Audio-Visual Weakly Supervised Approach for Apathy Detection in the Elderly

Garima Sharma, Jyoti Joshi, Radia Zeghari, Rachid Guerchouche

IJCNN 2020

OUTLINE

- Introduction
- Related Work
- Dataset
- Model Architecture
- Experiments and Results
- Conclusion
- Progress Report

Introduction

- It is found that there is a high chance of occurrence of apathy in various mental disorders such as 73% in Alzheimer's disease. [4]
- The diagnosis of apathy involves series of interviews, which are conducted by an expert to identify a change in a person's behaviour and loss of interest and activities.
- Studies also show that apathy is often misdiagnosed or confused with depression. [6]

[4] P. H. Robert, E. Mulin, P. Mallea, and R. David, "Apathy diagnosis, ´ assessment, and treatment in alzheimer's disease," CNS neuroscience & therapeutics, vol. 16, no. 5, pp. 263–271, 2010.
[6] L. Nobis and M. Husain, "Apathy in alzheimer's disease," Current opinion in behavioral sciences, vol. 22, pp. 7–13, 2018.

Introduction

- According to literature in psychology, apathy can be identified by analyzing the emotions. [13]
- In this paper, an automatic multimodal system is proposed to detect apathy.
- The proposed method contains four parts to effectively analyze the emotion exploiting the **facial expressions**, **facial action units**, **speech features** and **local facial motion information**.

[13] S. E. Starkstein, “Apathy and withdrawal,” International Psychogeriatrics, vol. 12, no. S1, pp. 135–137, 2000.

Related Work

Apathy Detection

- Along with facial expressions, speech of a person can also be used to detect the state of apathy. [14]
- In a recent study, Happy et al. [18] identified the state of apathy by using facial expressions and the local facial motion. Their method uses the positive and negative narration video as a set to extract emotion and local motion features.

[14] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, “Multimodal assistive technologies for depression diagnosis and monitoring,” Journal on Multimodal User Interfaces, vol. 7, no. 3, pp. 217–228, 2013.
[18] S. Happy, A. Dantcheva, A. Das, R. Zeghari, P. Robert, and F. Bremond, “Characterizing the state of apathy with facial expression and motion analysis,” in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE, 2019, pp. 1–8.

Related Work

Emotion Detection

- Barrett et al. [26] argued that facial expressions alone can’t estimate the emotion of a person.
- Audio and facial features are combined to identify different emotions in several studies [27].
- It has also been found that it is difficult to recognize facial expressions of an old person as compared to a young person due to the wrinkles and folds in the face [28].

[26] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, “Emotional expressions reconsidered: challenges to inferring emotion from human facial movements,” Psychological Science in the Public Interest, vol. 20, no. 1, pp. 1–68, 2019.
[27] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, “Audio-visual emotion recognition in video clips,” IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 60–75, 2017.
[28] M. Folster, U. Hess, and K. Werheid, “Facial age affects emotional ¨ expression decoding,” Frontiers in psychology, vol. 5, p. 30, 2014.

Related Work

Multiple Instance Learning

- In MIL, the training data is divided into multiple sets and one label is used for the complete set without having the labels for each data point [30].
- In a recent study, Xu et al. [36] proposed a weakly supervised deep learning based MIL method in medical image processing. Further, Zhu et al. [37] proposed a MIL based method with salient windows focused on unsupervised object detection.

[30] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, “Multiple instance learning: A survey of problem characteristics and applications,” Pattern Recognition, vol. 77, pp. 329–353, 2018.
[36] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. Eric, and C. Chang, “Deep learning of feature representation with multiple instance learning for medical image analysis,” in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 1626–1630.
[37] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, “Unsupervised object class discovery via saliency-guided multiple class learning,” IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 4, pp. 862–875, 2015.

Dataset

- The data used in this study is collected in **Nice Memory Research Center in Nice University Hospital**. Videos are recorded in an interview where a person is **narrating a positive and a negative experience** to a clinician.
- The videos are then labeled to be **apathetic or non-apathetic** by an expert after observing the person. The dataset contains videos of 70 participants among which 28 are apathetic and 42 are non apathetic. (apathy dataset)
- The images available in Faces dataset [41] are used first in the model architecture to pre-train and thus fine tune it on apathy dataset.

[41] N. C. Ebner, M. Riediger, and U. Lindenberger, “Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation,” Behavior research methods, vol. 42, no. 1, pp. 351–362, 2010.

Model Architecture

- Each video is divided into equal sized chunks. It helps in generating large number of clips from a limited number of original videos.
- After training a model from the features described below, chunk level prediction is converted into video level prediction using MIL.

Model Architecture

Visual Features-ElderFace

- To improve facial expression recognition in elderly, first the training is performed on a separate data of elderly people [41].
- This training is performed by initializing the model with VGGFace [42] 6th layer features.
- This pretrained model on separate elderly data is then used to extract features for the apathy dataset.

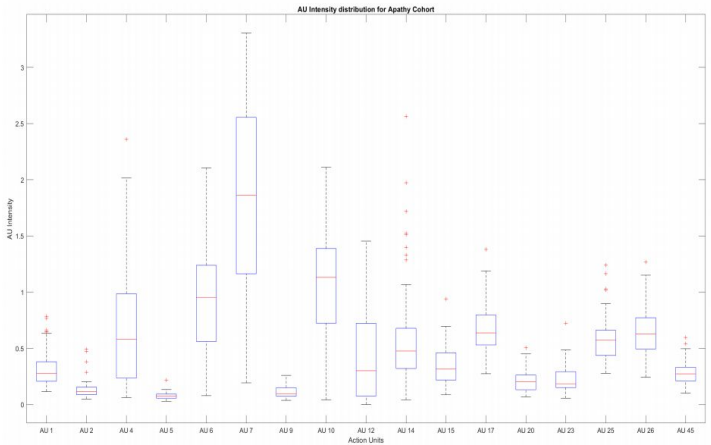
[41] N. C. Ebner, M. Riediger, and U. Lindenberger, “Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation,” Behavior research methods, vol. 42, no. 1, pp. 351–362, 2010.
[42] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” 2015.

Model Architecture

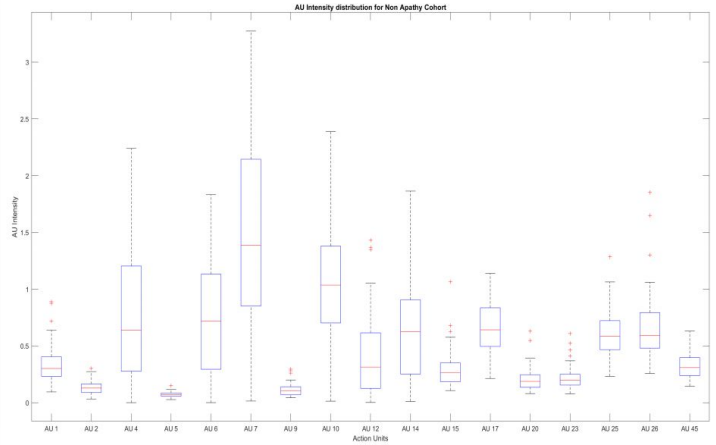
Visual Features-Action Units

- Action Units (AU) encode the small muscle movements in a face, which are useful in recognizing the facial expression of a person.
- The intensity of AU is computed from OpenFace 2.0 toolkit at frame level.

AU Number	AU Description
1	Inner brow raiser
2	Outer brow raiser
4	Brow lowerer
5	Upper lid raiser
6	Cheek raiser
7	Lid tightener
9	Nose wrinkler
10	Upper lip raiser
12	Lip corner puller
14	Dimpler
15	Lip corner depressor
17	Chin raiser
20	Lip stretcher
23	Lip tightener
25	Lips part
26	Jaw drop
45	Blink



(a) Action units distribution for apathy cohort

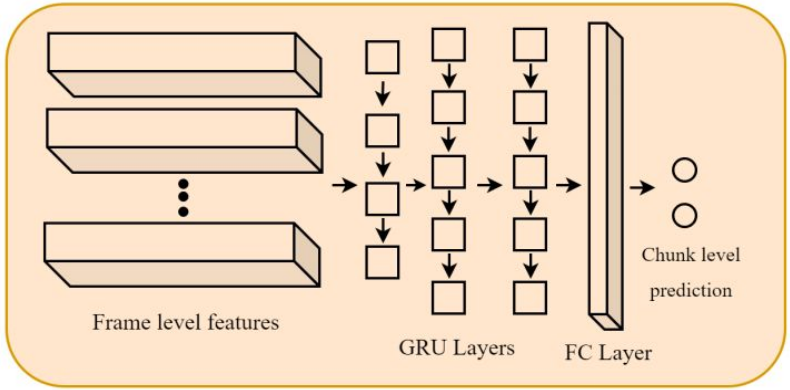


(b) Action units distribution for non apathy cohort

Model Architecture

Motion Features

- 68 2D facial landmarks are extracted from OpenFace 2.0. This represents and encodes local motion within the face.
- The above mentioned features are extracted and trained separately using a Gated Recurrent Unit (GRU) network to learn the temporal changes.



Model Architecture

Frequency	Energy	Spectral	Temporal features
Pitch	Shimmer	Alpha ratio	Rate of loudness peaks
Jitter	Loudness	Hammarberg Index	Mean length and standard deviation of voiced regions
Formant 1, 2, 3 frequency	Harmonic to noise ratio	Spectral Slope 0–500 Hz and 500–1500 Hz	Mean length and standard deviation of unvoiced regions
Formant 1		Formant 1, 2, and 3 relative energy	No. of continuous voiced regions per second
		Harmonic difference H1–H2 and H1–A3	

Audio Features

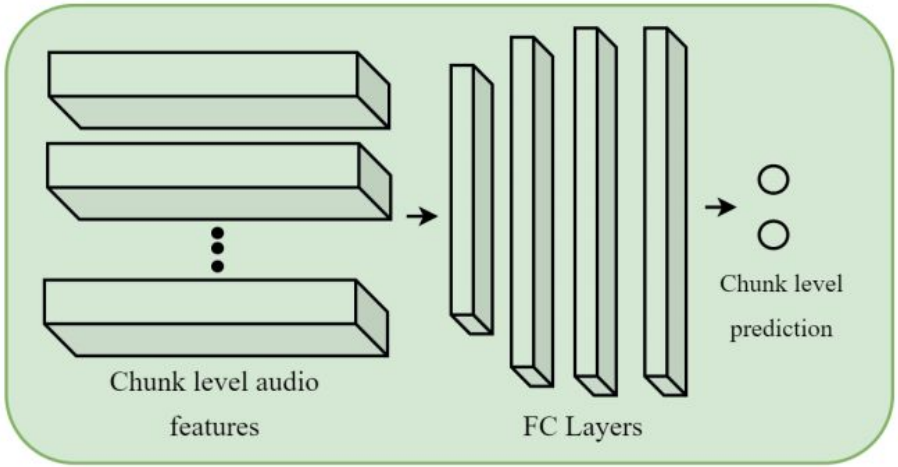
- The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [48] defines a minimalistic feature set which is widely used to extract features from audio signals to recognize the emotion.
- These features are found to be beneficial to encode the emotion of a person [49]. The features are extracted for each video using OpenSMILE [50] toolkit.

[48] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, ´ L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan et al., “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” IEEE transactions on affective computing, vol. 7, no. 2, pp. 190–202, 2015.
[49] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” in 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016, pp. 5200–5204.
[50] F. Eyben, M. Wollmer, and B. Schuller, “Opensmile: the munich ¨ versatile and fast open-source audio feature extractor,” in Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462.

Model Architecture

Audio Features

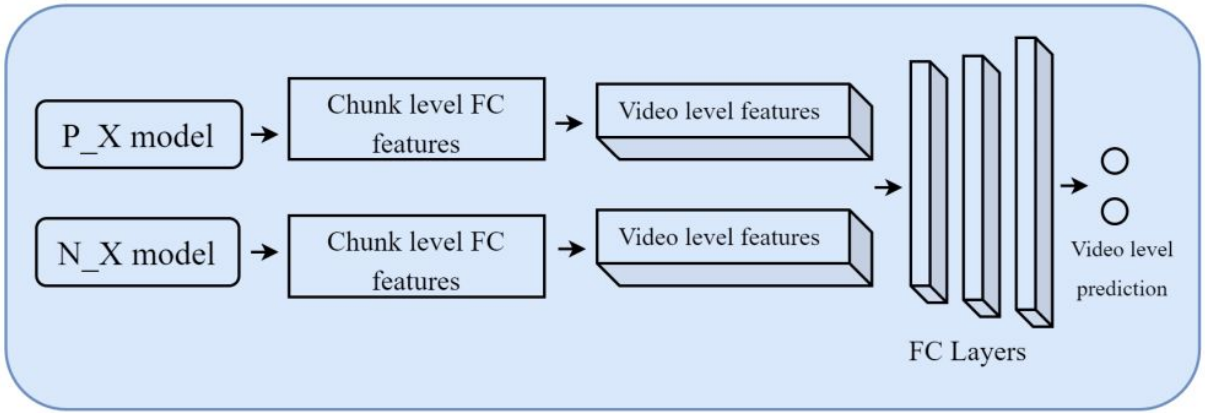
- A small deep neural network is trained which has 5 FC layers having size 128, 256, 1024 and 2048.



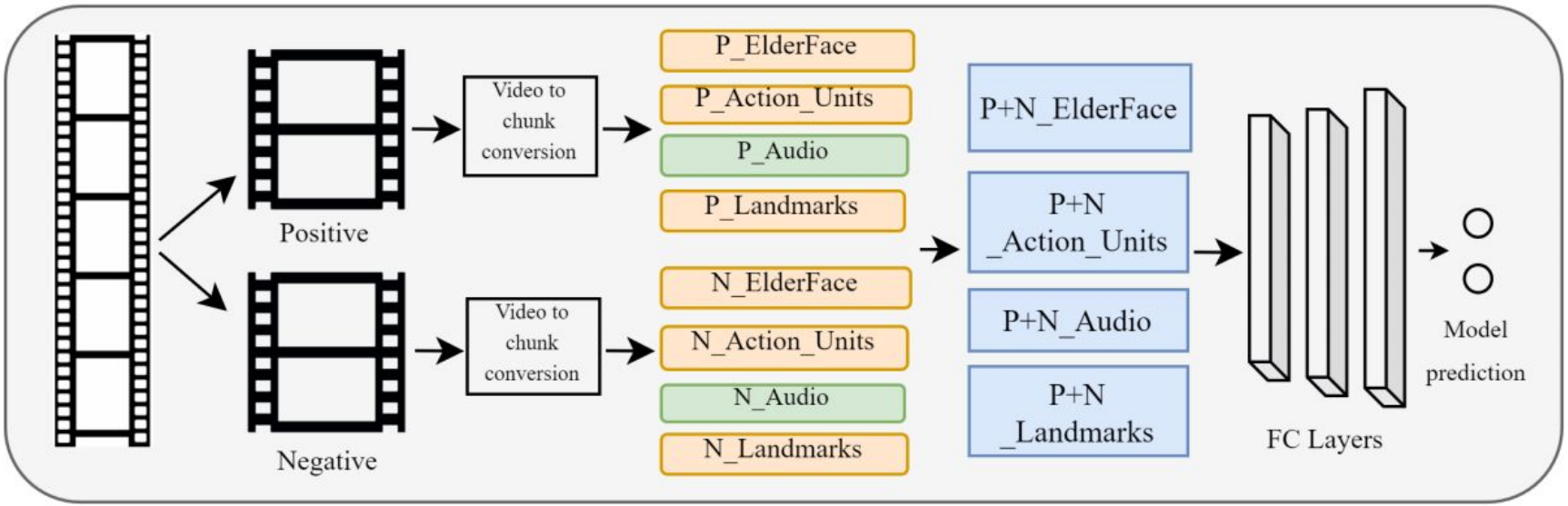
Model Architecture

Fusion of Features

- These 2D chunk level features are converted to 3D video level features and then are concatenated together.
- The final model obtained by this is used to make apathetic and non-apathetic predictions.



Model Architecture



Experiments and Results

Preprocessing

- The size of each chunk is fixed to be 32 frames.
- The idea of selecting only 32 frames is to keep a trade off between performance and computational efficiency.

TABLE III: Details of the data used in experiments.

	No. of participants	No. of videos	No. of chunks
Apathetic	28	56	2231
Non apathetic	42	84	3675
Total	70	140	5906

Experiments and Results

Results

Features	ElderFace			Action Units			Audio			Landmarks			Combined
	P	N	P+N	P	N	P+N	P	N	P+N	P	N	P+N	
Accuracy (%)	60.00	60.00	54.28	51.42	57.14	60.00	54.28	52.85	61.42	55.71	57.14	54.28	**75.71**
F1-score	0.602	0.603	0.517	0.510	0.571	0.602	0.535	0.529	0.617	0.560	0.575	0.547	**0.736**

- It is evident that ElderFace and Landmarks features significantly discriminate apathetic behaviour in positive or negative videos, separately yielding higher accuracy.
- The results also show that only expression level information is not sufficient given the complexity of the apathy detection task.

Experiments and Results

Results

Features	ElderFace			Action Units			Audio			Landmarks			Combined
	P	N	P+N	P	N	P+N	P	N	P+N	P	N	P+N	
Accuracy (%)	60.00	60.00	54.28	51.42	57.14	60.00	54.28	52.85	61.42	55.71	57.14	54.28	**75.71**
F1-score	0.602	0.603	0.517	0.510	0.571	0.602	0.535	0.529	0.617	0.560	0.575	0.547	**0.736**

- The largest increase in the accuracy before and after combining positive and negative narrations, is observed in the case of the audio features. It shows the efficiency of audio features in emotion and apathy detection.
- The fusion of multiple modalities produced a hike in performance. The result shows that each feature learned by different modality is complementary to other and are equally contributing for the task of automatic apathy detection.

Conclusion

- The proposed MIL based method exploits facial expressions, action units, facial landmarks and audio to detect apathetic and non-apathetic behaviour.
- Although, vision based facial expression recognition methods have achieved a very high accuracy, it is still difficult to use them to detect apathy in elderly.
- Another challenge is subjectivity issue. The intensity of expressions may vary for positive and negative videos. In future, the proposed network will be improved by considering the intensity of expressions.

Progress Report

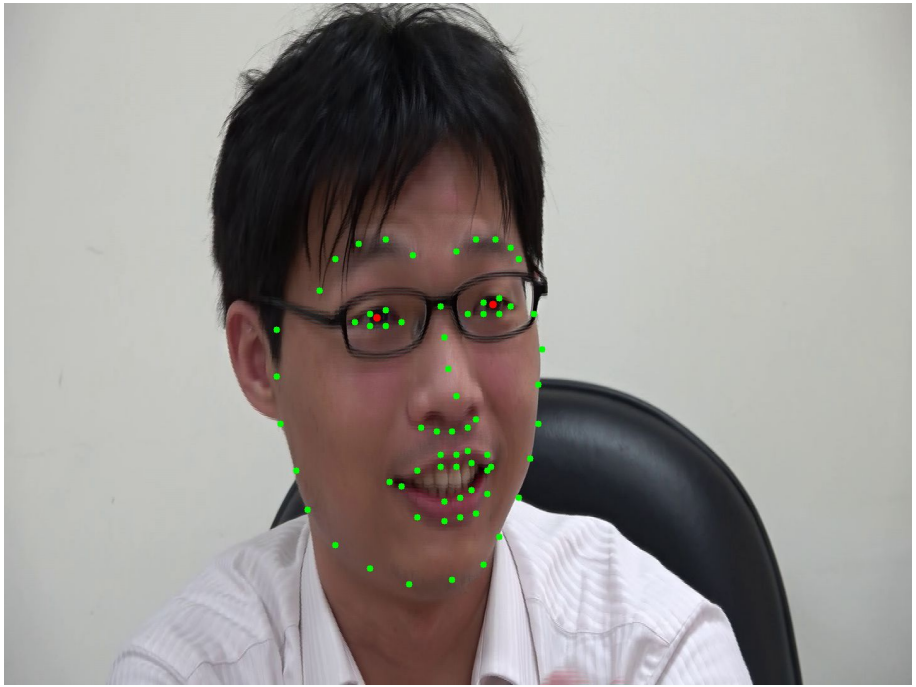
Datasets

- 24 videos
- Conversational Question Answering
- 臨床失智評估量表 CDR
- 0 健康 / 0.5 疑似輕微 / 1 輕度 / 2 中度
- CDR = 0 / 0.5 / 1 / 2 : 2 / 5 / 15 / 2 videos

GOAL

- Input Videos → **Model** → CDR Score (0/0.5/1/2)

Facial Landmarks & Pupils Detection



reference: Analyzing Facial and Eye Movements to Screen for Alzheimer's Disease

Global Features

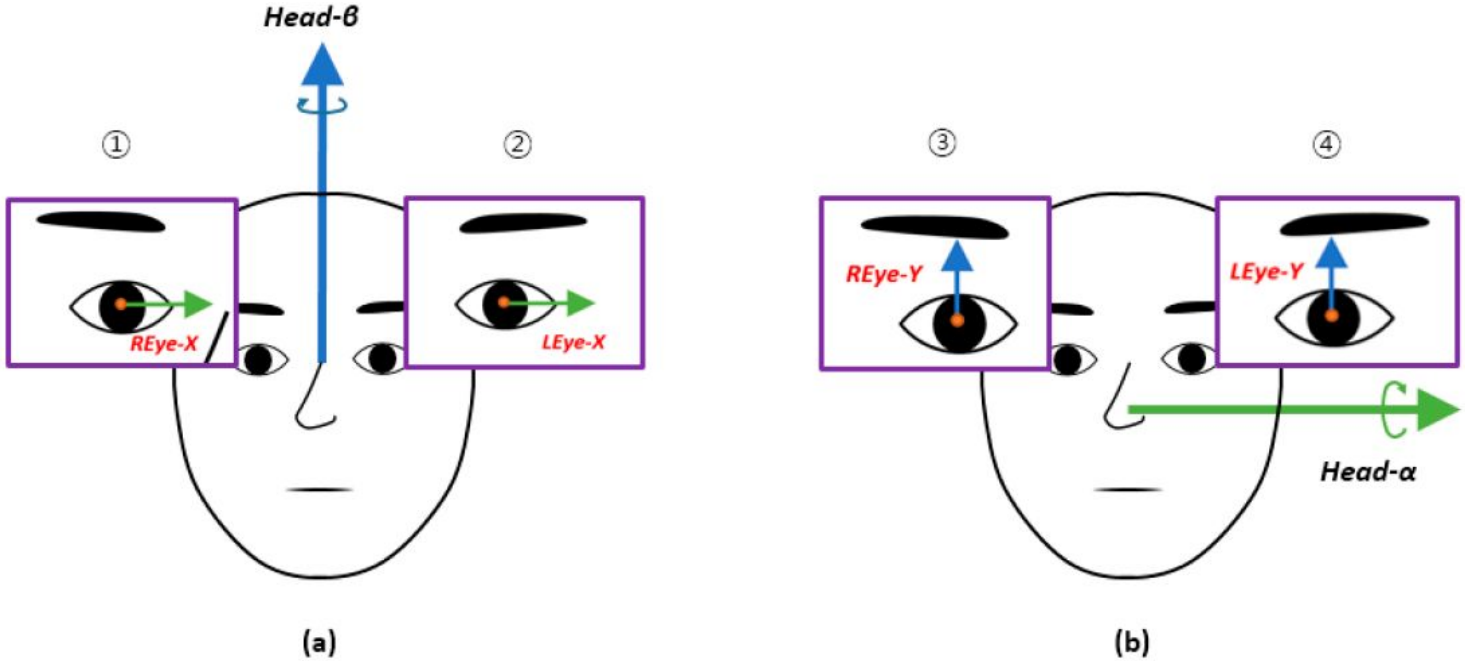


Figure 4. A pair of axes was used to obtain the correlation coefficient: (**a**) horizontal and (**b**) vertical.

Train SVM

Result

- True = [0. 2. 2. 1. 2. 2. 1. 1. 2. 3. 0. 2. 2. 1. 2. 2. 1. 1. 2. 3.]
- pred = [0. 1. 2. 1. 2. 1. 1. 1. 3. 0. 0. 1. 2. 2. 2. 1. 2. 1. 2. 0.]
- **Accuracy : 11/20**

Speech Features

- Segmented by QAs
- Extract MFCC Features
- Train SVM
  - True = [0. 2. 2. 1. 2. 2. 1. 1. 2. 3.]
  - pred = [0. 2. 2. 1. 2. 2. 2. 2. 2. 2.]
  - **Accuracy : 7/10**