An Active Speaker Detection Method in Videos using Standard Deviations of Color Histogram

Publisher: IEEE

Adekunle A. Akinrinmade; Emmanuel Adetiba; Joke A. Badejo; Oluwadamilola Oshin
All Authors

Abstract
Document Sections

I.

Introduction

II.

Related Works

III.

Methodology

- IV.

Experiment

- V.

Discussion

Show Full Outline
Authors
Figures
References
Keywords
Metrics

Abstract:
Active Speaker Detection (ASD) is a process that predicts who the speaker is amongst those whose faces appear in a video (if any) at any given point in time within the recorded video. This work presents a novel algorithm capable of detecting the active speakers in each video using the standard deviations of Color Histograms (CHs) computed at the mouth region from one frame to another. This paper relies on the assumption that the lips of an active speaker are in motion. They open and close and thus reveal the inner parts of the mouth, like the tongue, teeth, and the vocal cavity which are of diverse colors in the process of talking. It is possible to use already existing algorithms to detect the mouth region. This region can be analyzed during the speaking process for the changes in color activity, and this can be used to predict whether a user is speaking or not. If a person is not speaking, the lips are at rest the CH of such mouth regions such candidates would be stable. As a result, the standard deviations of such regions would be negligible. A threshold can be experimentally determined which is thus capable of predicting if a person is speaking or otherwise. This paper explores 53 online videos from Channels TV station, these videos were employed in the creation of 250 video clips. Each clip is between 15 to 60 seconds with a total of 3.6 hours. Each video contained the faces of at most two speakers in no particular order. Sometimes, only one of the speakers' faces appears, at other times both appear in the duration of the video. The status of the speakers whether active or not was

manually labeled to be used for the performance evaluation of the proposed algorithm. This method was able to predict the active speakers with an accuracy of 99.19%.
Published in: 2023 International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG)
Date of Conference: 05-07 April 2023
Date Added to IEEE *Xplore*: 22 May 2023
ISBN Information:
DOI: 10.1109/SEB-SDG57117.2023.10124488
Publisher: IEEE
Conference Location: Omu-Aran, Nigeria
Funding Agency:

I. Introduction

ASD plays an important role in human-computer/human-robot interactions [3] [4], audio-visual diarization [5], it allows a deaf audience to get better value from movies [6], video conferencing systems [7], in the automatic curation of the audio samples of videos where the faces of the subjects are visible [8], speaker naming where the identity of the speaker is also revealed [6], speech enhancement, video re-targeting during meetings [1] and a prerequisite used by artificial cognitive systems in the acquisition of different languages in a social setting [9]. ASD research is faced with some challenges, these include the presence of numerous people which leads to the variability in the choice of the possible speakers in the video, videos with poor resolutions [10], sometimes, the speaker in the video is off-the-screen [11], the faces of these speakers could be turned at angles making face detection difficult. Some mouth motions do not necessarily imply speaking, for example, chewing, grumbling, yawning, smiling, grunts, humming, sighs, coughing, and so on [1], [2] can confuse ASD algorithms