Suppressing Uncertainties for Large-Scale Facial Expression Recognition

Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao
ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab, Shenzhen
Institutes of Advanced Technology, Chinese Academy of Science
University of Chinese Academy of Sciences, China
Nanyang Technological University Singapore

CVPR 2020

Outline

- Introduction
- Related Work
- Self-Cure Network
- Experiments
- Conclusion
- Progress Report

Introduction

- In the past decades, researchers have made significant progress on facial expression recognition (FER) with algorithms and large-scale datasets, where datasets can be collected in laboratory or in the wild.
- However, for the large-scale FER datasets collected from the Internet, it is extremely difficult to annotate with high quality due to the uncertainties.

Introduction

- Generally, training with uncertainties of FER may lead to the following problems.
  - First, it may result in overfitting on the uncertain samples which may be mislabeled.
  - Second, it is harmful for a model to learn useful facial expression features.
  - Third, a high ratio of incorrect labels even makes the model disconvergence in the early stage of optimization.

Introduction

- To address these issues, we propose a simple yet efficient method, termed as **Self-Cure Network (SCN)**. The SCN consists of three crucial modules:
  1. self-attention importance weighting
  2. ranking regularization
  3. noise relabeling
- We elaborately design a rank regularization to supervise the SCN to learn meaningful importance weights, which also provides a reference for the relabeling module.
- We extensively validate our SCN on synthetic FER data and a new real-world uncertain emotion dataset (WebEmotion) collected from the Internet.

Related Work

Facial Expression Recognition

- According to the facial feature type, they can be grouped into engineered features and learning-based features.
- Engineered features
  - SIFT [34], HOG [6], Histograms of LBP [35], Gabor wavelet coefficients [26], etc.
- Learned features
  - Facial Action Units based CNN [27]
  - region-based attention networks [25] [42]

[34] Pauline C. Ng and Steven Henikoff. Sift: predicting amino acid changes that affect protein function. Nucleic Acids Research, 31(13):3812–3814, 2003.
[35] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing, 27(6):803 – 816, 2009.
[42] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. arXiv preprint:1905.04075, 2019.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005.
[25] Y. Li, J. Zeng, S. Shan, and X. Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE Transactions on Image Processing, 28(5):2439–2450, May 2019.
[26] Chengjun Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image Processing, 11(4):467–476, April 2002.
[27] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-inspired deep networks for facial expression feature learning. Neurocomputing, 159(C):126–136, 2015.

Related Work

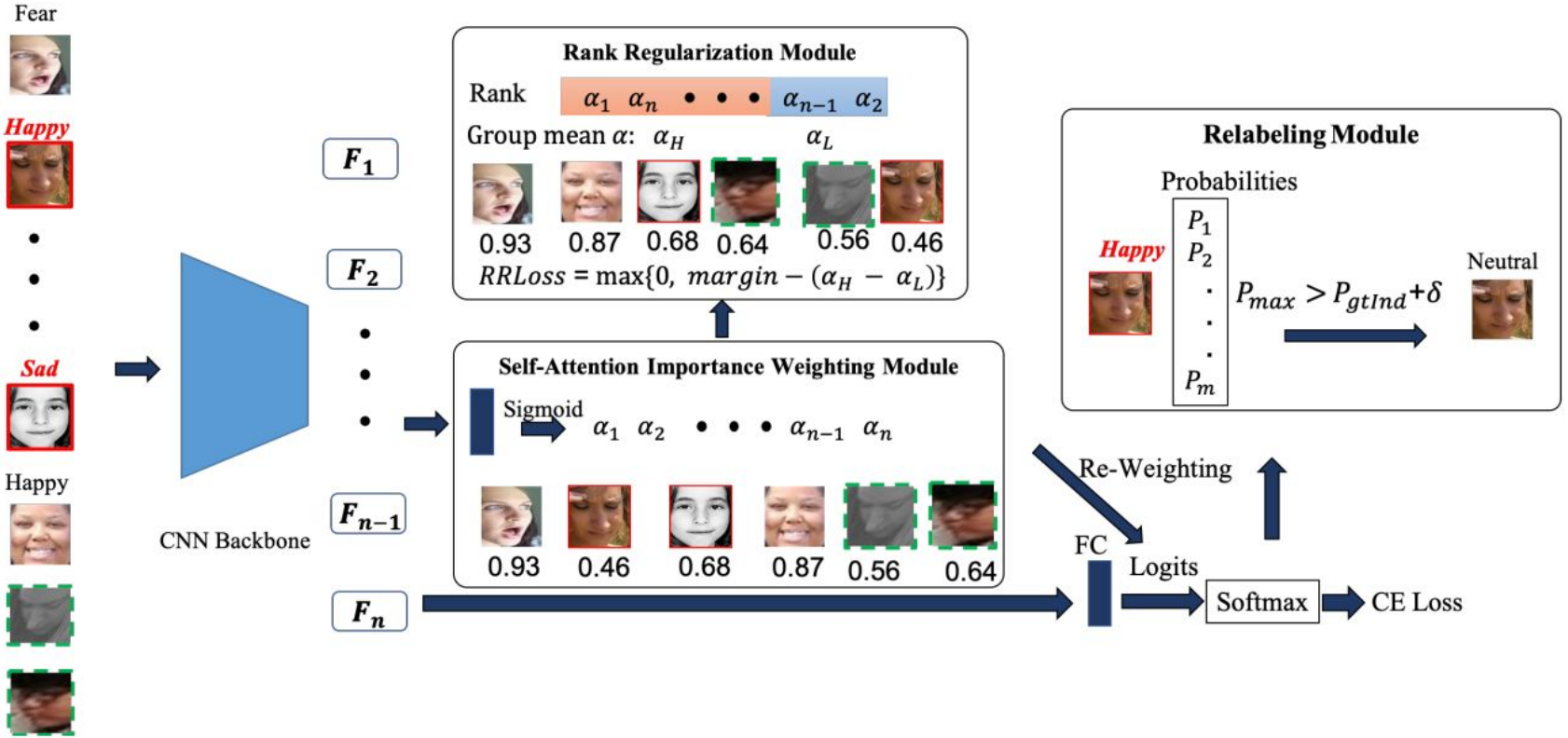
Learning with Uncertainties

- In order to handle noisy labels, one intuitive idea is to leverage a small set of clean data that can be used to assess the quality of the labels during the training process [40, 23, 8], or to estimate the noise distribution [36], or to train the feature extractors [3].
- For the FER task, Zeng et al. [43] first consider the inconsistent annotation problem among different FER datasets, and propose to leverage these uncertainties to improve FER.

[40] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In CVPR, pages 839–847, 2017.
[43] Jiabei Zeng, Shiguang Shan, Xilin Chen, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In ECCV, pages 222–237, 2018.

[3] Samaneh Azadi, Jiashi Feng, Stefanie Jegelka, and Trevor Darrell. Auxiliary image regularization for deep cnns with noisy labels. arXiv preprint:1511.07069, 2015.
[8] Mostafa Dehghani, Aliaksei Severyn, Sascha Rothe, and Jaap Kamps. Avoiding your teacher's mistakes: Training neural networks with controlled weak supervision. arXiv preprint 1711.00313, 2017.
[23] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In ICCV, pages 1910–1918, 2017.

Self-Cure Network



Self-Attention Importance Weighting

- The self-attention importance weighting module is comprised of a linear fully-connected (FC) layer and a sigmoid activation function.

$$\alpha_i = \sigma(\mathbf{W}_a^\top \mathbf{x}_i)$$

- In this paper, we choose the logit-weighted one of [17] which is shown to be more efficient.

$$\mathcal{L}_{WCE} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{\alpha_i \mathbf{W}_{y_i}^\top \mathbf{x}_i}}{\sum_{j=1}^{C} e^{\alpha_i \mathbf{W}_j^\top \mathbf{x}_i}}$$

- The self-attention weights in the above module can be arbitrary in (0, 1).

[17] Wei Hu, Yangyu Huang, Fan Zhang, and Ruirui Li. Noisetolerant paradigm for training face recognition cnns. In CVPR, pages 11887–11896, 2019.

Rank Regularization

- In the rank regularization module, we first rank the learned attention weights in descending order and then split them into two groups with a ratio β.
- The rank regularization ensures that the mean attention weight of high-importance group is higher than the one of low-importance group with a margin.

$$\mathcal{L}_{RR} = max\{0, \delta_1 - (\alpha_H - \alpha_L)\}$$

- where δ1 is a margin which can be a fixed hyper parameter or a learnable parameter.
- In training, the total loss function is L = γ LRR + (1 − γ) LWCE where γ is a trade-off ratio.

Relabeling

- A sample is assigned to a new pseudo label if the maximum prediction probability is higher than the one of given label with a threshold δ2.

$$y' = \begin{cases} l_{max} & \text{if } P_{max} - P_{gtInd} > \delta_2 \\ l_{org} & \text{otherwise,} \end{cases}$$

- In our system, uncertain samples are expected to obtain low importance weights thus to degrade their negative impacts with **re-weighting**, and then fall into the low importance group, and finally may be corrected as certain samples by **relabeling**.
- which is the reason why we call our method as **self-cured network**.

Experiments

Datasets

- RAF-DB [22]
  - 30000 images
  - 6 expressions
- FERPlus [4]
  - about 36000 images
  - 8 expressions
- AffectNet [32]
  - 450000 images
  - 8 expression
- The collected WebEmotion
  - 41,000 videos downloaded from YouTube

[4] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In ACM ICMI, 2016.
[22] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In CVPR, pages 2852–2861, 2017.
[24] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. TIP, 28(5):2439–2450, 2018.

Experiments

[14] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In ECCV, September 2018.
[46] Weihe Zhang, Yali Wang, and Yu Qiao. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In CVPR, June 2019.

Evaluation of SCN on Synthetic Uncertainties

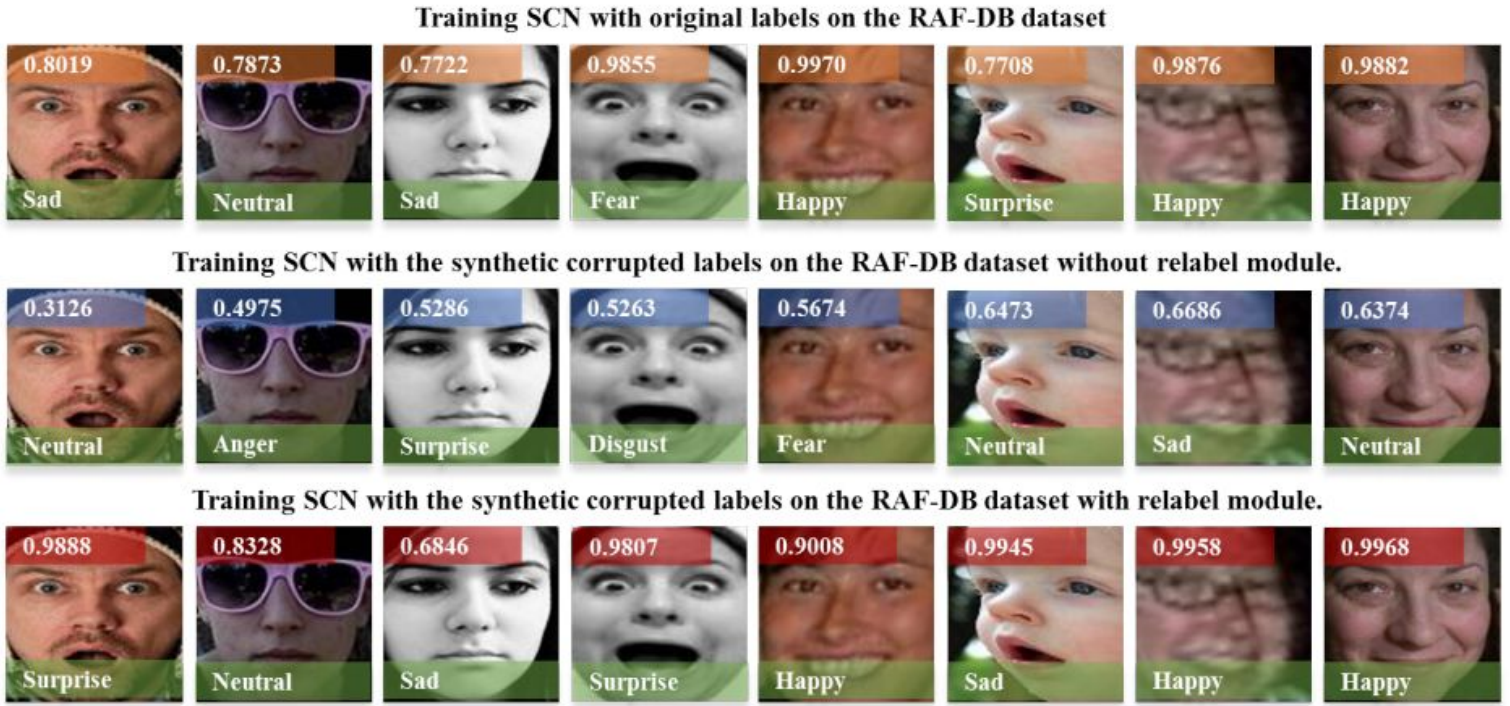
Pretrain	SCN	Noise(%)	RAF-DB	AffectNet	FERPlus
×	CurriculumNet [14]	10	68.5	-	-
×	MetaCleaner [46]	10	68.45	-	-
×	×	10	61.43	44.68	77.15
×	✓	10	**70.26**	**45.23**	**78.53**
×	CurriculumNet [14]	20	61.23	-	-
×	MetaCleaner [46]	20	61.35	-	-
×	×	20	55.5	41.00	71.88
×	✓	20	**63.50**	**41.63**	**72.46**
×	CurriculumNet [14]	30	57.52	-	-
×	MetaCleaner [46]	30	58,89	-	-
×	×	30	46.81	38.35	68.54
×	✓	30	**60.61**	**39.42**	**70.45**
✓	×	10	80.81	57.18	83.39
✓	✓	10	**82.18**	**58.58**	**84.28**
✓	×	20	78.18	56.15	82.24
✓	✓	20	**80.10**	**57.25**	**83.17**
✓	×	30	75.26	52.58	79.34
✓	✓	30	**77.46**	**55.05**	**82.47**

Pretrained model on Ms-Celeb-1M [15]

[15] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. CoRR, abs/1607.08221, 2016.

Experiments

Visualization of α in SCN



Experiments

Exploring SCN on Real-World Uncertainties

Our collected WebEmotion dataset consists of massive noises since the searching keywords are regarded as labels.

Table 3: The effect of SCN on WebEmotion for pretraining. The 2nd column indicates finetuning with or without SCN.

WebEmoition	SCN	RAF-DB	AffectNet	FERPlus
×	×	72.00	46.58	82.4
w/o SCN	×	78.97	56.43	84.20
w/o SCN	✓	80.42	57.23	85.13
SCN	✓	82.45	58.45	85.97

Experiments

Ablation Studies

Table 5: Evaluation of the three modules in SCN.

Weight	Rank	Relabel	RAF-DB	RAF-DB (pretrain)
×	×	×	72.00	84.20
×	×	✓	71.25	83.78
×	✓	×	74.15	85.14
✓	×	×	76.26	86.09
✓	✓	×	76.57	86.63
✓	✓	✓	78.31	87.03

Table 6: Evaluation of the ratio γ between RR-Loss and WCE-Loss.

0.2	0.3	0.5	0.6	0.8
76.12%	76.35%	**78.31%**	76.57%	71.75%

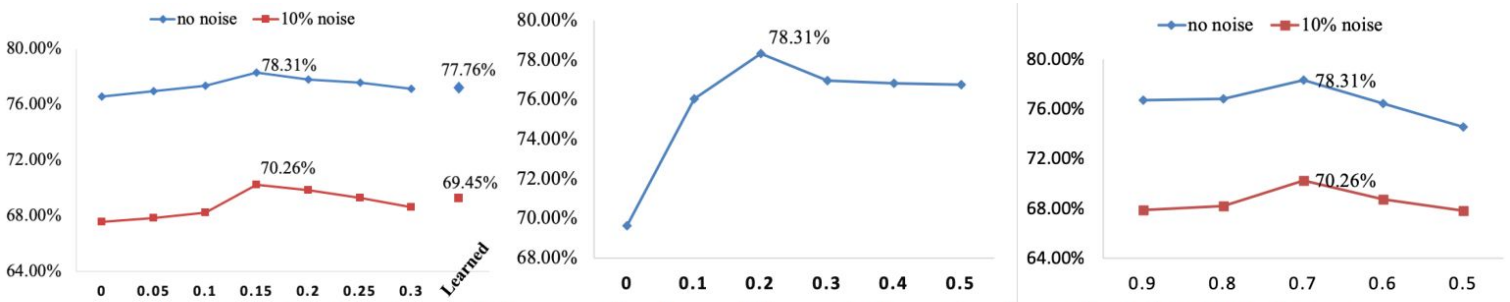


Figure 5: Evaluation of the margin δ_1 and δ_2, and the ratio β on the RAF-DB dataset.

Experiments

Comparison to the State of the Art

Table 7: Comparison to the state-of-the-art results. *These results are trained using label distributions. +Oversampling is used since AffectNet is imbalanced. ‡RAF-DB and AffectNet are jointly used for training. Note that IPA2LT tests with 7 classes on AffectNet.

(a) Comparison on RAF-DB.

Method	Acc.
DLP-CNN [22]	84.22
IPA2LT [43]	86.77
gaCNN [24]	85.07
RAN [42]	86.90
Our SCN (ResNet18)	**87.03**
Our SCN (ResNet18) ‡	**88.14**

(b) Comparison on AffectNet.

Method	mean Acc.
Upsample [32]	47.00
Weighted loss [32]	58.00
IPA2LT‡ [43] (7 cls)	55.71
RAN [42]	52.97
RAN+ [42]	59.5
Our SCN+(ResNet18)	**60.23**

(c) Comparison on FERPlus

Method	Acc.
PLD* [5]	85.1
ResNet+VGG [18]	87.4
SeNet50* [1]	88.8
RAN [42]	88.55
RAN-VGG16* [42]	89.16
Our SCN (ResNet18/IR50)	88.01/**89.35**

References

[1] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using crossmodal transfer in the wild. arXiv preprint arXiv:1808.05561, 2018.
[5] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In ACM ICMI, pages 279–283, 2016.
[18] Christina Huang. Combining convolutional neural networks for emotion recognition. In 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), pages 1–4, 2017.
[22] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In CVPR, pages 2852–2861, 2017.
[24] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. TIP, 28(5):2439–2450, 2018.
[32] Ali Mollahosseini, Behzad Hasani, Mohammad H Mahoor, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. TAC, 10(1):18–31, 2017.
[42] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. arXiv preprint:1905.04075, 2019.
[43] Jiabei Zeng, Shiguang Shan, Xilin Chen, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In ECCV, pages 222–237, 2018.

Conclusion

- This paper presents a self-cure network (SCN) to suppress the uncertainties of facial expression data thus to learn robust feature for FER.
- The SCN consists of three novel modules including self-attention importance weighting, ranking regularization, and relabeling.
- Our SCN achieves state-of-the-art results and can handle both synthetic and real-world uncertainties effectively.

Progress Report

Datasets

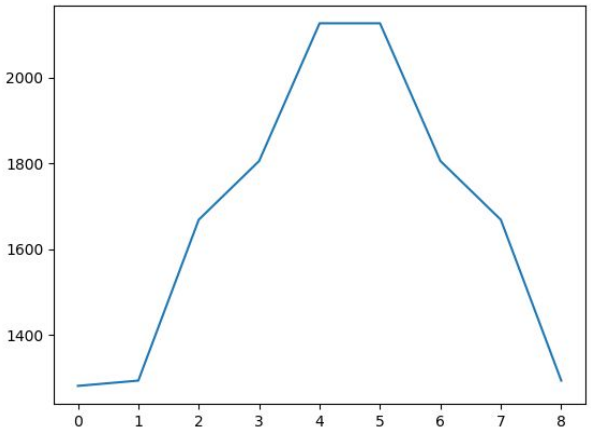
- 40 videos from 高榮
- Conversational Question Answering
- 臨床失智評估量表 CDR
- 0 健康 / 0.5 疑似輕微 / 1 輕度 / 2 中度
- CDR = 0.5 / 1 / 2 : 17 / 19 / 4 videos

GOAL

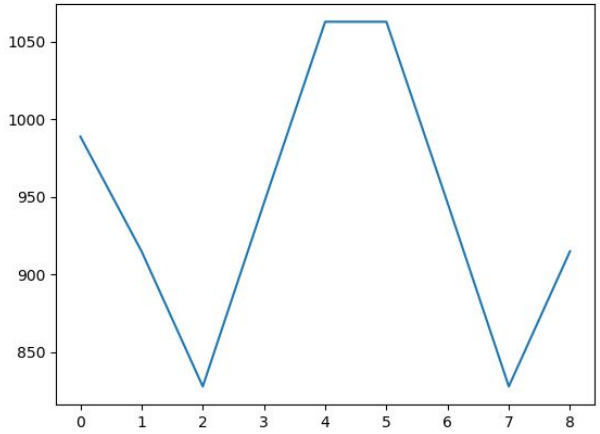
- Input Videos → **Model** → CDR Score (0.5 / 1 / 2 or 0.5 / 1+2)

Visual Features

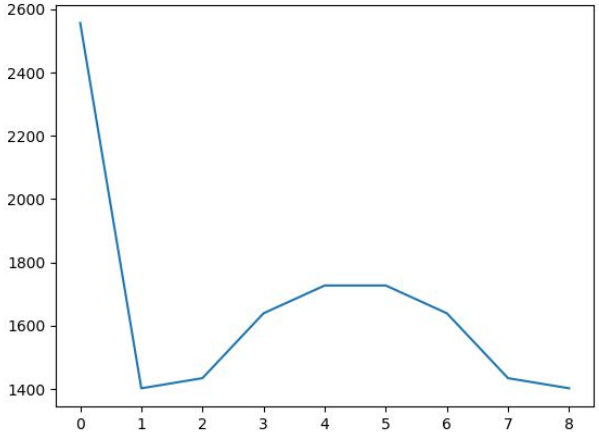
- Local Feature
  - Pupil Detection
  - P(x,y) → dx, dy → STFT (10 frames) → superposition



CDR = 0



CDR = 1



CDR = 2

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

Visual Features

- Global Feature
  - Pupils Detection & Head Pose Estimation
  - Pearson Correlation

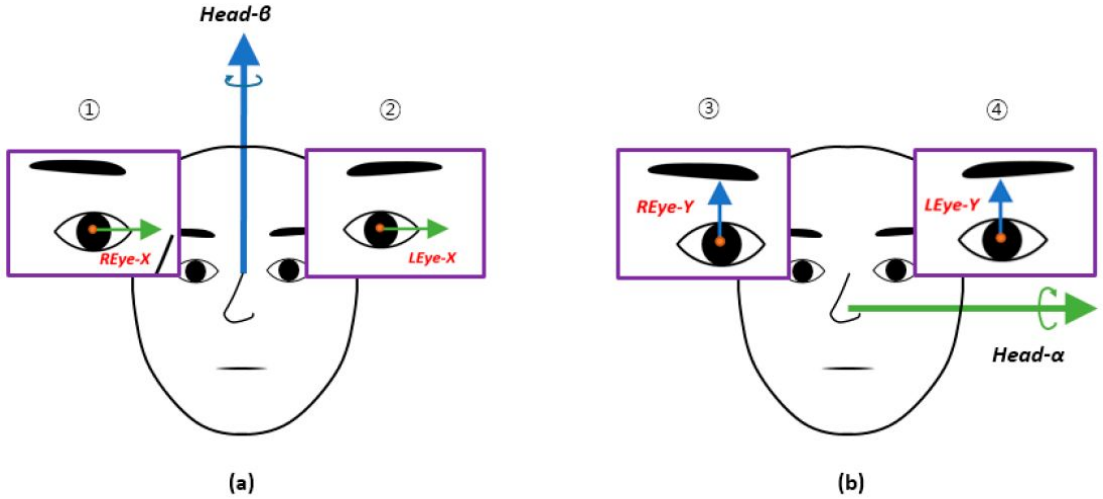


Figure 4. A pair of axes was used to obtain the correlation coefficient: (**a**) horizontal and (**b**) vertical.

Visual Features

- Local & Global Features
- Train SVM
  - version 1 - Label : 0 / 1 / 2 → CDR = 0.5 / 1 / 2
  - version 2 - Label : 0 / 1 → CDR = 0.5 / 1 + 2
- Result
  - version 1

    true = [1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 0. 0. 2. 1. 0. 2. 0. 0. 0. 0.]

    pred= [1. 0. 1. 0. 1. 1. 1. 2. 2. 1. 0. 0. 1. 1. 0. 1. 0. 0. 0. 1.]

    Accuray = 12/20 (60%)
  - version 2

    true = [1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 0. 0. 1. 1. 0. 1. 0. 0. 0. 0.]

    pred= [1. 1. 1. 0. 1. 1. 1. 1. 1. 1. 0. 0. 1. 1. 0. 1. 0. 1. 0. 1.]

    Accuray = 16/20 (80%)

Visual Features

- Data Augmentation (flip horizontally)
- Train SVM
  - version 1 - Label : 0 / 1 / 2 → CDR = 0.5 / 1 / 2
  - version 2 - Label : 0 / 1 → CDR = 0.5 / 1 + 2
- Result
  - version 1

    Accuray = 18/40 (45%)
  - version 2

    Accuray = 20/40 (50%)

Speech Features

- Mel-scale Frequency Cepstral Coefficients (MFCC)
- Train SVM
  - version 1 - Label : 0 / 1 / 2 → CDR = 0.5 / 1 / 2
  - version 2 - Label : 0 / 1 → CDR = 0.5 / 1 + 2
- Result
  - version 1
    true = [1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 0. 0. 2. 1. 0. 2. 0. 0. 0. 0.]
    pred= [0. 0. 0. 0. 0. 1. 0. 1. 0. 1. 0. 0. 0. 0. 0. 1. 0. 1. 0. 0.]
    Accuray = 8/20 (40%)
  - version 2
    true = [1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 0. 0. 1. 1. 0. 1. 0. 0. 0. 0.]
    pred= [0. 1. 0. 0. 1. 1. 1. 1. 0. 1. 1. 1. 1. 1. 0. 1. 0. 1. 1. 0.]
    Accuray = 11/20 (55%)

reference:
The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing
Opensmile: the munich versatile and fast open-source audio feature extractor

Speech Features

- GeMAPS defines a minimalistic feature set (Extracted using OpenSMILE).

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	

Speech Features

- Train SVM
  - version 1 - Label : 0 / 1 / 2 → CDR = 0.5 / 1 / 2
  - version 2 - Label : 0 / 1 → CDR = 0.5 / 1 + 2
- Result
  - version 1

    true = [1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 0. 0. 2. 1. 0. 2. 0. 0. 0. 0.]

    pred= [0. 1. 1. 1. 1. 1. 1. 0. 1. 1. 1. 0. 1. 1. 0. 1. 0. 0. 1. 1.]

    Accuray = 12/20 (60%)
  - version 2

    true = [1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 0. 0. 1. 1. 0. 1. 0. 0. 0. 0.]

    pred= [0. 1. 1. 0. 0. 1. 0. 1. 1. 0. 1. 1. 1. 1. 0. 1. 1. 1. 0. 1.]

    Accuray = 11/20 (55%)

Speech Features

- MFCC & GeMAPS
- Train SVM
  - version 1 - Label : 0 / 1 / 2 → CDR = 0.5 / 1 / 2
  - version 2 - Label : 0 / 1 → CDR = 0.5 / 1 + 2
- Result
  - version 1

    true = [1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 0. 0. 2. 1. 0. 2. 0. 0. 0. 0.]

    pred= [0. 0. 0. 0. 0. 0. 1. 1. 1. 0. 0. 0. 0. 1. 0. 0. 0. 1. 0. 0.]

    Accuray = 11/20 (55%)
  - version 2

    true = [1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 0. 0. 1. 1. 0. 1. 0. 0. 0. 0.]

    pred= [0. 1. 1. 0. 0. 1. 1. 0. 0. 1. 0. 0. 1. 1. 0. 1. 0. 1. 0. 0.]

    Accuray = 13/20 (65%)

Visual (augmented) & Speech Features

- Train SVM
  - version 1 - Label : 0 / 1 / 2 → CDR = 0.5 / 1 / 2
  - version 2 - Label : 0 / 1 → CDR = 0.5 / 1 + 2
- Result
  - version 1
    Accuray = 22/40 (55%)
  - version 2
    Accuray = 23/40 (57.5%)