NATIONAL CENTRE FOR SCIENTIFIC RESEARCH "DEMOKRITOS"
UNIVERSITY OF THE PELOPONNESE

Lecture Speech Analytics

P6 – Multimodal Analysis

Dimitros * Karousis * Karozis * Mastrapas * Nikoloutsakos

Project Description

Developing a video lecture classifier

- Process video and audio data.
- Extract features.
- Apply Machine Learning algorithm to train a classifier.
- Create an end-to-end process.

- *Hey classifier, what do you think about this lecture ? Is it boring, neutral or exciting ?*
- *Just a moment... Mmm, some segments are boring, some others are just neutral, but most of them are really exciting !*

Giving business dimensions to the project

Translating speech features into value.

- Conferences organisation: speakers assessment for improving the quality of the conferences.
- University Master programs: use speech analytics to make assessment of the lecturers.
- Improve customer experience: Speech analytics can analyze conversations via phone, email, text, webchat and social and turns them into searchable, structured data that businesses can use to gain insight into what customers feel or think (e.g. call center)

Project requirements

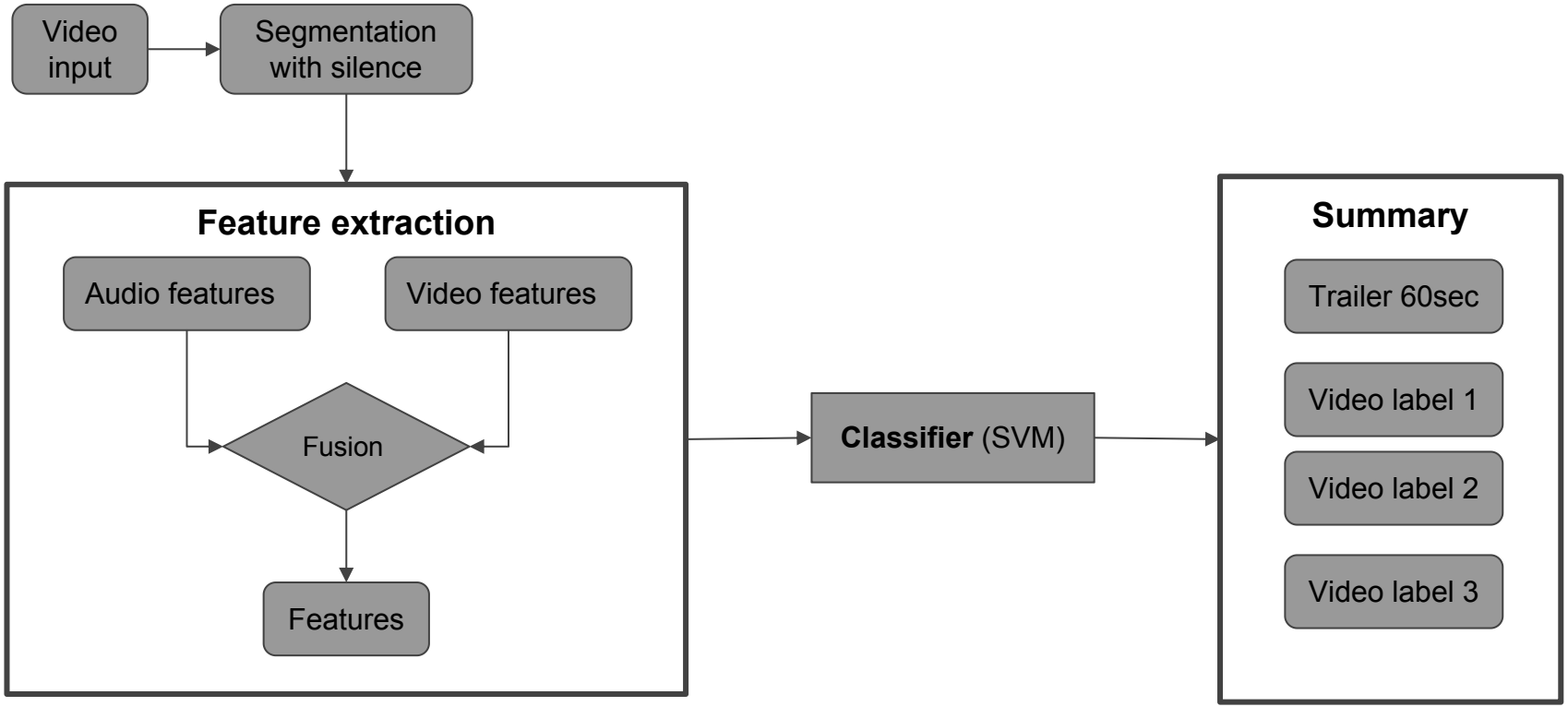
1. Annotate 10 recordings (videos) from the course in terms of: topic, arousal, tone variation, etc. Keep 5 as “external” test
2. Extract audio features (relevant to the tasks in (1))
3. Extract visual features (relevant to the tasks in (1))
4. Predict the annotated targets of (1) and evaluate using leave-one-video-out cross validation
5. Using the trained models, create a summarization tool and test it on the external test dataset

Tools used

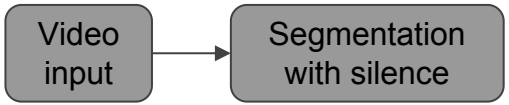
The project was created in Python 3 programming language and the code is available in a github public repository. The Trello task manager used in order the team member to collaborate more efficiently and a Nextcloud server was used to exchange big data.

Process – Methodology

Process


Video input
Segmentation with silence
Feature extraction
Audio features
Video features
Fusion
Features
Classifier (SVM)
Summary
Trailer 60sec
Video label 1
Video label 2
Video label 3

Process



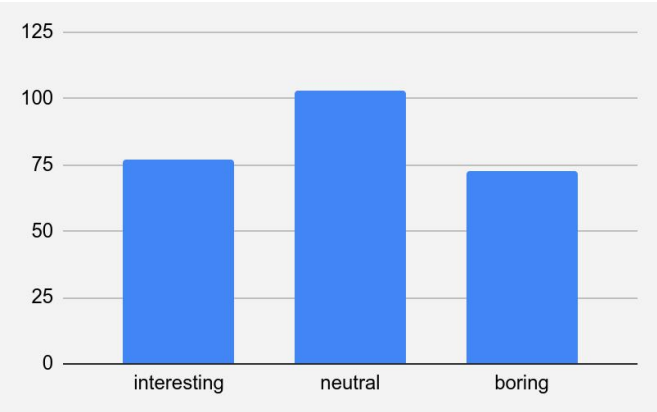
- Use pyAudioAnalysis to find silent points
- Produce segmented videos from silent to silent point (~20sec)

Train dataset

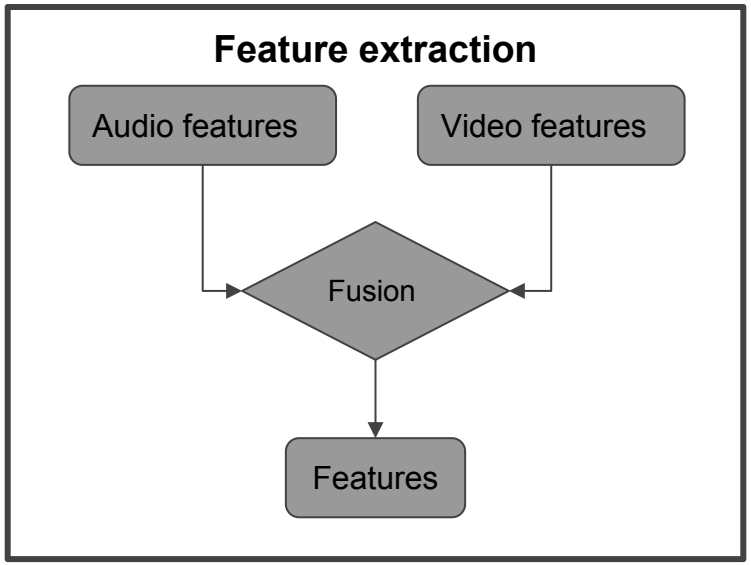
- Annotate parts for 1 class with 3 labels
  - Interest
  - Boring
  - Neutral
- 5 videos -1 50 parts / video

Classifier (SVM)

- Use SVM classifier



Process



Audio

pyAudioAnalysis -ı 64 features

Video

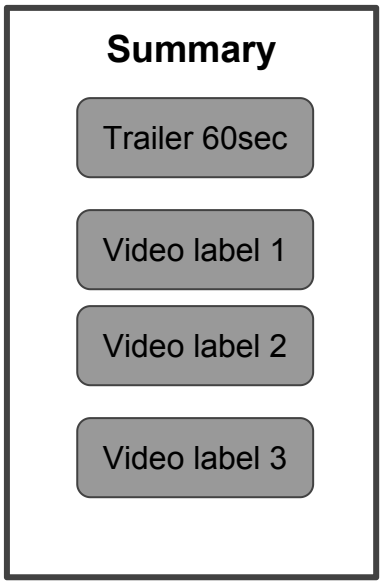
- ~4 frames for every segmented video
- Keep average value of features

VGG16 -ı over 25000 (i)

Random vector -ı size 100 (ii)

(i) x (ii) -ı 100 features

Process



User define parameters

- Percentage of each class
- Classes to be used
- Max duration

The average duration of parts is taken as the unit that will add up to ‘max duration’ parameter

Results

Results

Overall accuracy: 0.316

- Bad annotation (more or less expected).
- Small number of instances compared to the number of features.

(audio only: 0.28
Video only: 0.32)

Speaker 1

Actual / Predict	'boring'	'neutral'	'interesting'
'boring'	0	13	4
'neutral'	0	12	4
'interesting'	0	15	2

Speaker 2

Actual / Predict	'boring'	'neutral'	'interesting'
'boring'	1	0	7
'neutral'	2	0	26
'interesting'	0	0	14

Speaker 3

Actual / Predict	'boring'	'neutral'	'interesting'
'boring'	0	21	0
'neutral'	0	12	0
'interesting'	0	17	0

Future work

- More instances (now we run 250 instances with 164 features)
- Better selection of class type (now, arousal used)
- More annotation (with established rules among annotators)
- More extensive optimization & tuning (e.g. Grid search, Other models, deep in case of more data)

Questions

- Send data (already versioned)
- Try 10-fold validation
- Try without speaker independency

Demo

THANK YOU!