AN END-TO-END LEARNING APPROACH FOR MULTIMODAL EMOTION RECOGNITION: EXTRACTING COMMON AND PRIVATE INFORMATION

Fei Ma, Wei Zhang, Yang Li, Shao-Lun Huang, Lin Zhang

Tsinghua-Berkeley Shenzhen Institute, Tsinghua University
mf17@mails.tsinghua.edu.cn, wzhang17@mails.tsinghua.edu.cn, yangli@sz.tsinghua.edu.cn,
shaolun.huang@sz.tsinghua.edu.cn, linzhang@tsinghua.edu.cn

ABSTRACT

Multimodal emotion recognition is important for facilitating efficient interaction between humans and machines. To better detect emotional states from multimodal data, we need to effectively extract both the common information that captures dependencies among different modalities, and the private information that characterizes variations in each modality. However, existing works are mostly designed to pursue either one of these objectives but not both. In our work, we propose an end-to-end learning approach to simultaneously extract the common and private information for multimodal emotion recognition. Specifically, we use a correlation loss based on Hirschfeld-Gebelein-Rényi (HGR) maximal correlation and a reconstruction loss based on autoencoders to preserve the common and private information, respectively. Experimental results on eNTERFACE'05 database and RML database demonstrate the effectiveness of our proposed approach.

Index Terms— Multimodal emotion recognition, end-to-end learning, Hirschfeld-Gebelein-Rényi (HGR) maximal correlation, autoencoder

1. INTRODUCTION

Multimodal emotion recognition aims to distinguish human emotional states by integrating various modalities, such as physiological signals, text, facial expression and speech. It is crucial to efficiently implement human-machine interaction and can be used for many applications, such as depression diagnosis, online learning and advertisement [1]. During the last decade, while multimodal emotion recognition has attracted a great deal of attention from research community, it is still a challenging task and an open research problem.

The key challenge in multimodal emotion recognition is data fusion, which incorporates the information of each modality by learning the feature representations from multimodal data. A number of previous works [2, 3, 4, 5] have studied this problem. However, due to the shallow fusion structure used in these works, they are not able to relate the feature representations in one modality to those in other modalities, which limits the performance to some extent. To capture statistical interactions between modalities, a few correlation-based approaches have been proposed, among which the methods based on canonical correlation analysis (CCA) [6] are typical. For example, in [7], Li et al. employed the kernel canonical correlation analysis (KCCA) approach to fuse the hand-crafted features in reproducing kernel Hilbert space. In [8], Sarvestani et al. proposed a new feature fusion method based on sparse kernel probabilistic canonical correlation analysis (SKPCCA), which unifies the latent variables of two modalities into a feature vector with an acceptable size. In [9], Qiu et al. adopted deep canonical correlation analysis (DCCA) to learn the non-linear feature transformations of two modalities into a highly correlated space.

However, since CCA-based approaches often suffer from numerical issues, the feature dimensionalities of the aforementioned methods are restricted to be relatively small to ensure stable implementation [10]. This is undesirable in real-world settings where the dimensionalities of data should be large enough to handle complicated tasks [11]. Besides, most of the existing correlation-based works usually omit the effect of each modality on emotion recognition. In this case, the learned feature mappings using such strategies do not capture much dependencies and are not able to preserve the unimodal information reliably, which leads to performance degradation.

To clarify the above two issues, we introduce the concepts of common and private information for high-level multimodal learning. Common information is an important topic in statistics and information theory, which focuses on measuring the dependence of two variables [12]. We adopt it to describe the correlation between different modalities, which determines the effectiveness of integrating the various modal information. Private information is also significant for emotion recognition. It is observed in [13] that visual modality typically provides better performance than others. Under this circumstance, only maximizing correlation between different modalities may do harm to the recognition performance. We thus employ private information here to measure the sufficiency of feature mappings representing the unimoodal data. By extracting common and private information simultaneously, we

expect our approach to effectively capture the complicated dependencies among different modalities and fully represent the data source.

To achieve this, we propose an efficient system in an end-to-end learning framework. Firstly, We adopt convolutional neural network(CNN) and Long short-term memory(LSTM) to automatically learn the high-level affective features from the multimodal data directly. Such an end-to-end learning can increase the performance of emotion recognition over hand-crafted features [14]. Secondly, we use Hirschfeld-Gebelein-Rényi (HGR) maximal correlation [15, 16, 17], a well-known measure of dependence, to determine the maximally non-linear correlated feature mappings of different modalities. In [11], Wang et al. proposed a deep learning implementation of HGR maximal correlation with the Soft-HGR loss, which has high efficiency and stability. It is employed here to extract the common information by learning the highly non-linear feature transformations of the data input. We further capture the private information through the reconstruction loss of autoencoders to warranty the preservation of the unimodal information. Finally, the concatenated features are fed into the softmax classifier for classification. We conduct experiments on the two public emotional databases, eNTERFACE'05 [18] and RML [19, 20]. The results demonstrate that our proposed architecture performs better the state-of-the-art methods.

To summarize, our main contributions are as follows:

- We propose an end-to-end learning framework to automatically learn the high-level discriminative feature combinations from the raw data with CNN and LSTM.
- We design an efficient multimodal learning approach by extracting the common and private information with the correlation loss and reconstruction loss.
- Our experiments on eNTERFACE'05 database and RML database demonstrate the effectiveness of our multimodal emotion recognition system.

2. RELATED WORK

There are some typical approaches to extract the correlation among different modalities. In statistics, canonical correlation analysis (CCA) is a generalization from the Pearsons correlation [21]. It extracts the feature projections of two modalities, which are linearly maximally correlated. As an extension of CCA, deep canonical correlation analysis (DCCA) is introduced in [22] to learn the non-linear feature representations through deep neural networks. HGR maximal correlation is another well-known measure of dependence. It determines the maximally non-linear correlated feature mappings of two types of data. CCA can be considered as the realization of the HGR maximal correlation with the linear form [11]. Recently, Soft-HGR loss is introduced in [11] as a development of HGR maximal correlation to extract informative features for each data source with deep learning. It is much easier to implement than the conventional way and also presents better efficiency and stability than the traditional CCA methods.

The autoencoder architecture is also appealing for multimodal learning. For example, in [23], Ngiam et al. proposed a deep autoencoder to better learn the features for one modality when multiple modalities are present during feature learning. In [24], Wang et al. developed the deep canonically correlated autoencoders (DCCAE) from DCCA by optimizing the combination of nonlinear canonical correlation and the reconstruction errors of the autoencoders to learn the feature representations for downstream discriminative tasks.

It is worth noting that [1] also involves Soft-HGR loss for multimodal emotion recognition. However, our work is essentially different since that in [1], the private information is not investigated deeply and the structure of the neural netowrk used in this paper is too simple with the hand-crafted features as input, resulting in limited performance.

3. METHODOLOGY

Our aim is to learn the high-level affective feature representations for mutlimodal emotion recognition, satisfying the following objectives: (i) the highly non-linear correlation of the feature combinations among various modalities is learned, thus the single modality could embrace the common information from other modalities, (ii) the learned features of each modal data are sufficient to represent the corresponding modality. To achieve all the above goals, we propose an end-to-end multimodal architecture. The pipeline of our approach is shown in Fig. 1.

3.1. Problem formulations

In this section, we use the bimodal scenario to explain our architecture, which can be extended to more modalities. Given paired observations from bimodal data $\big\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}, z^{(i)}) \mid \mathbf{x}^{(i)} \in \mathbb{R}^{D_\mathbf{x}}, \mathbf{y}^{(i)} \in \mathbb{R}^{D_\mathbf{y}}, z^{(i)} \in \mathcal{Z} = \{1, 2, \cdots, |\mathcal{Z}|\}, i = 1, \cdots, m\big\}$. For example, let $\mathbf{x}$ and $\mathbf{y}$ denote the visual and audio data with dimensionalities $\mathbb{R}^{D_\mathbf{x}}$ and $\mathbb{R}^{D_\mathbf{y}}$ respectively and z denote the associated category label. Suppose $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_k(\mathbf{x})]^{\mathrm{T}}$ and $\mathbf{g}(\mathbf{y}) = [g_1(\mathbf{y}), g_2(\mathbf{y}), \cdots, g_k(\mathbf{y})]^{\mathrm{T}}$ represent the k-dimensional feature functions of $\mathbf{x}$ and $\mathbf{y}$ separately. The goal is to learn such feature representations from data sources such that the corresponding labels z can be correctly predicted.

3.2. Proposed model

The architecture of our system can be mainly divided into two components: the feature extraction module and the classifier module. Here, we combine the feature extraction and the classifier together, which makes the best use of the dependence between the learned feature representations and target information. The whole architecture is designed in the end-to-end fashion.

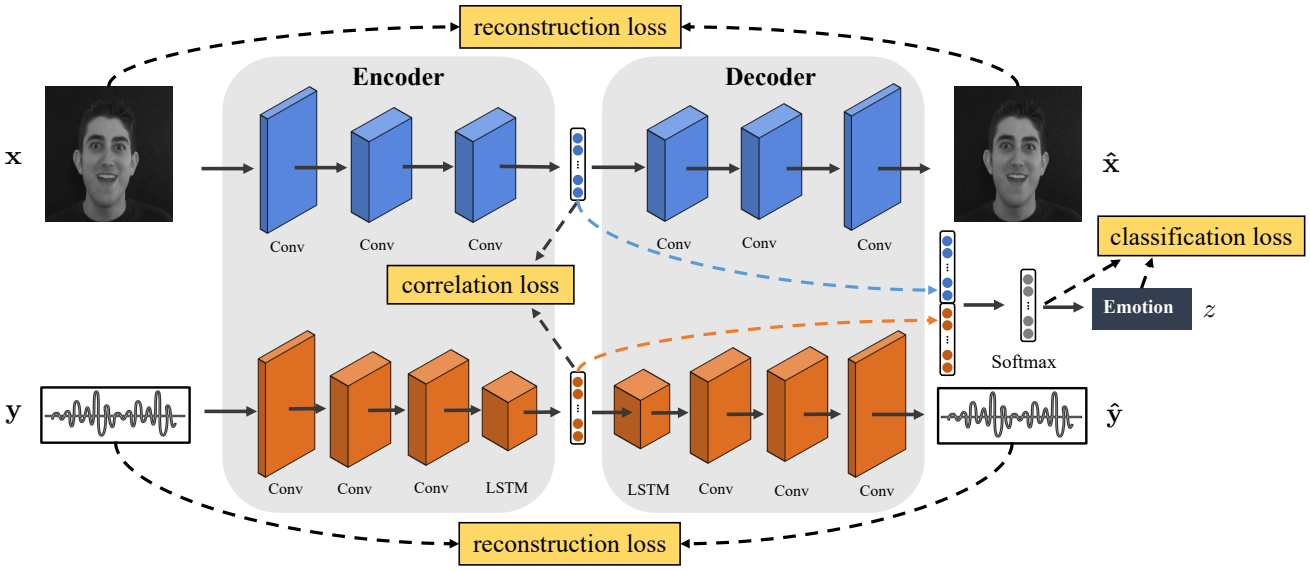

Fig. 1: The structure of our proposed system for multimodal emotion recognition. The full objective of the system consists of three parts: correlation loss, reconstruction loss and classification loss.

The feature extraction mudule is composed of two networks, one for visual data and another for audio data. Each network comprises an encoder that learns a latent representation of the data source and a decoder that reconstructs the input data from the representation. We utilize $\mathbf{f}$ to represent the visual encoder network. After we split video samples into 1-second segments, the key frames of every segment are resized into 128×128. They are then fed into the visual network to extract visual affective features from visual data. This network is composed of three 3×3 convolutional neural layers and a fully connected layer with 512 ReLU units sequentially. Each of the convolutional neural layer has 32, 64, 64 filters respectively, followed by a batch-normalization layer and a 2×2 max-pooling layer. The output of the last convolutional neural layer is flattened and sent into the fully connected layer.

Similarly, $\mathbf{g}$ represents the encoder network to extract audio features from audio data. The audio signals of the 1-second segments are used as input. At 22.05kHz this corresponds to a 22050-dimensional input vector. Inspired by [25], we set the network to be composed of three convolutional neural layers, a LSTM layer with 256 units and a fully connected layer with 512 ReLU units sequentially. Each of the convolutional neural layer has 64, 128, 256 filters respectively, followed by a batch-normalization layer and a max-pooling layer. The kernel size for the three layers is 8, 6, 6 and the corresponding max-pooling size is set to 10, 5, 3 separately. After the LSTM layer, the dropout strategy is used to prevent overfitting with probability 0.5. The output of the LSTM layer is sent into the fully connected layer.

The decoders for the visual and audio data leverage symmetric architectures to encoders, denoted as $\mathbf{f_{rec}}$ and $\mathbf{g_{rec}}$ separately. Here, the dimensionality of the latent feature representations is set to 512 for sufficient expressive ability. It is expected the representations discriminate emotions correctly with the learned common and private information. To extract such representations, we exploit two types of loss functions: Soft-HGR loss and reconstruction loss of autoencoders.

We further feed the concatenated feature mapping to the classifier module for classification. Correspondingly, the classification loss is added.

3.3. Model Learning

The optimization objective considers the feature extraction and the classifier simultaneously for better performance. We define the overall objective of the network as Eq.(1):

$$\mathcal{L} = \alpha \mathcal{L}_{corr} + \beta \mathcal{L}_{rec} + \gamma \mathcal{L}_{clf} \tag{1}$$

The first two terms are loss functions measuring the common and private information among various modalities respectively. The last term is the classification loss. The parameters α, β, γ are weighting factors which control the relative importance of each term.

As shown in Eq. (2), we adopt Soft-HGR loss [11] as the correlation loss to extract the HGR maximal correlation between visual and acoustic signals. From the perspective of information theory, minimizing Eq. (2) is equivalent to extracting the common information from multimodal data.

$$\begin{aligned} L_{corr} &= -\mathbb{E}[\mathbf{f}^{\mathrm{T}}(\mathbf{x})\mathbf{g}(\mathbf{y})] \\ &\quad + \frac{1}{2}\mathrm{tr}(\mathrm{cov}(\mathbf{f}(\mathbf{x}))\mathrm{cov}(\mathbf{g}(\mathbf{y}))) \end{aligned} \tag{2}$$

In the meanwhile, the mean of such information features are constrained to zero. It has been proved that optimizing the Soft-HGR loss preserves the same amount of information as the HGR maximal correlation in a more efficient way [11].

Private information is also crucial, which guarantees the learned representations can fully represent unimodal data for discriminative tasks. Inspired by [24] , we feed the visual and audio feature representations, $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{y})$, into the corresponding decoder networks, $\mathbf{f_{rec}}$ and $\mathbf{g_{rec}}$. The outputs are denoted as $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ respectively, i.e., $\hat{\mathbf{x}} = \mathbf{f_{rec}}(\mathbf{f}(\mathbf{x}))$ and $\hat{\mathbf{y}} = \mathbf{g_{rec}}(\mathbf{g}(\mathbf{y}))$. With the reconstruction errors shown in Eq. (3), this framework is expected to reconstruct each modality from its representation accurately so that private information can be incorporated.

$$\mathcal{L}_{rec} = \mathbb{E}[||\mathbf{x} - \hat{\mathbf{x}}||^2 + ||\mathbf{y} - \hat{\mathbf{y}}||^2] \tag{3}$$

The concatenated learned feature representation is then fed into a softmax classifier. Following [11], we use the cross-entropy loss function as the classification objective to incorporate the label information, expressed as Eq. (4):

$$\mathcal{L}_{clf} = -\mathbb{E}[\log P_{z|\mathbf{xy}}] \tag{4}$$

where

$$P_{z=j|\mathbf{xy}} = \frac{\exp([\mathbf{f}^{\mathrm{T}}(\mathbf{x}), \mathbf{g}^{\mathrm{T}}(\mathbf{y})]\boldsymbol{\theta}_j)}{\sum_{i=1}^{|\mathcal{Z}|} \exp([\mathbf{f}^{\mathrm{T}}(\mathbf{x}), \mathbf{g}^{\mathrm{T}}(\mathbf{y})]\boldsymbol{\theta}_i)} \tag{5}$$

In Eq. (5), $j = 1, \cdots, |\mathcal{Z}|$ and $[\mathbf{f}^{\mathrm{T}}(\mathbf{x}), \mathbf{g}^{\mathrm{T}}(\mathbf{y})]$ represents the concatenation of $\mathbf{f}^{\mathrm{T}}(\mathbf{x})$ and $\mathbf{g}^{\mathrm{T}}(\mathbf{y})$.

4. EXPERIMENTS

4.1. Dataset

In order to evaluate the effectiveness of our strategy, we conduct extensive experiments on the two public audio-visual emotional databases, including eNTERFACE'05 and RML.

The eNTERFACE'05 database is composed of 1293 English video clips from 44 subjects with 14 different nationalities. Each utterance is categorized into six basic emotional states, including anger, disgust, fear, happiness, sadness and surprise, which are almost balanced. The samples are recorded at a sampling rate of 48000 Hz with a frame rate of 25fps. The utterance durations vary from 1.12 seconds to 106.92 seconds. The average duration is 3.17 seconds.

The RML database contains 720 video samples from 8 subjects speaking six different languages. It contains the same six emotions as the eNTERFACE'05 database. Every emotion has 120 samples. The samples are recorded at a sampling rate of 22050 Hz with a frame rate of 30fps. The utterance durations vary from 2.13 seconds to 8.37 seconds and the average duration is 4.94 seconds. Each video sample is truncated to a length of 2 seconds as in [20].

4.2. Data Preprocessing and Experimental Setup

Since video clips of the two databases vary in time duration generally, we split each of them into segments of the same length and extract audio signals and images from them, following [26]. Before transforming clips into segments, the window size and window shift of the segments need to be determined. Although the best window size for emotion recognition is still not clear, it is reported in [27] that a segment longer than 0.25 seconds includes sufficient emotional information. We set the window size and window shift to be 1 seconds and 0.4 seconds respectively. The utterance-level label is used as the label for the transformed segments.

After dividing the video clips into segments, we take the central frame in each segment as the key frame and crop a gray face image with the size 128×128. For consistency, the samples of the eNTERFACE'05 database are downsampled to 22.05 kHz. The audio signals and images are converted into the WAV and JPEG formats respectively and fed into our proposed networks to classify the segment with a particular emotion. After the emotion probabilities of each segment are predicted, average results across all segments belonging to the same video clips are used to predict the utterance-level emotion labels.

We employ the 5-fold cross validation protocol on the original databases to robustly estimate the recognition performance of our proposed system. For each database, the folds are randomly selected. The segments which belong to the same video clip are assigned to the training data or test data together. We use the grid search technique to determine the parameters in Eq. (1). They are set as: $\alpha = 300, \beta = 1, \gamma = 500$. The learning rate is set to 0.0001. Adam algorithm is used as the optimizer. The average recognition accuracy(%) is reported as the final result.

4.3. Experimental Results and Analysis

We first show the effectiveness of our feature representations learning by comparing the performance of audio, visual and audio-visual emotion recognition on eNTERFACE'05 and RML databases, as shown in Table 1. It can be seen that the learned features in our fusion strategy help improve the emotion classification significantly compared to a single modality. To be specific, our system improves the accuracies on the eNTERFACE'05 database from 58.95% for audio features and 83.21% for visual features to the 85.43%. Similarly, the accuracies are improved from 72.44% for audio features and 80.77% for visual features to the 86.89% on the RML database.

Table 1: Recognition performance of our method.

	Audio	Visual	Audio-Visual
eNTERFACE'05	58.95	83.21	85.43
RML	72.44	80.77	86.89

Classification results for each emotion in the audio-visual case are illustrated in Table 2. It can be found that "disgust" and "surprise" are harder to be identified compared to others in the eNTERFACE'05 database but achieve higher accuracies in the RML database. On the contrary, "anger" obtains the highest accuracy in the eNTERFACE'05 database while it

is hard to be predicted correlctly in the RML database. This indicates audio-visual cues of the two databases are different.

Table 2: Classification results for each emotion.

	eNTERFACE'05	RML
Anger	93.42	83.33
Disgust	82.23	92.00
Fear	82.60	77.33
Happiness	87.41	92.67
Sadness	85.93	84.00
Surprise	82.23	92.00

In order to investigate the effect of feature dimension, we implement audio-visual emotion recognition with different feature dimensionalities. The experimental results on both databases are reported in Fig. 2. When the feature dimensionality is low, both of the overall recognition accuracies have an increasing tendency as it increases. However, when the feature dimensionality exceeds a certain value, the performance will not be improved or even degraded. This shows that the extracted features at higher dimension may contain redundant or noisy information [10]. Finally, the highest recognition accuracies are both achieved when the feature dimensionality is 512. This explains why we set the dimension of $\mathbf{f}$ and $\mathbf{g}$ at this level in our work.

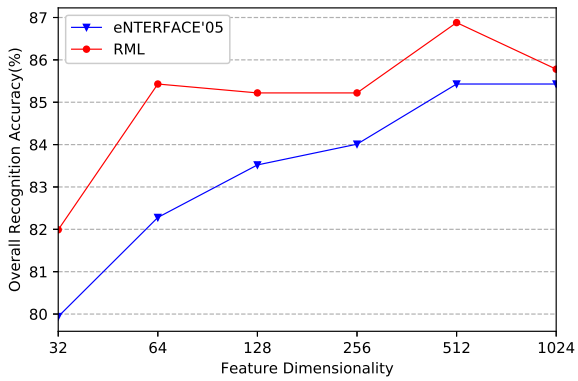

Fig. 2: The effect of feature dimensionality on the overall recognition accuracy.

We next compare our method against several state-of-the-art correlation-based multimodal learning methods including DCCA [22], DCCAE [24] and Soft-HGR [11]. To maintain consistency, the structure and parameters of their networks are set similarly as those of ours. These deep learning methods belong to the unsupervised approaches, which extract the feature transformations without including the supervised information. The learned representations are then fed into two-layer neural networks for classification. Fig. 3 shows the emotion recognition performance of these methods. The observations can be summarized as follows: (1) DCCAE achieves a higher accuracy than Soft-HGR and DCCA on our task, which confirms power of private information again. (2) Our method outperforms all of these unsupervised methods. The results show that the discriminative information might be lost when data are projected into lower dimensions without the guidance of supervised targets.

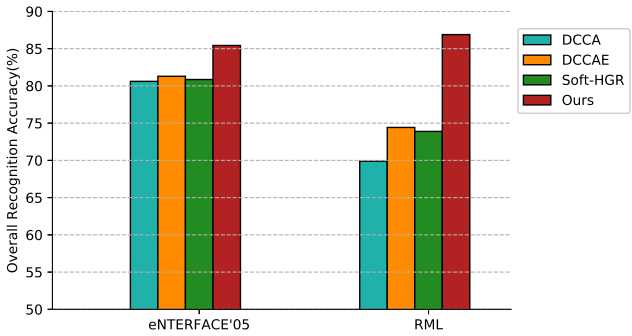

Fig. 3: Comparisons with correlation-based methods.

Finally, we compare our work with previous works on the eNTERFACE'05 and RML databases, as shown in Table 3. We can find that our method is competitive with the compared works [3, 4, 5, 10]. Besides, these works require careful selection of hand-crafted features while our method is implemented in an end-to-end way. Furthermore, all these works cannot sufficiently take the advantage of common and private information while our method can benefit from both together. This indicates that the power of our method may be gained from the high-level affective features preserving enough common and private information captured from the automatic feature learning procedure.

Table 3: Comparisons with previous works.

	Method	Accuracy(%)
eNTERFACE'05	Hossain et al., [4]	83.06
	Dobrišek et al., [5]	77.50
	Wang et al., [10]	72.47
	ours	**85.43**
RML	Fadil et al., [3]	79.72
	Wang et al., [10]	82.22
	ours	**86.89**

5. CONCLUSION

In this paper, we propose an efficient deep learning approach for multimodal emotion recognition. Specifically, we design deep networks with CNN and LSTM to learn the feature mappings from multimodal data in an end-to-end approach. Furthermore, the correlation loss based on HGR maximal correlation and the reconstruction loss of autoencoders are introduced to capture common and private information among different modalities respectively. The experimental results demonstrate our system outperforms the state-of-the-art methods. Such framework is highly flexible and can be further extended to handle the missing modality problems.

6. ACKNOWLEDGMENTS

The research of Shao-Lun Huang was funded by the Natural Science Foundation of China 61807021, Shenzhen Science and Technology Research and Development

Funds (JCYJ20170818094022586), and Innovation and entrepreneurship project for overseas high-level talents of Shenzhen (KQJSCX2018032714403783). The research of Lin Zhang was funded by Shenzhen Science and Technology Research and Development Funds (GJHZ20170314112258560).

7. REFERENCES

[1] Wei Zhang, Weixi Gu, Fei Ma, Shiguang Ni, Lin Zhang, and Shao-Lun Huang, "Multimodal emotion recognition by extracting common and modality-specific information," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2018, pp. 396–397.

[2] Muharram Mansoorizadeh and Nasrollah Moghaddam Charkari, "Multimodal information fusion application to human emotion recognition from face and speech," *Multimedia Tools and Applications*, vol. 49, no. 2, pp. 277–297, 2010.

[3] C Fadil, R Alvarez, C Martinez, J Goddard, and H Rufiner, "Multimodal emotion recognition using deep networks," in *VI Latin American Congress on Biomedical Engineering CLAIB 2014, Paraná, Argentina 29, 30 & 31 October 2014*. Springer, 2015, pp. 813–816.

[4] M Shamim Hossain and Ghulam Muhammad, "Audio-visual emotion recognition using multi-directional regression and ridgelet transform," *Journal on Multimodal User Interfaces*, vol. 10, no. 4, pp. 325–333, 2016.

[5] Simon Dobrišek, Rok Gajšek, France Mihelič, Nikola Pavešić, and Vitomir Štruc, "Towards efficient multi-modal emotion recognition," *International Journal of Advanced Robotic Systems*, vol. 10, no. 1, pp. 53, 2013.

[6] Harold Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[7] Bo Li, Lin Qi, and Lei Gao, "Multimodal emotion recognition based on kernel canonical correlation analysis," in *Electronics, Computer and Applications, 2014 IEEE Workshop on*. IEEE, 2014, pp. 934–937.

[8] Reza Rohani Sarvestani and Reza Boostani, "Ff-skpcca: Kernel probabilistic canonical correlation analysis," *Applied Intelligence*, vol. 46, no. 2, pp. 438–454, 2017.

[9] Jie-Lin Qiu, Wei Liu, and Bao-Liang Lu, "Multi-view emotion recognition using deep canonical correlation analysis," in *International Conference on Neural Information Processing*. Springer, 2018, pp. 221–231.

[10] Yongjin Wang, Ling Guan, and Anastasios N Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 597–607, 2012.

[11] Lichen Wang, Jiaxiang Wu, Shao-Lun Huang, Lizhong Zheng, Xiangxiang Xu, Lin Zhang, and Junzhou Huang, "An efficient approach to informative feature extraction from multimodal data," *arXiv preprint arXiv:1811.08979*, 2018.

[12] Aaron Wyner, "The common information of two dependent random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 163–179, 1975.

[13] Jingwei Yan, Wenming Zheng, Zhen Cui, Chuangao Tang, Tong Zhang, and Yuan Zong, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, 2018.

[14] Sefik Emre Eskimez, Zhiyao Duan, and Wendi Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *ICASSP*. IEEE, 2018, pp. 5099–5103.

[15] Hermann O Hirschfeld, "A connection between correlation and contingency," in *Mathematical Proceedings of the Cambridge Philosophical Society*. Cambridge University Press, 1935, vol. 31, pp. 520–524.

[16] Hans Gebelein, "Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung," *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 21, no. 6, pp. 364–379, 1941.

[17] Alfréd Rényi, "On measures of dependence," *Acta mathematica hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.

[18] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas, "The enterface05 audio-visual emotion database," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*. IEEE, 2006, pp. 8–8.

[19] Yongjin Wang and Ling Guan, "Recognizing human emotional state from audiovisual signals," *IEEE transactions on multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[20] Zhibing Xie and Ling Guan, "Multimodal information fusion of audio emotion recognition based on kernel entropy component analysis," *International Journal of Semantic Computing*, vol. 7, no. 01, pp. 25–42, 2013.

[21] Karl Pearson, "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.

[22] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255.

[23] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[24] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes, "On deep multi-view representation learning," in *International Conference on Machine Learning*, 2015, pp. 1083–1092.

[25] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller, "End-to-end speech emotion recognition using deep neural networks," in *ICASSP*. IEEE, 2018, pp. 5089–5093.

[26] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian, "Learning affective features with a hybrid deep model for audio-visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[27] Yelin Kim and Emily Mower Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," in *ICASSP*. IEEE, 2013, pp. 3677–3681.