Thermal Augmented Expression Recognition

Shangfei Wang, *Senior Member, IEEE*, Bowen Pan, Huaping Chen, and Qiang Ji, *Fellow, IEEE*

***Abstract*—Visible facial images provide geometric and appearance patterns of facial expressions and are sensitive to illumination changes. Thermal facial images record facial temperature distribution and are robust to light conditions. Therefore, expression recognition is enhanced by visible and thermal image fusion. In most cases, only visible images are available due to the widespread popularity of visible cameras and the high cost of thermal cameras. Thus, we propose a novel visible expression recognition method by using thermal infrared (IR) data as privileged information, which is only available during training. Specifically, we first learn a deep model for visible images and thermal images. Then we use the learned feature representations to train support vector machine (SVM) classifiers for expression classification. We jointly refine the deep models as well as the SVM classifiers for both thermal images and visible images by imposing the constraint that the outputs of the SVM classifiers from two views are similar. Thermal IR images during training are then exploited to construct better facial representations and expression classifiers from visible images. We extend the proposed thermal augmented expression recognition method for partially unpaired data, acknowledging that visible images and thermal images maybe not be recorded synchronously. Experimental resulton the MAHNOB laughter database demonstrate that the proposed thermal augmented expression recognition method can effectively exploit thermal IR images' supplementary role for visible facial expression recognition during training to obtain better facial representations and a better visible expression classifier. The proposed thermal augmented expression recognition method achieves state-of-the-art expression recognition performance for both paired and unpaired facial images.**

***Index Terms*—Deep models, facial expression recognition, privileged information, support vector machine, thermal images, visible images.**

I. Introduction

FACIAL expression recognition has gained attention in recent years due to its wide application in human–computer interaction. Considerable progress has been achieved in the field of facial expression using visible images and videos [1]. Although visible images contain geometric and appearance patterns of faces that are crucial for facial expression recognition, they are not attuned to illumination changes. Thermal infrared (IR) images, which record temperature distribution, are not sensitive to illumination conditions. Several researchers [1]–[5] have utilized expression recognition from thermal images.

Although IR images are receptive to illumination changes, they are specifically sensitive to the surrounding temperature and opaque to glass [6]. Additionally, they do not provide clear geometric and appearance patterns of faces as visible images do. Geometric patterns, appearance patterns, and temperature distribution are all crucial for facial expression recognition. Therefore, it is beneficial to combine visible images and IR images for facial expression recognition. This method is currently underutilized [7], and proves the potential of expression recognition by fusing visible and thermal images. However, the use of thermal imaging for facial expression recognition in real-world situations is questionable. In most cases, only visible images are available due to the prevalence and low cost of visible cameras compared to thermal IR cameras. Typically only visible images are available during testing. Images of two modalities are available during training. We propose using thermal images as privileged information, only needed during training, to construct a better expression classifier from visible images and better facial representations that can be constructed using visible images alone. The thermal images are not required during testing. Thus, our approach can be used in real-world applications.

Vapnik and Vashist [8] recently introduced the new learning model learning using privileged information, and proposed a discriminative model called support vector machine (SVM)+ [8]. The basic idea of SVM+ is to use privileged information to predict the slack variables in the SVM. Their experimental results of digit recognition [8] demonstrate that SVM+ classifiers trained on low-resolution digit images with the help of privileged information reach similar accuracies compared with results trained on high-resolution images with SVM. SVM+ assumes that privileged information and regular features share the same slacking variable. This assumption is potentially restrictive and unrealistic for some applications. Instead of using SVM+ directly, we propose a novel method, referred as deep two-view support vector machine, to classify expressions from visible images using thermal images with the similarity constraint. First, we adopt two deep networks: one to learn visible facial representations from visible images, and one to learn thermal facial representations from thermal images. Then we train two classifiers to recognize expressions from visible facial representation and thermal facial representation. During classifier training, we simultaneously refine the deep models and classifiers for both thermal

Manuscript received May 30, 2017; revised October 1, 2017; accepted December 18, 2017. Date of publication January 11, 2018; date of current version June 14, 2018. This work was supported in part by the National Science Foundation of China under Grant 61473270, Grant 91748129, Grant 61175037, and Grant 61228304, and in part by the Project from Anhui Science and Technology Agency under Grant 1508085SMF223. This paper was recommended by Associate Editor B. W. Schuller. *(Corresponding author: Shangfei Wang.)*

S. Wang, B. Pan, and H. Chen are with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China (e-mail: sfwang@ustc.edu.cn; bowenpan@mail.ustc.edu.cn; hpchen@ustc.edu.cn).

Q. Ji is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: qji@ecse.rpi.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2017.2786309

images and visible images, with the constraint that the outputs of the classifiers from visible images and thermal images are similar. We extend the proposed thermal-augmented visible expression recognition method for partially unpaired data. Instead of discarding unpaired data, we keep all available visible and thermal images to train the feature representations and classifiers. We then use the paired data to further tune the feature representations and classifiers through a similarity constraint. We successfully exploit all the paired and unpaired images for thermal-augmented visible expression classification. Experiments on benchmark databases demonstrate the effectiveness and superior performance of our proposed method for learning the visible facial representations and expression classifier with thermal images as privileged information for both paired images and partially unpaired images.

Compared with related work, our contributions are as follows.

First, we propose a novel visible expression recognition method by using thermal IR data as privileged information. The thermal IR data, required only during training, are explored to learn better visible facial representations and the expression classifier jointly.

Second, we extend the proposed thermal-augmented visible expression recognition method for partially unpaired data. Instead of discarding unpaired data, we successfully exploit all the paired and unpaired images through keeping all available visible and thermal images to learn representations and classifiers.

II. Related Work

Current related work recognizes expressions from visible facial images and videos. A comprehensive review of expression recognition from visible images can be found in [9]–[11]. Herein is a brief review of acquiring expression recognition from thermal images by integrating thermal IR data with visible spectrum images. We also survey a related learning model called multiview learning. Multiview learning requires data from multiple sources during training and testing, while the proposed thermal-augmented expression recognition method requires thermal and visible images during training, but only visible images during testing.

A. Expression Recognition From Thermal Images

Facial expression recognition in the thermal spectrum has received relatively little attention compared to facial expression recognition in the visible spectrum [11]. Depending on the wavelength, thermal IR images can be divided into short wave IR images (950–1700 nm), medium wave IR images (3000–5000 nm), and long wave IR images (LWIR, 8000–14000 nm). In this analysis, we only consider expression recognition from long wave IR images due to the vast heat energy emitted in LWIR sub-bands.

Current research of thermal facial expression recognition primarily encompasses two steps: 1) feature extraction and 2) classification. In feature extraction, most research proposes using temperature features embedded in thermal images. Wang *et al.* [7] proposed several statistical parameters, including mean, standard deviation, minimum, and maximum, as thermal features. Similarly, Pérez-Rosas *et al.* [12] used several statistical parameters, such as average temperature, overall minimum and maximum temperature, average of the minimum and maximum temperature observed in each frame, standard deviation, standard deviation for the minimum and maximum temperature observed in each frame, and the difference between the minimum and the maximum temperature. Liu and Yin [13] proposed a thermal video descriptor to capture the information of both skin temperatures and head motions using a bag-of-video-word model. Nguyen *et al.* [14] adopted the temperature data of the whole face as the features. Jarlier *et al.* [3] calculated the representative temperature maps by subtracting each temperature value of the images. This was obtained at the apex from the spatially corresponding temperature image obtained at the end of the baseline. Wesley *et al.* [6] first detected regions of interest (ROIs) at 13 fiducial regions on the critical faces evaluating all action units (AUs), and then applied principal component analysis (PCA) to extract the deviation from neutral expression at each of the corresponding ROIs. Nhan and Chau [15] defined time, frequency, and time-frequency features from the mean temperature time series of five ROIs in the periorbital, supraorbital and nasal regions. Khan *et al.* [16] extracted the facial thermal features using the PCA from thermal intensity values. Sharma *et al.* [17] proposed dynamic thermal patterns in histograms to capture thermal and spatio-temporal characteristics associated in thermal videos. Wesley *et al.* [6] defined the representative temperature maps by subtracting each temperature value of the image obtained at the apex from the spatially corresponding temperature image obtained at the end of the baseline.

In addition to defining specified thermal features, some research adopts features that have been proposed for visible expression recognition. For example, Sharma *et al.* [17] extracted local binary patterns on three orthogonal planes from thermal videos. Yoshitomi *et al.* [4] used discrete cosine transformation to extract thermal features. Poursaberi *et al.* [18] adopted a Gauss–Laguerre filter of circular harmonic wavelets to extract features from thermal images. Hernandez *et al.* [19] extracted texture features using the gray level co-occurrence matrix. Other than using temperature features or hand-crafted features designed for visible images, He *et al.* [5] proposed to learn thermal features using the deep Boltzmann machine.

After feature extraction, classifiers should be trained. Nguyen *et al.* [14] used the PCA eigen-space method based on class-features (EMC), and the PCA-EMC method to classify five expressions (i.e., neutral, fear, anger, happiness, and sadness) from facial temperatures. Jarlier *et al.* [3] adopted the K-nearest neighbor model to classify nine AUs or AU combinations and to differentiate their speed and strength of contraction. Wesley *et al.* [6] employed a feed-forward multilayer perceptron to recognize eight AUs and AU combinations. Nhan and Chau [15] distinguished baseline and affect states using a Fisher linear discriminant analysis classifier trained with feature subsets selected with a standard genetic algorithm. Khan *et al.* [16] proposed to classify pretended and evoked facial expressions using linear discriminants. Liu and Yin [13]

and Hernandez *et al.* [19] adopted a support vector machine for expression recognition. Pérez-Rosas *et al.* [12] classified positive, negative, and neutral affective states using the AdaBoost classifier with decision stumps. All of these studies demonstrate the feasibility and effectiveness of thermal images for expression recognition.

B. Expression Recognition by Fusing Thermal Images and Visible Images

Our research indicates that facial expression recognition using visible and IR images are currently underutilized in the field. Yoshitomi *et al.* [2] proposed to classify five expressions (i.e., anger, happiness, neutral, sadness, and surprise) through decision-level fusion of voices, visual, and IR facial images. Wang *et al.* [7] proposed to fuse visible features and thermal features for facial expression recognition from the decision and feature levels using Bayesian networks and SVMs. Sharma *et al.* [17] proposed detecting stress by feature-level fusion of visible and thermal videos using SVM. Wesley *et al.* [6] provided a comparative analysis of facial AU recognition from thermal videos, visual videos, and their fusion. The aforementioned works demonstrate the promise of fusing visible and thermal images to recognize expression and emotion.

Currently, fusing thermal and visible images to facilitate expression recognition requires two modalities during training and testing. Although visible cameras are widely used, thermal cameras are typically only available in laboratory situations due to their high prices. Therefore, the use of thermal imaging for facial expression recognition in real-world situations is questionable. Shi *et al.* [20] recently proposed a method of expression recognition from visible images with the help of thermal images, which are only available during training, but not available during testing. A new visible feature space is constructed using canonical correlation analysis (CCA) with the help of thermal images. An SVM is adopted as the classifier on the constructed visible feature space. Their proposed method integrates thermal images and visible images to learn a new representation of visible images during training. The new representation is expected to be more discriminative for expression recognition. Thermal images are only required during training to construct better visible image representation. During testing, only visible images are available. Their proposed method uses thermal images as privileged information. This maximizes the impact of thermal images and recognizes expression in real-world situations without increasing equipment cost.

Although the constructed representation reflects thermal IR images supplementary role for visible images, it has no direct relationship to target expression labels. To address this, we propose a new thermal augmented expression recognition method. This recognition method utilizes thermal images to construct a better representation and better classifier for expression recognition from visible images. This is accomplished by jointly refining the learned representations and classifiers using the similarity constraint on the mapping functions from visible representation to expressions and thermal representation to expressions. Deep networks are used to learn representations and SVMs are used as classifiers. Furthermore, we extend the proposed expression recognition method for partially unpaired images due to the evidence that thermal images and visible images may be not recorded synchronously in some cases.

C. Multiview Learning

The proposed thermal enhanced expression recognition method jointly optimizes image representations and classifiers from visible images and thermal images. This method models the inherent dependencies between thermal images and visible images for expression recognition performance improvement during training. During testing, the proposed thermal enhanced expression recognition method classifies expressions from visible images only. This method is related to multiview learning, which requires data from multiple sources during both training and testing. Comprehensive surveys on multiview learning can be found in [21] and [22].

One mainstream multiview learning algorithm is to locate common spaces of multiviews by synthesizing every views relevant information for classification. This is also referred to as multiview dimensionality reduction. CCA is a popular subset method that aims to find two bases, one for each view, that are optimal with respect to maximum correlations [23]. CCA is a way of measuring the linear relationship between two views. Kernel CCA (KCCA) is proposed after to extend CCA from the linear change to the nonlinear change by leveraging kernel method during CCA transformation [24]. Due to the recent emergence of big data and success of deep learning, several deep neural network (DNN)-based multiview representation learning algorithms are proposed. The training criteria of DNN-based multiview representation learning can be classified into two categories: learn representations in two views that are maximally correlated, and learn a compact representation that best reconstructs the inputs. The first category, learn representations in two views that are maximally correlated, involves deep CCA (DCCA) [25], which is a DNN version of CCA [26]. DCCA aims to learn feature representations of two views which are maximally correlated by applying DNNs as a nonlinear function to model real world data accurately through extracting high level representations [27]. DCCA has been proved more accurate than KCCA in nonlinear transformation tasks [26]. The other category of DNN-based multiview representation learning is to learn a compact representation that best reconstructs the inputs, such as multimodal deep belief network [28] and multimodal deep Boltzmann machine (MMDBM) [29].

Another platform of multiview learning is multiview supervised learning [21]. Unlike multiview dimensionality reduction, which leans subspace jointly from multiple views and ignores target labels, multiview supervised learning algorithms train classifiers from multiple views simultaneously. SVM2K is a typical method that globally optimizes two distinct SVMs, one in each of the two views, by using slack variables to measure the amount incongruent points between two predictions.

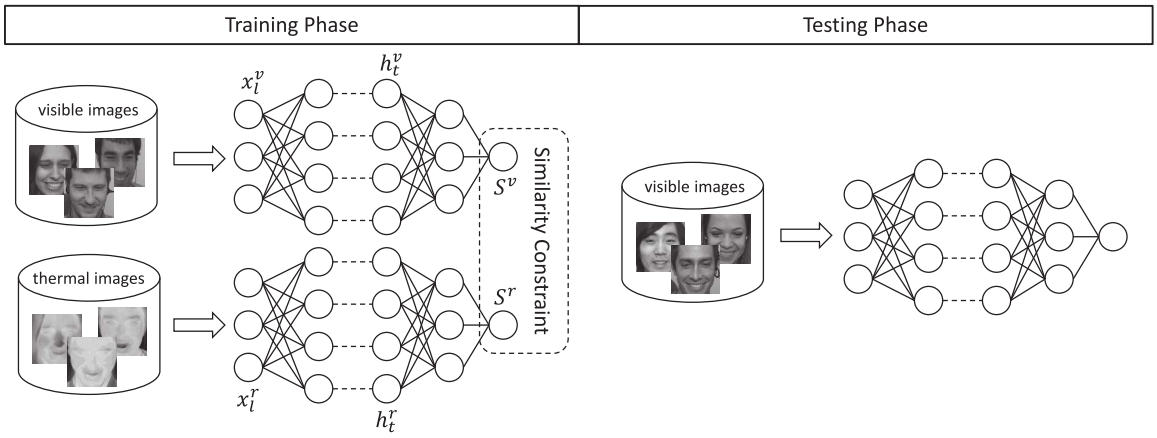


Fig. 1. Framework of our proposed approach.

Instead of directly adopting current multiview learning methods by using two view data during training and one view data during testing, we propose a new thermal enhanced expression recognition method from visible images. We learn two DNNs for visible images and thermal images, respectively, and use the learned feature representation to train SVMs for expression classification. DNNs and SVMs are then refined by constraining the similar outputs of the SVMs. This proposed method of exploring thermal images as privileged information during training learns better representations and classifiers for expression recognition from visible images.

III. Thermal-Augmented Expression Recognition

The framework of our approach is summarized in Fig. 1. Features are extracted from both thermal images and visible images by DNNs. During training, the expression classifier from visible images and the expression classifier from thermal images are jointly trained through the similarity constraint. During testing, the expression label of a visible facial image is predicted using only the DNN and the SVM classifier from visible images.

A. Problem Statement

We developed a method for expression recognition by using two views of data, i.e., visible images and thermal images, to represent the same category of expressions. Denote $v \in R^{|V|}$ and $t \in R^{|T|}$ as visible and thermal image data, respectively, where $|V|$ and $|T|$ represent dimension of visible and thermal image data. We assume two different categories of expressions, labeled as $y \in \{-1, 1\}$. The training data consists of visible image, thermal image and the corresponding facial expression labels, denoted as $D = \{v_l, r_l, y_l\}_{l=1}^{L}$, where L is the number of samples in training set. Our goal is training a predictive classifier by using thermal IR data as privileged information for the training set D, and making a likely prediction for an unknown testing sample $x = \{v\}$, which only contains visible view data.

B. General Solution

Within our two view expression recognition method, two duplicated network structures are applied for visible and thermal view, respectively. There are mainly two components for the network structure in each data view. The first component is a DNN used for feature extraction. The second component is a simple classifier used for making predictions. In order to combine the two separated structures together, a similarity constraint is imposed on the outputs of the two classifiers. This dictates that their outputs are similar values. We will discuss feature extraction, classifiers, and similarity constraint in the following sections. We will then introduce the general objective function of the proposed thermal enhanced visible expression recognition.

1) Feature Extraction: The DNN consists of multiple layers of restricted Boltzmann machine (RBM). We adopt two deep networks for nonlinearly transforming two views of data: 1) visible image data and 2) thermal image data. As shown in Fig. 1, the output of deep networks can be computed layer by layer through a series of nonlinear functions. In the deep network for visible image data, $t \in \{1, 2, \ldots, T\}$ represents the index of layers of the network, n_t^v represents the number of nodes at tth layer. For an input sample $v_l \in R^{|V|}$ $(l = 1, \ldots, L)$, n^v represents the number of input nodes. Output of the first layer of the network can be computed as $h_1^{v,l} = s(\langle w_1^v, v_l \rangle + b_1^v) \in R^{n_1^v}$, where $\{w_1^v, b_1^v\}$ are parameters of the first layer of deep network for the first view of data, $s : R^{|V|} \to R^{n_1^A}$ is the nonlinear function of the transformation. h_1^v is the input of the second layer of the network. The output of the second layer can be computed similarly. The final layer $h_T^{v,l}$ is obtained as: $h_T^{v,l} = s(<w_T^v, h_{T-1}^{v,l}> + b_T^v) \in n_T^v$, where $h_T^{v,l}$ is the learned representation of visible image data and can be computed as the final projection. The deep network for thermal image data is similarly calculated by replacing v into r, and $h_T^{r,l} = f^r(x_l^r)$ can be computed as the final projection. The outputs of the top layers, h_T^v and h_T^r, can be viewed as the abstract feature representation because they are obtained by nonlinear transformation through layers.

2) Classification: A simple classifier is applied to the deep network. This simple classifier takes extracted features, i.e., the top layer of DNN, as its input and outputs the final

prediction. The proposed method can use an arbitrary classifier, which generally first projects input into a real value, $f_x : R^{|X|} \rightarrow R$, and then outputs the prediction by a decision function, $d_x : R \rightarrow \{-1, 1\}$. Denote $\text{loss}_v(y, \hat{y})$ and $\text{loss}_r(y, \hat{y})$ as the loss function of the two classifiers of visible and thermal views, respectively. The function of the two classifiers projects feature representation h_T^v and h_T^r into real values $f_v(h_T^v)$ and $f_r(h_T^r)$.

3) Two Views Similarity Constraint: At this point we find the prediction of the two views of data $f_v(h_T^v)$ and $f_r(h_T^r)$. We assume their classification information are similar since they are projections of the same expression category. Motivated by Farquhar *et al.*'s [30] work, we introduce a similar constraint of $f_v(h_T^v)$ and $f_r(h_T^r)$ to refine the performance of the related classifiers. The similarity constraint of $f_v(h_T^v)$ and $f_r(h_T^r)$ can be expressed as

$$\left|f_v\left(h_T^v\right) - f_r\left(h_T^r\right)\right| \leq \eta_l + \epsilon \tag{1}$$

where, η_l is the slack variable to measure the amount that lth sample fails to meet ϵ similarity between projection $f_v(h_T^v)$ and $f_r(h_T^r)$.

4) General Objective Function: To evaluate model effectiveness, three aspects must be considered at the same time: 1) the loss of visible classifier; 2) the loss of thermal classifier; and 3) the similarity constraint. A successful model must meet these three conditions. First, the model must be able to make an accurate prediction for the visible view because it is the main task of expression recognition. Second, the model must be able to make an accurate prediction for the thermal view. Thermal images are used as privileged information. A good prediction from thermal images can help to improve the performance of visible view classification. Third, the similarity constraint must be imposed on the outputs of the two views. These two networks are interactive during the training phase. Combined together, the objective function of the whole structure is defined as

$$\begin{aligned} F(\Theta_v, \Theta_r) = {} & C^v \sum_{l=1}^{L} \text{loss}_v\left(y_l, f_v\left(h_T^v\right)\right) + C^r \sum_{l=1}^{L} \text{loss}_r\left(y_l, f_r\left(h_T^r\right)\right) \\ & + D \sum_{l=1}^{L} \left(f_v\left(h_T^v\right) - f_r\left(h_T^r\right)\right)^2 \end{aligned} \tag{2}$$

where Θ_v and Θ_r are parameters of the two network structures, and C^v, C^r, D control the weights of visible classifier's loss, thermal classifier's loss, and similarity constraint.

C. Deep Two-View Support Vector Machine

We adopt the maximum margin classifier SVM upon the DNN, which is similar to the framework in [31] and use deep network for representation learning for each view of data. By introducing a similarity constraint, the proposed method optimizes image representations and SVM classifiers for both views of data simultaneously. The proposed method is referred to as deep two-view SVM. We add a layer at the top of each deep network containing one node, denoted as S^v and S^r, shown in Fig. 1. The value of S^v and S^r is linear combination of the output of deep networks

$$S^v = \left\langle w_S^v, h_T^v \right\rangle + b_S^v; \quad S^r = \left\langle w_S^r, h_T^r \right\rangle + b_S^r \tag{3}$$

where w_S^v, b_S^v, w_S^r, and b_S^r denotes weights and biases of the added layer of deep networks. These are also parameters of the SVM classifier for both views of data. We define the constraint of similarity between S^v and S^r as

$$\left|S^v - S^r\right| \leq \eta_l \tag{4}$$

where η_l is slack variable that measures the amount of dissimilarity of sample l from two views.

Combining the similarity constraint defined above and SVM constraint, the optimization of our model is expressed as

$$\begin{aligned} \min_{\Theta_v, \Theta_r} \quad & L = \frac{1}{2}\left\|w_S^v\right\|_2^2 + \frac{1}{2}\left\|w_S^r\right\|_2^2 \\ & \quad + C^v \sum_{l=1}^{L} \left(\xi_l^v\right)^2 + C^r \sum_{l=1}^{L} \left(\xi_l^r\right)^2 + D \sum_{l=1}^{L} (\eta_l)^2 \\ \text{s.t.} \quad & \left|S^v - S^r\right| \leq \eta_l, \quad y_l \cdot S^v \geq 1 - \xi_l^v \\ & y_l \cdot S^r \geq 1 - \xi_l^r, \quad \xi_l^v \geq 0, \ \xi_l^r \geq 0, \ \eta_l \geq 0 \\ & \text{all for } 1 \leq l \leq L \end{aligned} \tag{5}$$

where $\Theta_v = \{\theta^v, w_S^v, b_S^v\}$ and $\Theta_r = \{\theta^r, w_S^r, b_S^r\}$ are parameters of the extended deep networks, ξ_l^v, ξ_l^r are slack variable of common SVMs, and η_l is the slack for similarity constraint. According to [32], using squared slacks is slightly better, thus in (5) we use $(\xi_l^v)^2$, $(\xi_l^r)^2$, and $(\eta_l)^2$ instead. By substituting the slack variables from the constraint into the objective function, (5) can be rewritten as the minimization of

$$\begin{aligned} F(\Theta_v, \Theta_r) = {} & \frac{1}{2}\left\|w_S^v\right\|_2^2 + \frac{1}{2}\left\|w_S^r\right\|_2^2 + C^v \sum_{l=1}^{L} \left[1 - y_l \cdot S^v\right]_+^2 \\ & + C^r \sum_{l=1}^{L} \left[1 - y_l \cdot S^r\right]_+^2 + D \sum_{l=1}^{L} \left(S^v - S^r\right)^2 \end{aligned} \tag{6}$$

where $[\cdot]_+$ is the hinge function.

Back propagation is used for optimization. The key is to compute the gradient of objective function, (6), with respect to all weight and bias parameters at each layer of deep networks, including the added layer. For parameters corresponding to visible view, the gradient of objective function (6) to the weight and bias at the top layer (SVM parameters) can be computed as

$$\begin{aligned} \frac{\partial F}{\partial w_S^v} &= \frac{\partial F}{\partial S^v} \cdot \frac{\partial S^v}{\partial w_S^v} = w_S^v + \delta_S^v \cdot h_T^v \\ \frac{\partial F}{\partial b_S^v} &= \frac{\partial F}{\partial S^v} \cdot \frac{\partial S^v}{\partial b_S^v} = \delta_S^v \cdot \mathbf{1} \end{aligned} \tag{7}$$

where δ_S^v is the backward error at the top layers that will be passed to previous layer, and $\mathbf{1}$ is an all-encompassing vector

$$\delta_S^v = \frac{\partial F}{\partial S^v} = -2C^v \cdot Y \times \left[1 - Y \times S^v\right]_+ + 2D \cdot \left(S^v - S^r\right). \tag{8}$$

For other layers, the gradients can be computed with a normal back propagation algorithm

$$\frac{\partial F}{\partial w_t^v} = \sum_{l=1}^{L} \frac{\partial F}{\partial h_t^{v,l}} \cdot \frac{\partial h_t^{v,l}}{\partial w_t^v} \tag{9}$$

all for $t = 1, \ldots, T$. For parameters corresponding to thermal image data, all derivatives are similar.

All deep network parameters are initialized through a "pretraining" process using contrastive divergence algorithm, according to [33]. Weights of SVM are randomly initialized. During testing, the decision function of our model is $d(x) = \text{sign}(f(x))$, where $f(x)$ can be estimated using both views or any one of them. The proposed method is formulated for binary classification, and can be extended for multiclass problems using the one-vs-rest approach.

D. Extension to Partially Unpaired Visible and Thermal Data

Visible images and thermal images may be not always collected synchronously. We extend the proposed method for partially unpaired visible and thermal images to account for this. Suppose we have a dataset including paired samples and unpaired samples. There are three portions in the dataset, which are paired samples denoted as $D_1 = \{v_l^{(1)}, r_l^{(1)}, y_l^{(1)}\}_{l=1}^{L_1}$, visible unpaired samples denoted as $D_2 = \{v_l^{(2)}, y_l^{(2)}\}_{l=1}^{L_2}$ and thermal unpaired samples denoted as $D_3 = \{r_l^{(3)}, y_l^{(3)}\}_{l=1}^{L_3}$. Both paired and unpaired samples are used for extracting features by DNN during pretraining phase. During back propagation phase, only paired samples are used for fine-tuning. Thus, we have an objective function shown as

$$\begin{aligned} F(\Theta_v, \Theta_r) = {} & \frac{1}{2}\left\|w_S^v\right\|_2^2 + \frac{1}{2}\left\|w_S^r\right\|_2^2 + C^v \sum_{l=1}^{L}\left[1 - y_l^{(1)} \cdot S^{(1)v}\right]_+^2 \\ & + C^r \sum_{l=1}^{L}\left[1 - y_l^{(1)} \cdot S^{(1)r}\right]_+^2 + D\sum_{l=1}^{L}\left(S^{(1)v} - S^{(1)r}\right)^2. \end{aligned} \tag{10}$$

Compared to (6), samples involved in (10) are only paired ones. The superscript (1) of a variable means that it is related to paired samples $D_1 = \{v_l^{(1)}, r_l^{(1)}, y_l^{(1)}\}_{l=1}^{L_1}$.

In the back propagation algorithm, the backward error at the top layers is slightly modified from (8)

$$\begin{aligned} \delta_S^v = \frac{\partial F}{\partial S^{(1)v}} = {} & -2C^v \cdot Y^{(1)} \times \left[1 - Y^{(1)} \times S^{(1)v}\right]_+ \\ & + 2D \cdot \left(S^{(1)v} - S^{(1)r}\right). \end{aligned} \tag{11}$$

After the model is fine-tuned, only visible data are used for computing the performance during testing phase. Although there are fewer available paired samples for fine-tuning, a large number of unpaired samples for pretraining remain. This guarantees appropriate features can be learned by DNN and achieve good performance.

The aforementioned discussion details how expression recognition with paired data can be regarded as a special case of expression recognition with partially unpaired data. When the unpaired sample sets are empty, the former is equal to the latter. The algorithm of our proposed model for partially unpaired visible and thermal data are outlined in Algorithm 1.

Algorithm 1 Expression Recognition for Partially Unpaired Visible and Thermal Data

Input: paired training samples $D_1 = \{v_l^{(1)}, r_l^{(1)}, y_l^{(1)}\}_{l=1}^{L_1}$
unpaired visible training samples $D_2 = \{v_l^{(2)}, y_l^{(2)}\}_{l=1}^{L_2}$
unpaired thermal training samples $D_3 = \{r_l^{(3)}, y_l^{(3)}\}_{l=1}^{L_3}$
visible testing sample $x = \{v\}$

Output: Optimal networks f_v and f_r
Prediction of the testing sample $d(x)$

pre-training phase

1. Pre-train the DNN of visible view layer by layer with data D_1 and D_2.
2. Pre-train the DNN of thermal view layer by layer with data D_1 and D_3.

back propagation phase

1. Minimise the objective function Eq.10 with pre-trained parameters Θ_v, Θ_r and data D_1 by gradient descent.
2. Output networks f_v and f_r with optimal parameters Θ_v^* and Θ_r^*.

testing phase

Output the prediction of the visible testing sample x by decision function $d(x) = sign(f_v(x))$.

IV. Experiments

A. Experimental Condition

One known work utilizes thermal image data as privileged information to help facial expression recognition from visible images, mentioned in Section I. Shi *et al.* [20] proposed using CCA to construct visible features with the help of thermal images, and adopting SVM as a classifier to recognize expressions from visible images. Their experiments were conducted on the NVIE database and the Equinox database. The number of image samples adopted in their experiments is insufficient for deep model training. Instead of using identical data, we conducted experiments of laughter versus speech and laughter versus posed laughter discrimination from images on the MAHNOB laughter database [34] to validate the effectiveness of the proposed method.

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 additionally asked to produce posed laughter and to speak in their native language, which were recorded via visible and thermal videos. Since the MAHNOB database does not provide visible and thermal images for other expressions except for laughter, we cannot conduct on expression recognition on this database. In our experiment, two subdatasets were used: 1) the laughter versus speech dataset and 2) the laughter versus posed laughter dataset. The lack of thermal videos of three subjects in laughter versus speech dataset and eight subjects in the laughter versus posed laughter dataset led us to select paired visible and thermal videos. Nineteen and 14 subjects from the laughter versus speech dataset and the laughter

versus posed laughter dataset were selected for the laughter versus speech discrimination and the laughter versus posed laughter discrimination, respectively. Image samples were collected by selecting synchronized visible and thermal frames. We then located the facial area for every frame in the visible and thermal videos using the face detection algorithm implemented in OpenCV [35]. Facial images were converted to gray scale and resized to 28×28. We adopted a down-sampling for subjects whose ratio of negative samples to positive samples was greater than two to balance the data. As a result, 8252 positive class laughter images and 12 914 negative class speech images were obtained from the laughter versus speech dataset. 2124 positive class laughter images and 1437 negative class posed-laughter images were obtained from the laughter versus posed laughter dataset. We conducted the laughter versus speech discrimination experiment and the spontaneous laughter versus posed laughter discrimination experiment.

In the experiments, we set the network structure of visible images as the same as that of thermal images, and used two hidden layers. We kept the weight of visible classifiers loss C^v equal to the weight of thermal classifiers loss C^r. The model selection was adopted to determine hyper parameters. Specifically, the number of hidden units, the weight of visible classifiers loss C^v, the weight of thermal classifiers loss C^r and the weight of similarity constraint D were all determined by grid search. A leave-one-subject-out cross-validation methodology was adopted. Accuracy and F1-score were employed as performance metrics.

We evaluated the proposed method for totally paired data by comparing the proposed thermal augmented expression recognition method with two methods that recognize visible images only. One uses SVM to directly recognize expressions from visible images. The other adopts DNN to learn visible image representation with SVM as the classifier. We compared the proposed method with several state-of-the-art multiview learning algorithms, including SVM2K, MMDBM, DCCA, and DCCAE. For MMDBM, DCCA, and DCCAE, with SVM used as the classifier. For the four multiview learning algorithms, the codes provided by Srivastava and Salakhutdinov [29], Farquhar *et al.* [30], and Wang *et al.* [25] were adopted. The four multiview learning algorithms are notably used as learning with privileged information. This requires both thermal images and visible images during training. Only visible images are used during testing.

For experiments with partially unpaired data, both paired and unpaired images were used for pretraining, while only paired images were used for fine-tuning. All of the selected samples from the MAHNOB laughter database were paired. We simulated unpaired data by splitting a paired sample $\{v_i, r_i, y_i\}$ into two unpaired samples, i.e., $\{v_i, y_i\}$ and $\{r_i, y_i\}$. The split samples were exclusively used for pretraining. The ratio of the unpaired data was set to 10%, 20%, 30%, 40%, and 50%, respectively. Each experiment with a certain unpaired sample ratio was repeated ten times. The mean, variance of the accuracy, and F1-score were employed as performance metrics.

We evaluated the proposed thermal augmented expression recognition from visible images for partially unpaired data by comparing the proposed method with the method that discards all unpaired data and uses our identical network structure. We compared the proposed method with the method that recognizes expressions from visible images using only DNN and SVM. In the single view recognition experiments all the visible images were used to pretrain the DNN model, but only the visible images with corresponding thermal images were used to fine tune the SVM classifier and DNN.

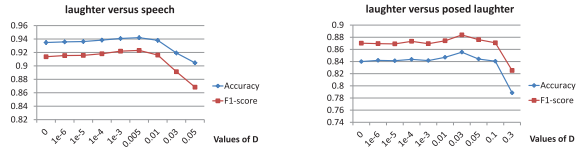


Fig. 2. Impact of D on performance.

B. Experimental Results and Analysis for Paired Data

1) Performance Versus Similarity Constraint: As shown in (6), D is a critical parameter, since it represents the weight of similarity constraint, and control the interaction between two networks during training. Theoretically, if D is equal to 0, the network of visible images and that of thermal images do not interact with each other. Thus, thermal network cannot help visible network for classification. If D is very larger, the similar constrain may be too strong to obtain good performance. We further experimentally analyze the impact of D to classification performance as shown in Fig. 2.

From Fig. 2, we find that when D gradually increases, the performance goes up at an early stage. After D is larger than the optimal value, the performance degenerates quickly. This observation is consistent with the theoretical analysis.

2) Analyses of Expression Recognition: Experimental results of the laughter versus speech recognition and the spontaneous laughter versus posed laughter recognition from paired data are shown in Tables I and II, respectively.

We obtain the following observations from Tables I and II.

The proposed method performs more accurately and has a higher F1-score compared to the two methods that recognize expressions from visible images only. Specifically, the accuracy of our proposed method in the laughter versus posed laughter discrimination is 17.50% and 1.57% higher than the SVM and single view structure with only visible data, respectively. SVM trains a simple classifier with raw pixel values of visible images as its input. In single view using DNN with SVM, features are extracted by DNN in the pretraining phase instead of using raw pixel values directly. Conversely the proposed method extracts features from both visible and thermal views by DNNs and fine-tunes features by training two networks simultaneously. Thermal data are viewed as privileged information and assist the visible network to construct better features through a similarity constraint between outputs of two SVMs. The proposed method achieves better performance than two methods with visible images only. This demonstrates that the proposed thermal augmented expression recognition method can effectively exploit thermal IR images supplementary role for visible facial expression recognition

TABLE I
RESULTS OF THE LAUGHTER VERSUS SPEECH DISCRIMINATION (WITH LAUGHTER AS A POSITIVE CLASS AND SPEECH AS A NEGATIVE CLASS)

Experiment	TN	FP	FN	TP	Accuracy(%)	F1-score
SVM (visible images only)	12314	600	1425	6827	90.43	0.8708
SVM2K (multi-view)	12328	586	1234	7018	91.40	0.8852
MMDBM + SVM (multi-view)	11720	1194	4006	4246	75.43	0.6202
DCCA + SVM (multi-view)	12093	821	2054	6198	86.42	0.8117
DCCAE + SVM (multi-view)	10333	969	1573	5873	86.44	0.8221
Single view structure (visible images only)	12476	438	837	7415	93.98	0.9208
Ours (multi-view)	12490	424	816	7436	**94.14**	**0.9230**

TABLE II
RESULTS OF THE LAUGHTER VERSUS POSED LAUGHTER DISCRIMINATION (WITH LAUGHTER AS A POSITIVE CLASS AND POSED LAUGHTER AS A NEGATIVE CLASS)

Experiment	TN	FP	FN	TP	Accuracy(%)	F1-score
SVM (visible images only)	643	794	344	1780	68.04	0.7578
SVM2K (multi-view)	641	796	426	1698	72.40	0.7761
MMDBM + SVM (multi-view)	254	1183	152	1972	62.51	0.7471
DCCA + SVM (multi-view)	694	743	501	1623	65.07	0.7229
DCCAE + SVM (multi-view)	746	690	439	1685	68.29	0.7491
Single view structure (visible images only)	1077	360	211	1913	83.97	0.8701
Ours (multi-view)	1088	349	166	1958	**85.54**	**0.8838**

during training, thereby obtaining better facial representations and a better visible expression classifier.

The proposed method also performs more accurately and has a higher F1-score compared to four state-of-the-art multiview methods. Specifically, the accuracy of our proposed method in the laughter versus posed laughter discrimination is 13.14%, 23.03%, 20.47%, and 17.25% higher than SVM2K, MMDBM, DCCA, and DCCAE, respectively. SVM2K trains classifiers for recognition directly and it is unable to capture complex structures in visible and thermal images the way a deep network can. MMDBM learns feature representation in the common space of two views and tries to minimize the average reconstruction error of the data in an unsupervised manner. The average reconstruction error is a poor metric to evaluate whether the features are suitable or not. DCCA/DCCAE captures relations between visible and thermal images. Feature representations learned by DCCA/DCCAE are not specific enough for expression recognition. Compared with these four multiview methods, the proposed method is able to learn feature representation by complex nonlinear transformations of two views, and then refine the feature by training deep networks and SVMs simultaneously with the supervision of target expression labels.

SVM2K further outperforms three state-of-the-art multiview methods. In laughter versus posed laughter discrimination, the accuracy of SVM2K is 9.89%, 7.33%, and 4.11% higher than MMDBM, DCCA, and DCCAE, respectively. SVM2K focuses on training a classifier for recognition directly in a supervised manner. MMDBM, DCCA, and DCCAE only learn feature representation without considering labels. It is more important in the expression recognition tasks to consider the relationship between the two views of data and labels. SVM2K performs better than the other three multiview methods in this distinction.

We see better performance in the former experiment for the laughter versus speech and laughter versus posed laughter discrimination tasks. It is evidently easier to capture the difference between laughter and speech. With respect to the laughter versus posed laughter discrimination task, the only difference between the two classes is spontaneous and posed laughter, which is harder to classify.

3) Analyses of the Learned Representations: We further demonstrated the effectiveness of features learned by DNN by investigating and visualizing the learned features of the data in 2-D plots with PCA. Three visualizations of feature space were evaluated: the feature space of input images, the top layer of pretraining DNN and the top layer of fine-tuning DNN. For each feature space, there are two kinds of visualization plots. One is a 2-D plot with respect to all samples tagged with its target labels and shows the distribution of two kinds of samples in a specific feature space. The other is a 2-D plot with respect to the samples tagged with subject IDs, showing the variance of different individuals. In the latter 2-D plot each subject ID is printed at the centroid of its samples.

The visualizations given in Figs. 3 and 4, respectively, show that in the feature space of raw data, samples labeled by two classes are naturally blended together. This demonstrates the difficulty of recognition tasks. Centroids of different subjects are dispersive, indicating a high variance of individuals. In the feature space learned by pretraining DNN, samples labeled by two classes are still blended. This is due to the unsupervised method of pretraining that excludes true label information. Compared with raw data space, samples with different subjects begin to blend slightly, though outliers are present. Samples with different subjects tend to be more clustered in feature space learned by fine-tuning DNN. In visible and thermal views samples with different classes are clearly separated. The plot demonstrates that the proposed deep two-view support vector machine learns facial representations that are discriminative to expression recognition and robust to individual variance. These features are beneficial to our top SVM classifiers and promote better recognition performance.

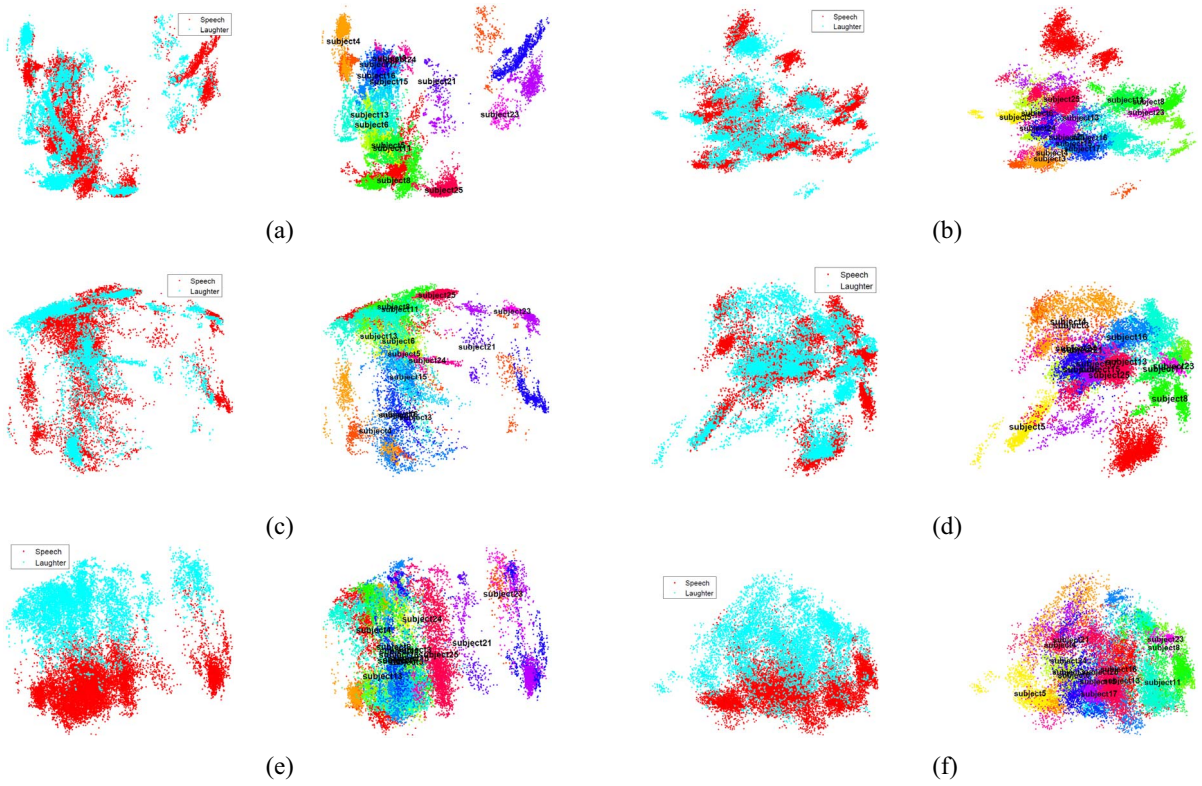


Fig. 3. PCA visualizations of the laughter versus speech dataset. (a) Visible view raw data. (b) Thermal view raw data. (c) Visible view pretraining DNN features. (d) Thermal view pretraining DNN features. (e) Visible view fine-tuning DNN features. (f) Thermal view fine-tuning DNN features.

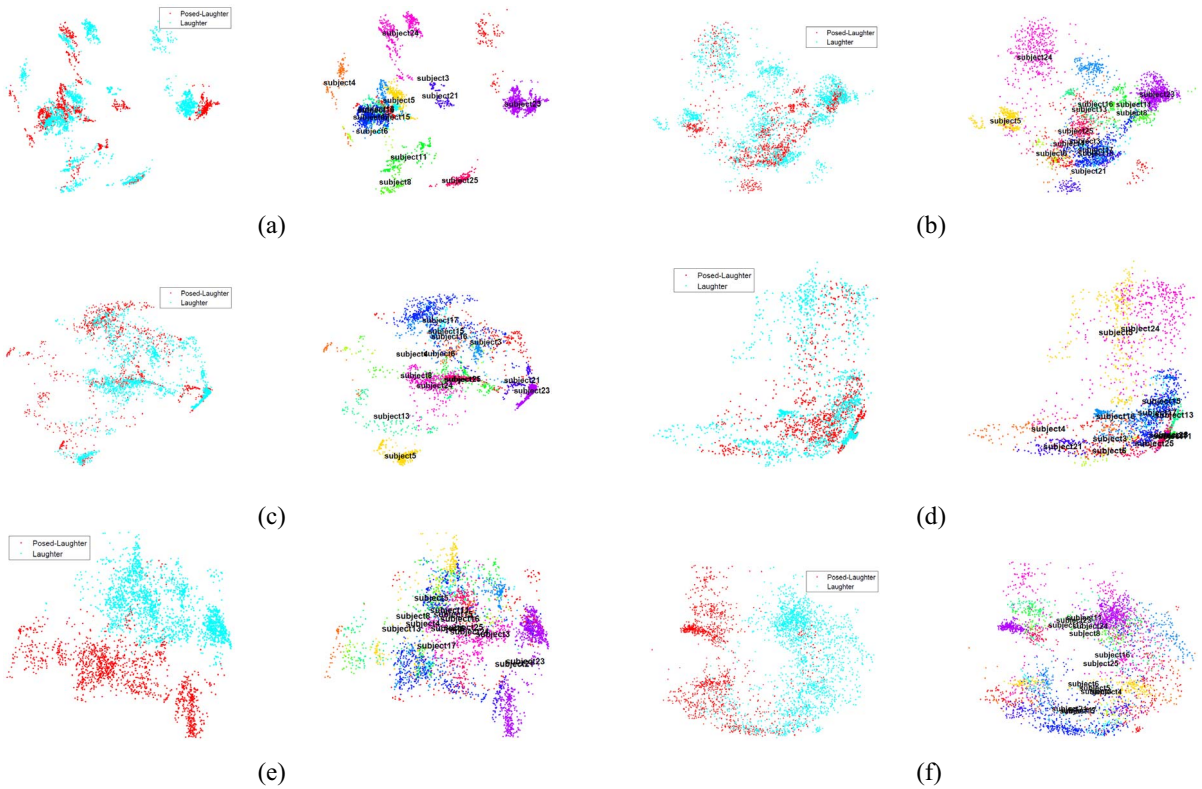


Fig. 4. PCA visualizations on the laughter versus posed laughter dataset. (a) Visible view raw data. (b) Thermal view raw data. (c) Visible view pretraining DNN features. (d) Thermal view pretraining DNN features. (e) Visible view fine-tuning DNN features. (f) Thermal view fine-tuning DNN features.

4) Analyses of the Misclassified Samples: Although the proposed method achieves good performance on expression recognition, it still misclassifies some samples. To analyze the root cause of the failure cases of the proposed method, we manually checked the misclassified samples, and found that most of them are difficult to discriminate from images. Take

TABLE III
COMPARISON RESULTS OF OUR PROPOSED METHOD WITH RELATED WORKS

Experiment	Accuracy(%)	F1-Laughter	F1-Speech
NN (Petridis et al. [34], audio + visual)	90.1	0.865	0.922
BLR (Rudovic et al. [36], audio + visual)	92.7	0.905	0.943
PF (Petridis et al. [37], audio + visual)	-	0.893	0.932
Ours (visual + thermal)	**94.14**	**0.9230**	**0.9527**

The accuracy of PF method is not reported in [37].

TABLE IV
RESULTS OF LAUGHTER VERSUS SPEECH DISCRIMINATION WITH PARTIALLY UNPAIRED DATA

unpaired samples ratio	Single view structure(visible images only)		Two views structure(paired-only)		Ours	
	Accuracy(%)	F1-score	Accuracy(%)	F1-score	Accuracy(%)	F1-score
10%	92.10±1.09	0.8966±0.0148	92.65±0.60	0.9035±0.0082	**93.39**±0.64	**0.9132**±0.0087
20%	92.06±0.89	0.8961±0.0123	92.24±1.02	0.8967±0.0153	**93.40**±0.62	**0.9133**±0.0087
30%	91.98±0.69	0.8944±0.0086	92.53±0.80	0.9025±0.0111	**93.23**±0.41	**0.9111**±0.0056
40%	91.44±0.83	0.8873±0.0113	91.46±0.68	0.8871±0.0093	**92.98**±0.55	**0.9074**±0.0079
50%	91.58±1.08	0.8896±0.0144	92.22±0.71	0.8982±0.0085	**92.84**±0.98	**0.9052**±0.0142

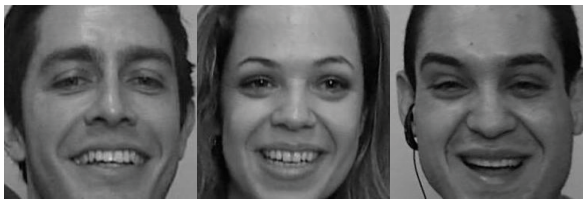


Fig. 5. Challenging samples from the laughter versus speech dataset.

laughter versus speech discrimination for example, some challenging samples are speaking samples, but the shape of the subject's month look as if the subject is really laughing, as shown in Fig. 5. Although the proposed method successfully explores thermal images to help discriminate laughter versus speech from visible images in most cases, for these challenging samples, we may require extra modality, such as audio, to distinguish.

5) Comparison With Related Works: Experiments conducted on the MAHNOB laughter database are primarily for the laughter versus speech discrimination using both audio and visual information. Therefore, we have to compare the proposed method with related work that uses different data channels. The comparison is only for reference due to different architectures and different data channels. In Table III, we compare our results with three related works: [34], [36], and [37]. In these three works facial point information for visual data and Mel frequency cepstral coefficients for audio data were extracted for classification. They differ only in their classification methods. Neural network, bimodal log-linear regression, and prediction-based fusion methods were applied, respectively.

Compared with the performance of the first three rows in Table III, the recognition accuracy and F1-score of our method excels. Visual and audio features in the alternate methods are hand-crafted and typically limited. Conversely, our features are first extracted by DNN and then are fine-tuned in the back propagation phase. These stable features are more suitable for expression recognition tasks.

In the experiments of related works, a sample is composed of a paired audio and video episode. We used a paired visual and thermal frame as a sample in our experiment. Although the proposed method does not leverage audio signals and temporal information for the laughter versus speech discrimination as the related work did, the proposed method still outperformed the related works. This demonstrates that thermal images can be used to help discriminate laughter versus speech from visible images rather than audio signals. While the proposed method requires thermal and visible images during training, it exclusively requires visible images during testing. The compared related works require audio and video during both training and testing. Our proposed method is more practical and requires less information during testing.

C. Experimental Results and Analyses for Partially Unpaired Data

Experimental results of expression recognition with partially unpaired data are shown in Tables IV and V with respect to the laughter versus speech and laughter versus posed laughter discrimination. From the results, we can find the following observations.

In two expression recognition tasks, the proposed method outperforms the single-view structure in accuracy and F1-score. In the single-view method only visible data were used for classification, resulting in a limited top SVM classifier. When we compared our proposed method with the single-view method, we found that thermal data are used as privileged information. This impacts the visible SVM classifier through the similarity constraint between the outputs of two SVMs. Our method performs better in this case.

The proposed method outperforms the same network structure with only paired samples in the majority of cases. In the paired-only method with a higher unpaired sample ratio, fewer paired samples can be used for pretraining. Paired and unpaired samples used for pretraining are not influenced by the unpaired sample ratio. This is due to the fact that paired and unpaired samples are always used in the training set. This proves that including more samples in the pretraining phase will result in more robust features learned by DNNs, enhancing performance in this currently unsupervised method.

TABLE V
RESULTS OF LAUGHTER VERSUS POSED LAUGHTER DISCRIMINATION WITH PARTIALLY UNPAIRED DATA

unpaired	Single view structure(visible images only)		Two views structure(paired-only)		Ours	
samples ratio	Accuracy(%)	F1-score	Accuracy(%)	F1-score	Accuracy(%)	F1-score
10%	79.46±1.93	0.8390±0.0125	**80.54**±1.49	**0.8454**±0.0106	80.16±0.90	0.8429±0.0078
20%	78.66±1.76	0.8299±0.0160	79.03±2.31	0.8376±0.0131	**80.20**±1.47	**0.8433**±0.0115
30%	78.24±2.27	0.8288±0.0156	79.53±2.22	**0.8394**±0.0142	**79.68**±1.10	0.8392±0.0092
40%	78.66±2.53	0.8305±0.0156	79.36±1.88	0.8366±0.0133	**79.74**±1.45	**0.8398**±0.0117
50%	78.35±2.53	0.8291±0.0167	78.01±3.21	0.8309±0.0227	**79.71**±1.70	**0.8403**±0.0131

The performance of the three methods in both expression recognition tasks generally decreases with higher unpaired sample ratios. This is a reasonable outcome to expect when fewer available samples are used for both pretraining and back propagation phases. With higher unpaired sample ratios, the performance of the two views structure decreases drastically while performance of the other two methods decreases marginally. More samples used for pretraining can compensate for fewer samples used for back propagation.

V. CONCLUSION

We propose a deep two-view learning approach for expression recognition aided by IR images. Visible and thermal facial images represent two expression views. Feature representation requires two deep networks trained for visible and thermal image data. Two support vector machines are trained for classification. Layer-wise RBM training method is applied during deep network pretraining. Deep network refining and SVM parameter learning are integrated into a single optimization problem through the similarity constraint between outputs of two SVMs. Thermal image data are used as privileged information to help expression recognition fine tune visible image data. During testing, only visible image data are used for classification. The proposed approach is extended to utilize paired and unpaired images. Experimental results on the MAHNOB laughter database demonstrate that the proposed method can capture relationships among two view data and their labels, and can effectively recognize expressions on visible images.

REFERENCES

[1] V. Bettadapura, "Face expression recognition and analysis: The state of the art," *Comput. Sci.*, 2012.

[2] Y. Yoshitomi, S.-I. Kim, T. Kawano, and T. Kilazoe, "Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face," in *Proc. 9th IEEE Int. Workshop Robot Human Interact. Commun. (RO MAN)*, Osaka, Japan, 2000, pp. 178–183.

[3] S. Jarlier *et al.*, "Thermal analysis of facial muscles contractions," *IEEE Trans. Affect. Comput.*, vol. 2, no. 1, pp. 2–9, Jan./Jun. 2011.

[4] Y. Yoshitomi, T. Asada, R. Kato, and M. Tabuse, "Facial expression recognition using facial expression intensity characteristics of thermal image," *J. Robot. Netw. Artif. Life*, vol. 2, no. 1, pp. 5–8, 2015.

[5] S. He, S. Wang, W. Lan, H. Fu, and Q. Ji, "Facial expression recognition using deep Boltzmann machine from thermal infrared images," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact. (ACII)*, Geneva, Switzerland, 2013, pp. 239–244.

[6] A. Wesley, P. Buddharaju, R. Pienta, and I. Pavlidis, "A comparative analysis of thermal and visual modalities for automated facial expression recognition," in *Proc. Int. Symp. Vis. Comput.*, 2012, pp. 51–60.

[7] S. Wang, S. He, Y. Wu, M. He, and Q. Ji, "Fusion of visible and thermal images for facial expression recognition," *Front. Comput. Sci.*, vol. 8, no. 2, pp. 232–242, 2014.

[8] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Netw.*, vol. 22, nos. 5–6, pp. 544–557, 2009.

[9] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image Vis. Comput.*, vol. 30, no. 10, pp. 683–697, 2012.

[10] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[11] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016.

[12] V. Pérez-Rosas, A. Narvaez, M. Burzo, and R. Mihalcea, "Thermal imaging for affect detection," in *Proc. 6th Int. Conf. Pervasive Technol. Related Assistive Environ. (PETRA)*, 2013, pp. 1–4. [Online]. Available: http://doi.acm.org/10.1145/2504335.2504374

[13] P. Liu and L. Yin, "Spontaneous facial expression analysis based on temperature changes and head motions," in *Proc. FG*, Ljubljana, Slovenia, 2015, pp. 1–6.

[14] H. Nguyen, K. Kotani, F. Chen, and B. Le, "Estimation of human emotions using thermal facial information," in *Proc. Int. Conf. Graph. Image Process. Soc. Opt. Photon.*, 2014, pp. 361–368.

[15] B. R. Nhan and T. Chau, "Classifying affective states using thermal infrared imaging of the human face," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 979–987, Apr. 2010.

[16] M. M. Khan, R. D. Ward, and M. Ingleby, "Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature," *ACM Trans. Appl. Percept.*, vol. 6, no. 1, p. 6, 2009.

[17] N. Sharma, A. Dhall, T. Gedeon, and R. Goecke, "Modeling stress using thermal facial patterns: A spatio-temporal approach," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact. (ACII)*, Geneva, Switzerland, 2013, pp. 387–392.

[18] A. Poursaberi, S. Yanushkevich, and M. Gavrilova, "An efficient facial expression recognition system in infrared images," in *Proc. 4th Int. Conf. Emerg. Security Technol.*, Cambridge, U.K., 2013, pp. 25–28.

[19] B. Hernandez, G. Olague, R. Hammoud, L. Trujillo, and E. Romero, "Visual learning of texture descriptors for facial expression recognition in thermal imagery," *Comput. Vis. Image Understand.*, vol. 106, nos. 2–3, pp. 258–269, 2007.

[20] X. Shi, S. Wang, and Y. Zhu, "Expression recognition from visible images with the help of thermal images," in *Proc. 5th ACM Int. Conf. Multimedia Retrieval*, Shanghai, China, 2015, pp. 563–566.

[21] S. Sun, "A survey of multi-view machine learning," *Neural Comput. Appl.*, vol. 23, nos. 7–8, pp. 2031–2038, 2013.

[22] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Inf. Fusion*, vol. 38, pp. 43–54, Nov. 2017.

[23] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, vol. 2. New York, NY, USA: Wiley, 1958.

[24] S. Akaho, "A kernel method for canonical correlation analysis," in *Proc. Int. Meeting Psychometric Soc. (IMPS)*, vol. 40, 2006, pp. 263–269.

[25] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, 2015, pp. 1083–1092.

[26] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. 30th Int. Conf. Mach. Learn.*, Atlanta, GA, USA, 2013, pp. 1247–1255.

[27] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[28] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, Bellevue, WA, USA, 2011, pp. 689–696.

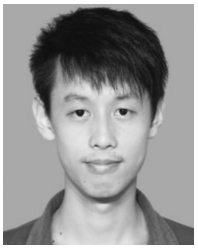
[29] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2222–2230.

[30] J. Farquhar, D. Hardoon, H. Meng, J. S. Shawe-Taylor, and S. Szedmak, "Two view learning: SVM-2K, theory and practice," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 355–362.
[31] Y. Tang, "Deep learning using linear support vector machines," *Comput. Sci.*, 2013.
[32] S.-X. Zhang, C. Liu, K. Yao, and Y. Gong, "Deep neural support vector machines for speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, 2015, pp. 4275–4279.
[33] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[34] S. Petridis, B. Martinez, and M. Pantic, "The MAHNOB laughter database," *Image Vis. Comput.*, vol. 31, no. 2, pp. 186–202, 2013.
[35] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, 2001, pp. I-511–I-518.
[36] O. Rudovic, S. Petridis, and M. Pantic, "Bimodal log-linear regression for fusion of audio and visual features," in *Proc. 21st ACM Int. Conf. Multimedia*, Barcelona, Spain, 2013, pp. 789–792.
[37] S. Petridis, V. Rajgarhia, and M. Pantic, "Comparison of single-model and multiple-model prediction-based audiovisual fusion," in *Proc. Joint Conf. Facial Anal. Animat. Auditory Vis. Speech Process.*, 2015, pp. 457–462.

Shangfei Wang (SM'15) received the B.S. degree in electronic engineering from Anhui University, Hefei, China, in 1996 and the M.S. degree in circuits and systems and the Ph.D. degree in signal and information processing from the University of Science and Technology of China (USTC), Hefei, in 1999 and 2002, respectively.

From 2004 to 2005, she was a Post-Doctoral Research Fellow with Kyushu University, Fukuoka, Japan. In 2011 and 2012, she was a Visiting Scholar with Rensselaer Polytechnic Institute, Troy, NY, USA. She is currently an Associate Professor with the School of Computer Science and Technology, USTC. She has authored or co-authored over 90 publications. Her current research interests include affective computing and probabilistic graphical models.

Dr. Wang is a member of the ACM.



Bowen Pan received the B.S. degree in computer science from Sichuan University, Chengdu, China, in 2016. He is currently pursuing the M.S. degree in computer science with the University of Science and Technology of China, Hefei, China.

His current research interest includes affective computing.

Huaping Chen received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, 1997.

He is a Professor with the School of Computer Science and Technology, University of Science and Technology of China. His current research interests include evolutionary algorithms, nonlinear programming, project scheduling, and resource scheduling problems.

Qiang Ji (F'15) received the Ph.D. degree in electrical engineering from the University of Washington, Seattle, WA, USA.

He is currently a Professor with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute (RPI), Troy, NY, USA. From 2009 to 2010, he served as a Program Director with the National Science Foundation (NSF), Arlington, VA, USA, where he managed NSFŠs computer vision and machine learning programs. He also held teaching and research positions with the Beckman Institute, University of Illinois at Urbana–Champaign, Urbana, IL, USA, the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, the Department of Computer Science, University of Nevada, Reno, NV, USA, and the Air Force Research Laboratory, Rome, NY, USA. He currently serves as the Director of the Intelligent Systems Laboratory, RPI. He has published over 230 papers in peer-reviewed journals and conferences. His current research interests include computer vision, probabilistic graphical models, machine learning, and their applications in various fields.

Prof. Ji was a recipient of multiple awards for his work. He is an editor on several related IEEE and international journals and he has served as the general chair, the program chair, the technical area chair, and a program committee member for numerous international conferences/workshops. He is a fellow of IAPR.