Project Overview

Introduction:
This project focuses on finding the speaker of a group discussion by performing machine learning using the common view focus among a group as a feature. Specific applications can be auto-recording of a seminar, or lectures with Q&A interaction. Moreover, if we take the key concept out of the project, which is finding the common view focus of a group, this idea could be used to classify injured person or persons among a group of people with the common knowledge that a person at common view focus is generally different from others.

Background:
Studies on group behavior were mostly about abnormal crowd behavior classification. Studies on group speaker detection were about voice classification and lip motion. Here we propose a classification based on vision and group behavior.

Aim and Objectives:
We aim to develop a functional algorithm that can determine the speaking person in a group discussion with vision data. The objective of focusing on this topic is to explore a new feature of human group interaction for machine and programs to extract. With this kind of technology developed, artificial intelligence could be applied to a wider range of works that need to deal with a group of persons.

Methodology

Pre-processing:
Before feature extraction, Face estimation cannot give 100% accuracy, so there will be false detections and missing faces errors. Our algorithm requires full knowledge of every face with pose in every frame to make precise classifications. We proposed the pre-processing flow on the right. After identifying a cluster, we use linear interpolation to fill in the missing faces.

Common view Focus Feature:
1. Depth position:
2. Distance and Angle

Pre-processing:
Before feature extraction, Face estimation cannot give 100% accuracy, so there will be false detections and missing faces errors. Our algorithm requires full knowledge of every face with pose in every frame to make precise classifications. We proposed the pre-processing flow on the right. After identifying a cluster, we use linear interpolation to fill in the missing faces.

Common view Focus Feature:
1. Depth position:
   Faces at different depth give different size of bounding boxes. We take the inverse of size of bounding boxes to get face depth.
2. Distance and Angle:

Classification:
1. Regression
   To find the common focus of the group, we need to extract the distance between a person and other participants’ view focus. We also force a penalty on the faces that others don’t look at by adding the largest possible angle or distance calculated by the face different of the two faces. In terms of the angle, we make the normalized angle to be one, which is the maximum normalized angle could achieve.

   (a) Movement of mouth: count the pixels which value is lower than a threshold. Threshold is defined by $|x-y| < \tau$.
   (b) Other features include face pixel difference, mouth pixel difference, etc. totally 8 features are selected.

2. AdaBoost
   AdaBoost, short for Adaptive Boosting, is a machine learning algorithm based on the idea of boosting. It tends to make a set of weak classifiers into a single strong learner.

Results: