Facial Expression Recognition Enhanced by Thermal Images through Adversarial Learning

Bowen Pan and Shangfei Wang*
Key Lab of Computing and Communication Software of Anhui Province
University of Science and Technology of China
Hefei, Anhui, P.R.China
bowenpan@mail.ustc.edu.cn;sfwang@ustc.edu.cn

ABSTRACT

Currently, fusing visible and thermal images for facial expression recognition requires two modalities during both training and testing. Visible cameras are commonly used in real-life applications, and thermal cameras are typically only available in lab situations due to their high price. Thermal imaging for facial expression recognition is not frequently used in real-world situations. To address this, we propose a novel thermally enhanced facial expression recognition method which uses thermal images as privileged information to construct better visible feature representation and improved classifiers by incorporating adversarial learning and similarity constraints during training. Specifically, we train two deep neural networks from visible images and thermal images. We impose adversarial loss to enforce statistical similarity between the learned representations of two modalities, and a similarity constraint to regulate the mapping functions from visible and thermal representation to expressions. Thus, thermal images are leveraged to simultaneously improve visible feature representation and classification during training. To mimic real-world scenarios, only visible images are available during testing. We further extend the proposed expression recognition method for partially unpaired data to explore thermal images' supplementary role in visible facial expression recognition when visible images and thermal images are not synchronously recorded. Experimental results on the MAHNOB Laughter database demonstrate that our proposed method can effectively regularize visible representation and expression classifiers with the help of thermal images, achieving state-of-the-art recognition performance.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**;

KEYWORDS

facial expression recognition; privileged information; adversarial learning

ACM Reference Format:
Bowen Pan and Shangfei Wang. 2018. Facial Expression Recognition Enhanced by Thermal Images through Adversarial Learning. In *MM '18: 2018 ACM Multimedia Conference, Oct. 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3240508.3240608

1 INTRODUCTION

Automatic facial expression recognition has attracted increasing attention in recent years due to its wide range of applications in human-computer interaction. Most research on facial expression recognition detects expressions from visible images, which provide geometric and appearance patterns but are sensitive to light conditions. Thermal images record facial temperature distribution and are not influenced by illumination changes. Therefore, successfully incorporating thermal images with visible images could result in more robust facial expression recognition in the wild [2].

Feature-level fusion [11, 19] or decision-level fusion [20] are the most intuitive approaches for integrating thermal and visible images. The former concatenates visible features and thermal features into one vector and then feeds them to a classifier for expression recognition. The latter combines the recognition results of visible and thermal images. Either approach requires visible and thermal images during both training and testing phases. However, thermal cameras are expensive and are infrequently used in real-life situations. Visible cameras are less expensive and offer higher resolution, so they are more widely used in daily life. The lack of thermal cameras prevents the adoption of feature-level fusion or decision-level fusion methods in real scenarios.

To address this, two works recently proposed to use thermal images as privileged information [14] during training to help visible images construct a better feature representation or expression classifier. Shi *et al.* [12] proposed to leverage thermal images to construct a new visible feature space through canonical correlation analysis (CCA) during training. A support vector machine (SVM) is adopted as the classifier on the constructed visible feature space. Only visible images are available during testing. Their method leverages thermal images to construct a new representation of visible images, which is expected to be more discriminative for expression recognition. Although their constructed

*Dr. Shangfei Wang is the corresponding author.

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240608

representation reflects thermal infrared images' supplementary role for visible images, it has no direct relationship to target expression labels. Wang *et al.* [17] proposed to utilize thermal images to construct a better feature representation and classifier for expression recognition from visible images. Two deep networks are used to learn representations from visible and thermal images, and then two SVMs are adopted as classifiers to recognize expressions from the learned visible representation and thermal representation. The learned representations and classifiers are jointly refined using the similarity constraint on the mapping functions from visible representation to expressions and thermal representation to expressions. Their method maximizes the benefit of thermal images to visible expression recognition through both feature representation and classification without increasing equipment cost in real-world situations. However, the learned visible representation is only regulated by target expression labels. It is independent from thermal representation. Since visible images and thermal images record the same face from different spectral bands, they share some common characteristics which represent the same facial expressions. This common representation should aid in expression recognition from one modality.

Inspired by current advances in adversarial learning [4, 6, 21], He *et al.* [5] proposed a modality classifier to construct statistically indistinguishable representations of multiple modalities through adversarial learning, and successfully applied their proposed method to unsupervised cross-modal retrieval. Unlike CCA, which can find maximally correlated representations from two modalities but cannot guarantee distribution similarity between the learned representations, He *et al.*'s method leverages adversarial learning to regularize the distribution of the learned representations and ensure their statistical indistinguishability. Since their proposed method is for unsupervised cross-modal retrieval, the learned representation has no direct relationship with target labels.

In this paper, we propose a novel facial expression recognition method enhanced by thermal images. Our method leverages thermal images to construct better visible feature representation and classifiers during training through adversarial learning and similarity constraints. Specifically, we learn two deep neural networks for expression classification from visible and thermal images. In addition to the similarity constraint on the mapping functions from visible representation to expressions and thermal representation to expressions as in [17], we propose an adversarial loss to enforce statistical similarity between the learned representations of two modalities. A discriminator is introduced to discriminate between the two modalities based on the learned representation. The two deep neural networks generate a modality-invariant representation and attempt to confuse the discriminator during training. In the training phase, the two deep neural networks and the discriminator are optimized alternately. During testing, the expression label of an unknown visible facial image is predicted by the visible neural network.

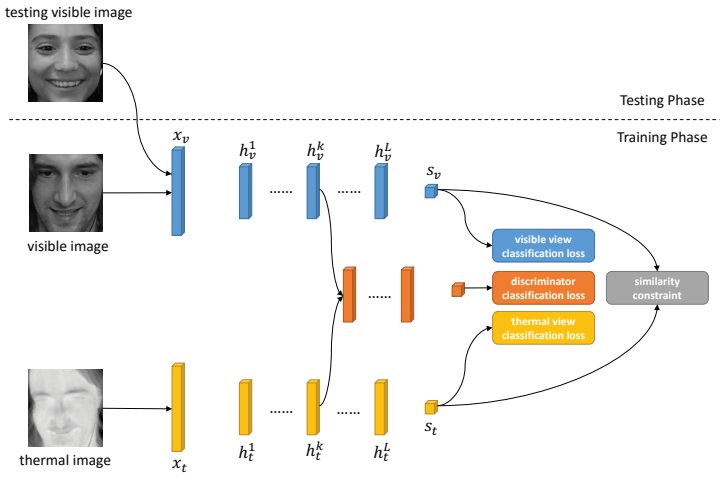


Figure 1: The framework of our proposed method for facial expression recognition.

Compared to related work, we are the first to introduce adversarial learning into thermally enhanced facial expression recognition. We propose to regularize the learned representations and classifiers jointly through adversarial learning and similarity constraints, and achieve state-of-the-art performance on expression recognition.

2 METHODOLOGY

The framework of the proposed approach consists of two deep neural networks and a discriminator, as shown in Figure 1. Two deep neural networks are used to learn feature representations and classifiers from visible and thermal images. A discriminator is introduced to distinguish the learned visible representation from the learned thermal representation. The two deep neural networks try to generate a modality-invariant representation and to confuse the discriminator. Through adversarial learning, the learned visible and thermal representations are expected to have similar distributions. Classification losses are used to make the predicted labels from visible representation and thermal representation closer to the ground truth labels. The similarity constraint enforces the similarity between the mapping functions from visible representation to expressions and thermal representation to expressions. During training, the two deep networks and the discriminator are optimized alternately. During testing, the expression label of an unknown visible facial image is predicted by the visible neural network.

2.1 Problem Statement

Let $D = \left\{x_v^{(i)}, x_t^{(i)}, y^{(i)}\right\}_{i=1}^{N}$ denote a training set containing N training samples. For the i^{th} sample, $x_v^{(i)} \in \mathbb{R}^{|V|}$ and $x_t^{(i)} \in \mathbb{R}^{|T|}$ represent the synchronous visible and thermal images respectively, where $|V|$ and $|T|$ represent the dimensions of visible and thermal images. $y^{(i)} \in \{0, 1\}$ is the ground

truth of the i^{th} sample. In our work, we only tackle the problem of expression recognition for binary classification. Given training set D, our goal is to learn an expression classifier from the visible images by using thermal images as privileged information. Thus, for a testing sample $\{x_v^{\text{test}}\}$ which only contains visible image, we can use the learned visible classifier to make a better prediction.

2.2 Proposed Method

Two duplicated neural networks, $f_v(x_v;\theta_v)$ and $f_t(x_t;\theta_t)$, are applied for visible and thermal views respectively. As shown in Figure 1, the structure of the neural networks consists of an input layer for image data, L hidden layers for non-linear transformation, and a linear output layer for prediction. We denote $s_v, s_t \in \mathbb{R}$ as the output values of the visible and thermal networks respectively.

$$\begin{aligned} s_v &= f_v(x_v;\theta_v) \\ s_t &= f_t(x_t;\theta_t) \end{aligned} \tag{1}$$

2.2.1 Supervised Classification Loss. The supervised classification loss measuring the error between the prediction and the ground truth must be defined in order to evaluate the effectiveness of our model. Consider a single training sample $\{x_v, x_t, y\}$. We obtain two predictions from visible and thermal images for a single sample, since there are two networks in our framework. We denote $\hat{y}_v$ and $\hat{y}_t$ as the probability of a sample belonging to the positive class predicted by visible and thermal networks respectively. The values of $\hat{y}_v$ and $\hat{y}_t$ can be calculated by Equation 2.

$$\begin{aligned} \hat{y}_v &= \sigma(s_v) \\ \hat{y}_t &= \sigma(s_t) \end{aligned} \tag{2}$$

where $\sigma(z) = \frac{1}{1+\exp(-z)}$ is the sigmoid function.

Let $L_c(y, \hat{y})$ be the binary cross entropy loss function as shown in Equation 3. The supervised classification losses of the visible and thermal views can be written as $L_c(y, \hat{y}_v)$ and $L_c(y, \hat{y}_t)$, respectively.

$$L_c(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y}) \tag{3}$$

2.2.2 Two-View Similarity Constraint. The idea of the similarity constraint originates with SVM2K [3]. The similarity constraint assumes that since s_v and s_t are projections of the same expression category, the information used for classification is similar. It means that the values of s_v and s_t should be similar. We introduce a similarity constraint loss as in [17], and penalize the large squared difference between s_v and s_t, defined in Equation 4.

$$L_s(s_v, s_t) = (s_v - s_t)^2 \tag{4}$$

2.2.3 Adversarial Loss. Motivated by He *et al.*'s work [5] in cross-modal retrieval, we introduce adversarial loss to regularize image representations and force a similar distribution for visible and thermal representations. For the k^{th} hidden layer representation of visible or thermal view, i.e., h^k, we introduce a discriminator $f_d(h^k;\theta_d)$ to determine the source view of the representation. The superscript k represents the id of the hidden layer in a neural network rather than the id of the training sample. We treat visible view as a positive class and thermal view as a negative class. Therefore, the adversarial loss for the k^{th} hidden layer representation is defined in Equation 5.

$$L_{d1}(h^k) = \begin{cases} L_c(1, f_d(h^k)) & \text{if } h^k \text{ is from visible view} \\ L_c(0, f_d(h^k)) & \text{if } h^k \text{ is from thermal view} \end{cases} \tag{5}$$

The discriminator minimizes $L_{d1}(h^k)$, while the visible and thermal networks try to maximize it. By applying the flip label technique, the adversarial loss is modified as Equation 6. Thus, maximizing $L_{d1}(h^k)$ is equivalent to minimizing $L_{d2}(h^k)$ while training visible and thermal networks.

$$L_{d2}(h^k) = \begin{cases} L_c(0, f_d(h^k)) & \text{if } h^k \text{ is from visible view} \\ L_c(1, f_d(h^k)) & \text{if } h^k \text{ is from thermal view} \end{cases} \tag{6}$$

2.2.4 Overall Loss Function. The two neural networks and the discriminator are optimized alternately. For a training set containing N training samples, the overall loss function when training two neural networks is defined in Equation 7.

$$\begin{aligned} L_1(\theta_v, \theta_t) &= C_1 \cdot \sum_{i=1}^{N} L_c(y^{(i)}, \hat{y}_v^{(i)}) + C_2 \cdot \sum_{i=1}^{N} L_c(y^{(i)}, \hat{y}_t^{(i)}) \\ &+ C_3 \cdot \sum_{i=1}^{N} L_s(s_v^{(i)}, s_t^{(i)}) + C_4 \cdot \sum_{i=1}^{N} \left(L_{d2}(h_v^{k(i)}) + L_{d2}(h_t^{k(i)}) \right) \\ &+ \Omega(f_v) + \Omega(f_t) \end{aligned} \tag{7}$$

This equation combines all of the aforementioned loss terms together and assigns a weight coefficient to each term. The regularization terms for the visible network $\Omega(f_v)$ and thermal network $\Omega(f_t)$ are added to reduce overfitting.

When training the discriminator, the optimization of the loss function is defined as Equation 8.

$$L_2(\theta_d) = C_d \cdot \sum_{i=1}^{N} \left(L_{d1}(h_v^{k(i)}) + L_{d1}(h_t^{k(i)}) \right) + \Omega(f_d) \tag{8}$$

where $\Omega(f_d)$ is the regularization term of the discriminator network and C_d is the weight coefficient of the classification loss of the discriminator.

Given the loss functions in Equation 7 and 8, the two neural networks and the discriminator are optimized alternately by gradient descent.

2.2.5 Inference. For an unknown visible image x_v^{test}, we feed it into the visible network and obtain the prediction $\hat{y}_v^{\text{test}}$ via Equation 1 and 2.

2.3 Extension for partially unpaired visible and thermal images

In application scenarios, visible and thermal images may not be synchronously recorded. Unpaired visible and thermal images are collected more frequently than synchronously recorded images. In order to fully exploit all of the paired and unpaired images to construct thermally enhanced representation and classifiers for visible facial expression recognition, we have extended the proposed thermally enhanced facial

expression recognition method for partially unpaired visible and thermal images in this section.

Formally, there are three datasets in the setting of partially unpaired visible and thermal images: paired visible and thermal images denoted as $D_1 = \left\{x_v^{(i)}, x_t^{(i)}, y^{(i)}\right\}_{i=1}^{N_1}$, unpaired visible images denoted as $D_2 = \left\{x_v^{(m)}, y^{(m)}\right\}_{m=1}^{N_2}$, and unpaired thermal images denoted as $D_3 = \left\{x_t^{(n)}, y^{(n)}\right\}_{n=1}^{N_3}$. We use all data to pre-train the visible and thermal neural networks. Then, the paired data is adopted to enforce the similarity between the mapping functions from visible representation to expressions and thermal representation to expressions through the similarity constraint. All data are employed to make the distribution of the learned visible representation more similar to the distribution of the thermal representation through adversarial learning.

Specifically, the extended thermally enhanced facial expression recognition method consists of two phases. In the first phase, the visible and thermal neural networks are trained with all of the available visible and thermal images, respectively. Concretely, we train the visible neural network with the labeled visible images $\left\{x_v^{(i)}, y^{(i)}\right\}_{i=1}^{N_1+N_2}$ from D_1 and D_2. Similarly, the thermal neural network is trained with the labeled thermal images $\left\{x_t^{(i)}, y^{(i)}\right\}_{i=1}^{N_1+N_3}$ from D_1 and D_3.

In the second phase, the paired dataset D_1 is used to finetune the model through similarity constraint and adversarial loss, which is similar to the original learning procedure. The remaining unpaired visible images $\left\{x_v^{(m)}\right\}_{m=1}^{N_2}$ and thermal images $\left\{x_t^{(n)}\right\}_{n=1}^{N_3}$ as well as their transformed hidden features, $\left\{h_v^{k(m)}\right\}_{m=1}^{N_2}$ and $\left\{h_t^{k(n)}\right\}_{n=1}^{N_3}$, can be still used in the adversarial learning. The loss functions of the visible and thermal networks and the discriminator are modified as Equations 9 and 10.

$$L_1(\theta_v, \theta_t) = C_1 \cdot \sum_{i=1}^{N_1} L_c(y^{(i)}, \hat{y}_v^{(i)}) + C_2 \cdot \sum_{i=1}^{N_1} L_c(y^{(i)}, \hat{y}_t^{(i)}) + C_3 \cdot \sum_{i=1}^{N_1} L_s(s_v^{(i)}, s_t^{(i)}) + C_4 \cdot \left[\sum_{i=1}^{N_1} \left(L_{d2}(h_v^{k(i)}) + L_{d2}(h_t^{k(i)}) \right) + \sum_{m=1}^{N_2} L_{d2}(h_v^{k(m)}) + \sum_{n=1}^{N_3} L_{d2}(h_t^{k(n)}) \right] + \Omega(f_v) + \Omega(f_t) \tag{9}$$

$$L_2(\theta_d) = C_d \cdot \left[\sum_{i=1}^{N_1} \left(L_{d2}(h_v^{k(i)}) + L_{d2}(h_t^{k(i)}) \right) + \sum_{m=1}^{N_2} L_{d2}(h_v^{k(m)}) + \sum_{n=1}^{N_3} L_{d2}(h_t^{k(n)}) \right] + \Omega(f_d) \tag{10}$$

Equation 9 and 10 indicate that both paired and unpaired images can be used in the adversarial learning, which ensures that modality-irrelevant representation can be learned from sufficient data.

3 EXPERIMENT

3.1 Experimental Conditions

Several expression databases provide both visible and thermal images, including the Equinox database[1], the NVIE database [16], the MMSE database [22], and the MAHNOB Laughter database [8]. The first two databases are small-scaled, and do not provide enough data to train deep networks. The third database is not publically available. Therefore, we conducted experiments on the MAHNOB Laughter database. The MAHNOB Laughter database consists of audio, visible videos, and thermal videos of spontaneous laughter from 22 subjects captured while the subjects watched funny video clips. Subjects were also asked to produce posed laughter and to speak in their native languages; this was recorded via visible and thermal cameras. Since the MAHNOB database does not provide visible and thermal images for expressions other than laughter, we cannot conduct expression category recognition on this database. In our experiment, two sub-datasets were used: the laughter versus speech dataset and the spontaneous laughter versus posed laughter dataset. Following the same experimental conditions as [17], 19 subjects from the laughter versus speech dataset were selected for discrimination, and 14 subjects from the spontaneous laughter versus posed laughter dataset were selected for discrimination.

Following the same experimental conditions as [17], we used the face detection algorithm implemented in OpenCV [15] to locate the facial area for every frame in the visible and thermal videos. Facial images were converted to grayscale and resized to 28×28 pixels. We balanced the data by adopting a down-sampling for subjects whose ratio of negative samples to positive samples was greater than two. As a result, 8252 laughter images and 12914 speech images were obtained from the laughter versus speech dataset, and 2124 spontaneous laughter images and 1437 posed laughter images were obtained from the spontaneous laughter versus posed laughter dataset.

In our experiments, both the visible and thermal image networks had the same three-layer structure. We built a discriminator with two layers on the second hidden layer representation of the visible and thermal views. All parameters of the visible and thermal networks were pre-trained by layer-wise restricted Boltzmann machines (RBMs), while the parameters of the discriminator were randomly initialized. The weight coefficient of visible classification loss C_1 was equal to the weight of thermal classification loss C_2 in order to maintain balance between visible and thermal views. Model selection was adopted to determine hyper parameters. Specifically, the number of hidden units and all of the weight coefficients were determined by grid search. A leave-one-subject-out cross-validation methodology was adopted.

[1] This database is not available now.

Accuracy and F1-score are employed as performance metrics.

We used six methods to conduct experiments on the laughter versus speech dataset and the spontaneous laughter versus posed laughter dataset. First, we trained a single visible neural network with only visible images. For the second method, a CNN consisting of five learnable layers, including two convolutional layers and three fully connected layers, is trained with only visible images. The visible CNN is randomly initialized and trained from scratch. The third method uses adversarial loss but lacks a similarity constraint. The fourth method is identical to Wang *et al.*'s method [17], and includes the similarity constraint but lacks adversarial loss. The fifth method learns visible representation in an unsupervised manner through adversarial learning with the help of thermal images first. The weight coefficients C_1, C_2 and C_3 are equal to 0 in Equation 7, and then an SVM classifier is trained from the learned visible representation. The sixth method is our proposed method.

To evaluate the extended method for partially unpaired data, we construct experiments on laughter versus speech recognition and spontaneous laughter versus posed laughter recognition using partially unpaired visible and thermal images. Following the same experimental conditions as [17], we simulated unpaired data by splitting a paired sample $\left\{x_v^{(i)}, x_t^{(i)}, y^{(i)}\right\}$ into two unpaired samples, $\left\{x_v^{(i)}, y^{(i)}\right\}$ and $\left\{x_t^{(i)}, y^{(i)}\right\}$. The ratios of the unpaired data were set to 10%, 20%, 30%, 40% and 50%. We compare our method to Wang *et al.*'s method [17], which uses similarity constraint without adversarial loss.

3.2 Experimental Results and Analysis

3.2.1 Analysis of facial expression recognition with paired data. Experimental results of laughter versus speech recognition and spontaneous laughter versus posed laughter recognition are shown in Table 1 and Table 2, respectively. From the tables, we find the following observations:

First, the proposed method outperforms all the compared methods. Specifically, the accuracy of our method on the spontaneous laughter versus posed laughter dataset is 6.26%, 14.38%, 4.69%, 2.19% and 9.55% higher than the visible neural network, visible CNN, the method without adversarial loss, the method without similarity constraint, and the method using unsupervised adversarial loss, respectively. Our method learns feature representation in a supervised manner with a supervised classification loss added in the loss function. The classification error propagates back through the backpropagation algorithm and guides two-view networks to learn suitable feature representation for expression recognition. For the method using unsupervised adversarial loss, feature representations are mainly learned in an unsupervised manner via adversarial two-view feature extractors. Without the guidance of ground truth labels, features learned by adversarial two-view feature extractors may not be specific enough for facial expression recognition.

Secondly, our method achieves better performance by using both the similarity constraint and adversarial loss. The method without adversarial loss and the method without similarity constraint each apply only the adversarial loss or the similarity constraint. The method without similarity constraint outperforms the method without adversarial loss, demonstrating that the effectiveness of adversarial loss is greater than that of the similarity constraint. Specifically, the accuracy of the method lacking the similarity constraint is 1.37% higher than the method without adversarial loss on the laughter versus speech database, and 2.50% higher on the spontaneous laughter versus posed laughter dataset. The similarity constraint is imposed on the scalar space (target label space) while the adversarial loss is imposed on the vector space (image representations space). Intuitively, the larger the space is, the greater the effectiveness of using the constraint or regularization. Additionally, adversarial loss happens in the early stages. View-irrelevant features can be learned early on, leading to better classifiers based on the learned representation. By contrast, the similarity constraint takes effect at the final stage, when there is greater freedom in the feature space of the visible and thermal views. Therefore the adversarial loss on the image representations has more of an effect on the target label than the similarity constraint, although the optimal method utilizes both.

Thirdly, the visible neural network outperforms the visible CNN on both datasets. Specifically, the accuracies of the visible neural network are 1.12% and 8.12% higher than those of the visible CNN on the laughter versus speech dataset and the spontaneous laughter versus posed laughter dataset, respectively. The visible neural network and visible CNN have roughly equal parameters. However, the visible neural network is pre-trained with layer-wise RBMs, while the CNN is initialized randomly and trained from scratch. Thus, the visible neural network achieves superior performance over the CNN.

3.2.2 Analysis of facial expression recognition with unpaired data. Experimental results of expression recognition with partially unpaired data are shown in Figure 2. From Figure 2, we can find the following observations:

First, our method outperforms the method without adversarial loss on both datasets. Only paired data are used to fine tune the method without adversarial loss, because both supervised classification loss and similarity constraints are defined for paired data. Both paired and unpaired images are used for adversarial learning in our method, so more robust features can be learned and better performance can be obtained.

Secondly, as the unpaired sample ratio increases, the performance of both methods decreases. The performance degradation is expected, since a higher unpaired sample ratio means that there are fewer paired training samples.

3.2.3 Analysis of adversarial loss. As seen in Equation 7, C_4 is a critical parameter in our model since it represents the weight of adversarial loss and controls its effect. Theoretically, if C_4 is equal to 0, there is no constraint imposed

Table 1: Experimental results on the laughter versus speech discrimination dataset.

Methods	TN	FP	FN	TP	Accuracy(%)	F1-Score
visible neural network [17]	12476	438	837	7415	93.98	0.9208
visible CNN	12329	585	927	7325	92.86	0.9064
w/o adversarial loss [17]	12490	424	816	7436	94.14	0.9230
w/o similarity constraint	12541	373	577	7675	95.51	0.9417
unsupervised adversarial loss + SVM	12576	338	1209	7043	92.69	0.9010
MMDBM + SVM [17]	11720	1194	4006	4246	75.43	0.6202
DCCA + SVM [17]	12093	821	2054	6198	86.42	0.8117
DCCAE + SVM [17]	10333	969	1573	5873	86.44	0.8221
NN [8] (audio + visible)	-	-	-	-	90.1	0.865
BLR [10] (audio + visible)	-	-	-	-	92.7	0.905
PF [9] (audio + visible)	-	-	-	-	-	0.893
Ours	12546	368	739	7513	**95.77**	**0.9451**

Table 2: Experimental results on the laughter versus posed laughter discrimination dataset.

Methods	TN	FP	FN	TP	Accuracy(%)	F1-score
visible neural network [17]	1077	360	211	1913	83.97	0.8701
visible CNN	722	715	145	1979	75.85	0.8215
w/o adversarial loss [17]	1088	349	166	1958	85.54	0.8838
w/o similarity constraint	1284	153	273	1851	88.04	0.8968
unsupervised adversarial loss + SVM	989	448	240	1884	80.68	0.8456
MMDBM + SVM [17]	254	1183	152	1972	62.51	0.7471
DCCA + SVM [17]	694	743	501	1623	65.07	0.7229
DCCAE + SVM [17]	746	690	439	1685	68.29	0.7491
Ours	1222	215	173	1951	**90.23**	**0.9177**

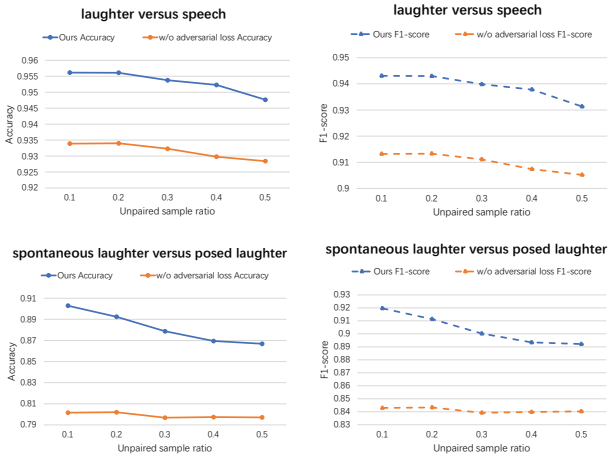


Figure 2: Experimental results with partially unpaired data.

on the visible or thermal representations. Thus, the distributions of the visible and thermal representations may be far away from each other. If C_4 is very large, the regularization of adversarial loss may be too strong to obtain good performance. We analyze the impact of varying C_4 on the classification performance as shown in Figure 3.

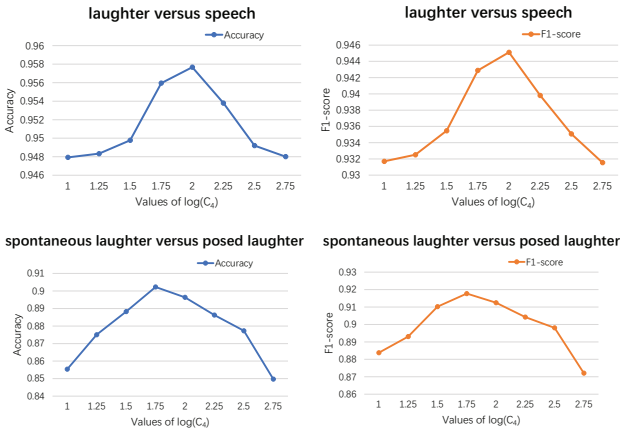
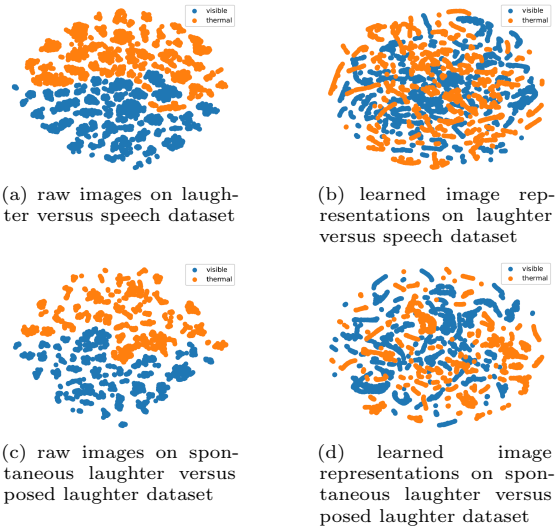


Figure 3: Impact of C_4 on performance.

From Figure 3, we find that as C_4 gradually increases, performance initially increases until it reaches a peak. When C_4 is larger than the optimal value, the performance degenerates. This observation is consistent with our theoretical analysis.

3.2.4 Adversarial feature representations visualization. To further evaluate the effect of the adversarial loss introduced in our model, we visualize the learned representation of the hidden layer with t-SNE embedding [7]. We plot the distribution of the visible and thermal images as well as the learned visible and thermal representations, as shown in Figure 4.



(a) raw images on laughter versus speech dataset

(b) learned image representations on laughter versus speech dataset

(c) raw images on spontaneous laughter versus posed laughter dataset

(d) learned image representations on spontaneous laughter versus posed laughter dataset

Figure 4: A t-SNE embedding of the distribution of the visible/thermal images and the learned visible/thermal representations.

Figure 4 shows the clear boundary between the raw visible and thermal image data. The learned visible and thermal representations are blended together. This phenomenon indicates that the distributions of the learned visible and thermal representations are nearing each other as the adversarial loss regularizes the data.

3.2.5 Comparison with related works. First, we compare our method to thermally enhanced expression recognition work. Currently, only two works address thermally enhanced expression recognition. Shi *et al.* [12] conducted experiments on the Equinox database and the NVIE database. Wang *et al.* [17] adopted the MAHNOB Laughter database. Therefore, we compare our work to Wang *et al.* [17] for both paired and unpaired samples as discussed in Sections 3.2.1 and 3.2.2. The experimental analyses in those sections demonstrate the superiority of the proposed method, since it leverages both adversarial and similarity loss, while Wang *et al.*'s method only employs similarity loss.

Since our proposed method is a multi-view learning method, we compare it with three mainstream multi-view learning methods including multimodal deep Boltzmann machine (MMDBM) [13], deep canonical correlation analysis (DCCA) [1] and deep canonically correlated autoencoders (DCCAE) [18]. The goal of multi-view learning methods is to find a common space of multi-view data with some unsupervised learning objective. The new feature representations in the common space are used to train a classifier like SVM. MMDBM, for example, learns the common space from multi-view data by inputting the learned features from each view as the visible nodes of the top-layer RBM. Through the layer-wise training of RBM, the hidden nodes of the top-layer RBM represent the common space. DCCA is a deep version of CCA, which maximizes the correlations between two learned representations. Compared to DCCA, DCCAE has an extra constraint: an autoencoder regularization term. Like most multi-view learning methods, MMDBM, DCCA, and DCCAE learn the feature representations in the common subspace in an unsupervised manner with the goal of minimizing the reconstruction error or maximizing the correlation between visible and thermal views. Our methods learn feature representations in the hidden space in a supervised and adversarial manner. This ensures that the learned features are suitable for facial expression recognition and that the distributions of visible and thermal representations are close to each other. Results of the multi-view learning method experiments are listed in Table 1. The accuracy of our method on the spontaneous laughter versus posed laughter dataset is 27.72%, 25.16%, and 21.94% higher than the accuracies achieved by MMDBM, DCCA and DCCAE, respectively.

Thirdly, we compare our method to related expression recognition works utilizing the MAHNOB Laughter database. Works using the MAHNOB Laughter database primarily distinguish laughter from speech by fusing audio and visual signals, as in [8], [10] and [9]. Since the experimental conditions of these methods are different from ours, we list them in Table 1 for reference only. In these works, facial points from visible images and Mel frequency cepstral coefficients for audio data were extracted for classification. From Table 1, we find that our method achieves the best performance. The compared methods extract feature representations from audio and visible views, which are fused to train the classifier. In our method, thermal image data is used as privileged information to construct better visible feature representation and a better classifier by utilizing the similarity constraint and adversarial learning. The improved results of our method may indicate the potential of thermal imagery for laughter versus speech discrimination.

4 CONCLUSION

In this paper, we propose a facial expression recognition method enhanced by thermal images, which are leveraged as privileged information to construct better visible feature representation and improved classifiers. Two deep neural networks are learned for expression classification from visible and thermal images. A similarity constraint is imposed to regulate the mapping functions from visible representation to expressions and thermal representation to expressions. A discriminator is built upon the visible and thermal feature representations in order to learn modality-irrelevant features. Experimental results on the MAHNOB Laughter database demonstrate that our proposed method can regularize visible

representations and expression classifiers effectively using adversarial learning and similarity constraints and the help of thermal images. This results in state-of-the-art performance on expression recognition tasks.

ACKNOWLEDGMENTS

This work has been supported by the National Science Foundation of China (Grant No. 61473270, 917418129, 61727809), and the major project from Anhui Science and Technology Agency (1804a09020038).

REFERENCES

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

[2] Vinay Bettadapura. 2012. Face expression recognition and analysis: the state of the art. *arXiv preprint arXiv:1203.6722* (2012).

[3] Jason Farquhar, David Hardoon, Hongying Meng, John S Shawe-taylor, and Sandor Szedmak. 2006. Two view learning: SVM-2K, theory and practice. In *Advances in neural information processing systems*. 355–362.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[5] Li He, Xing Xu, Huimin Lu, Yang Yang, Fumin Shen, and Heng Tao Shen. 2017. Unsupervised cross-modal retrieval through adversarial learning. In *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 1153–1158.

[6] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2016. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint* (2016).

[7] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[8] Stavros Petridis, Brais Martinez, and Maja Pantic. 2013. The MAHNOB laughter database. *Image and Vision Computing* 31, 2 (2013), 186–202.

[9] Stavros Petridis, Varun Rajgarhia, and Maja Pantic. 2015. Comparison of Single-model and Multiple-model Prediction-based Audiovisual Fusion. In *The Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing*. 457–462.

[10] Ognjen Rudovic, Stavros Petridis, and Maja Pantic. 2013. Bimodal log-linear regression for fusion of audio and visual features. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 789–792.

[11] Nandita Sharma, Abhinav Dhall, Tom Gedeon, and Roland Goecke. 2013. Modeling stress using thermal facial patterns: A spatio-temporal approach. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 387–392.

[12] Xiaoxiao Shi, Shangfei Wang, and Yachen Zhu. 2015. Expression recognition from visible images with the help of thermal images. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 563–566.

[13] Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*. 2222–2230.

[14] Vladimir Vapnik and Akshay Vashist. 2009. A new learning paradigm: Learning using privileged information. *Neural networks* 22, 5 (2009), 544–557.

[15] Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, Vol. 1. IEEE, I–I.

[16] Shangfei Wang, Zhilei Liu, Siliang Lv, Yanpeng Lv, Guobing Wu, Peng Peng, Fei Chen, and Xufa Wang. 2010. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia* 12, 7 (2010), 682–691.

[17] Shangfei Wang, Bowen Pan, Huaping Chen, and Qiang Ji. 2018. Thermal Augmented Expression Recognition. *IEEE Transactions on Cybernetics* (2018).

[18] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. 1083–1092.

[19] Avinash Wesley, Pradeep Buddharaju, Robert Pienta, and Ioannis Pavlidis. 2012. A comparative analysis of thermal and visual modalities for automated facial expression recognition. In *International Symposium on Visual Computing*. Springer, 51–60.

[20] Yasunari Yoshitomi, Sung-Ill Kim, Takako Kawano, and Tetsuro Kilazoe. 2000. Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In *Robot and Human Interactive Communication, 2000. RO-MAN 2000. Proceedings. 9th IEEE International Workshop on*. IEEE, 178–183.

[21] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE Int. Conf. Comput. Vision (ICCV)*. 5907–5915.

[22] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. 2016. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3438–3446.