL LANDMARK METADATA EXTRACTION

For the extraction of the facial landmark metadata, for each frame of the video sequence we first invoke the SF3D face detector from [1] in order to detect the ROIs in the frame which correspond to a face. For each detected face, we extract its facial landmarks with the 2D-FAN algorithm proposed in the work [2]. So for each face in each frame of the video, we retrieve a face descriptor containing the facial landmarks and some other data. In order to build a trajectory of the facial landmarks of one person over time, we have to group together all face descriptors belonging to the same person. For that, we currently employ a simple clustering strategy, based on the location of the extracted facial landmarks in the image.

III. DETECTOR ALGORITHM

For each person occurring in the video, we do now a statistical analysis of its facial landmark trajectory over time in order to infer whether that person is speaking or not and to detect in which time intervals the person does speak. When a person is speaking, he/she is repeatedly opening and closing the mouth. So we design our analysis method to detect this temporal pattern, based on the extracted landmarks for the mouth (visualized in yellow and red in Fig. 1). The basic idea is to measure the deviation in y -direction of the landmarks located at the lips and accumulate them over a few seconds, with a high deviation (significant lip movement) indicating that the person is speaking currently. As a preprocessing step, we have to *normalize* the 68-dimensional landmark vectors,



Fig. 1. Speaking person detector with visualized facial landmarks. Image courtesy of IgelTV.

thereby making them invariant to their spatial position and the face height. For this, we shift each vector so that its coordinate origin $(0, 0)$ is located at the middle of the mouth. Additionally, we scale each vector by the estimated face height (calculated from the top and bottom facial landmarks). We construct now a 1-D vector t from the normalized y -coordinate of the facial landmark located at the top of the upper lip for each frame. In the same way, we construct a 1-D vector b from the respective facial landmark located at the bottom of the lower lip. In order to calculate the deviation of both vectors over an accumulation period of a few seconds, we calculate now significantly blurred versions of both vectors $\tilde{t} = \text{smooth}(t)$ and $\tilde{b} = \text{smooth}(b)$. The deviation vectors are now calculated as $d = \text{abs}(t - \tilde{t})$ and $e = \text{abs}(b - \tilde{b})$, and a combined deviation vector c is calculated as the average of d and e . We retrieve the time intervals where a speaking person is detected now by thresholding the combined deviation vector c , using a user-specified threshold T .

ACKNOWLEDGMENT

This work was supported by European Union's Horizon 2020 research and innovation programme under grant number 951911 - AI4Media.

REFERENCES

- [1] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3fd: Single shot scale-invariant face detector," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [2] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1021–1030.