Information Sciences 533 (2020) 60–71

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Discriminative deep multi-task learning for facial expression recognition

Hao Zheng [a], Ruili Wang [b,*], Wanting Ji [b], Ming Zong [b], Wai Keung Wong [c], Zhihui Lai [d], Hexin Lv [e]

[a] School of Information Engineering, Nanjing Xiaozhuang University, Nanjing 211171, China
[b] School of Natural and Computational Sciences, Massey University, Auckland, New Zealand
[c] Institute of Textiles Clothing, The Hong Kong Polytechnic University, Hong Kong, China
[d] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518055, China
[e] Institute of Information Technology, Zhejiang Shuren University, Hangzhou 321028, China

ARTICLE INFO

Article history:
Received 17 November 2019
Received in revised form 16 April 2020
Accepted 20 April 2020
Available online 7 May 2020

Keywords:
Deep multi-task learning
Discriminative
Facial expression recognition

ABSTRACT

Deep multi-task learning (DMTL) is an efficient machine learning technique that has been widely utilized for facial expression recognition. However, current deep multi-task learning methods typically only consider the information of class labels, while ignoring the local information of sample spatial distribution. In this paper, we propose a discriminative DMTL (DDMTL) facial expression recognition method, which overcomes the above shortcomings by considering both the class label information and the samples' local spatial distribution information simultaneously. We further design a siamese network to evaluate the local spatial distribution through an adaptive reweighting module, utilizing the class label information with different confidences. In addition, by taking the advantage of the provided local distribution information of samples, DDMTL is able to achieve acceptable results even if the number of training samples is small. We implement experiments on three facial expression datasets. The experimental results demonstrate that DDMTL is superior to the state-of-the-art methods.

1. Introduction

Facial expression recognition is a challenging task in computer vision [1,22,46,47]. In [15], Ekman developed the first facial expression recognition system, named the Facial Action Coding System. Since then, many machine learning techniques, such as scale-invariant feature transform (SIFT) [3], histograms of oriented gradients (HOG) [13], local binary patterns (LBP) [33], and local Gabor binary patterns (LGBP) [49], were introduced to facial expression recognition. These facial expression recognition methods achieved acceptable performances in processing small training sets. However, the generalization ability of these methods is limited. In other words, these methods were not easily adjusted to recognize new facial expressions.

Recently, some deep neural networks were developed for image processing [11,34,40,42,46] and facial expression recognition [18,36,37], such as Convolutional Neural Networks (CNNs). Since the convolution and pooling layers of a CNN are capable of extracting multi-level features from facial images, these CNN-based methods have obtained acceptable performance for facial expression recognition when the sample is adequate. However, these methods are incapable of making full use of the information in facial images. To address this issue, some methods [9,16,21,24,32] have been subsequently developed for

* Corresponding author.
E-mail address: Ruili.wang@massey.ac.nz (R. Wang).

https://doi.org/10.1016/j.ins.2020.04.041
0020-0255/

facial expression recognition. However, due to the subtlety and complexity of human facial expressions, these methods do not perform well under real-world conditions.

More recently, deep learning based multi-task learning, named deep multi-task learning (DMTL), is utilized to address the above issue in facial expression recognition. Multi-task learning [8] is a machine learning paradigm, which leverages valuable information contained in multiple correlated tasks to improve all correlated tasks' performance. DMTL is to operate multi-task learning in deep neural networks. Many DMTL-based methods were developed for facial expression recognition [30,43,48], in which facial features can be recognized using DMTL to deduce facial information. However, the existing DMTL-based facial expression recognition methods typically only consider the class label information, while ignoring the local information of sample spatial distribution.

In this paper, we develop a Discriminative Deep multi-task Learning (DDMTL) method to address the above issue in facial expression recognition, which is able to consider both the class label information and the local information of sample spatial distribution simultaneously. We further design a weight-shared siamese network to evaluate the local distribution through an adaptive reweighting module, utilizing the class label information with different confidence. Since the proposed network is a two-channel siamese network with shared weights, it can be regularized via contrastive loss. This can prevent the focus of DDMTL only on the class label information. Therefore, DDMTL can obtain acceptable test performance and achieve satisfactory experimental results even when the number of training samples is small. A simplified flowchart of DDMTL is illustrated in Fig. 1.

The main contributions of this paper include:

1) A novel discriminative DMTL (DDMTL) method is proposed, including a discriminative softmax loss and a contrastive loss. Notably, this is the first study that adds discriminative local spatial distribution information to the softmax loss.
2) An adaptive reweighting module is developed, which utilizes the label information with different confidence.
3) DDMTL considers both the information of class labels and the local information of sample spatial distribution simultaneously, and overcomes the challenges in facial expression recognition. Experimental results demonstrate that DDMTL is superior to the state-of-the-art.

The rest of this paper is organized as follows. Section 2 presents related work. The proposed DDMTL method is detailed in Section 3. Section 4 conducts experiments to validate the effectiveness of DDMTL. Section 5 is the conclusion of this paper.

2. Related work

2.1. Multi-task learning

Multi-task learning is a machine learning paradigm that leverages valuable information contained in multiple correlated tasks to improve all correlated tasks' performance. It is a kind of transfer learning by sharing information among different tasks and exploiting the similarity among these tasks. Lukaszr et al. [26] proposed a multi-task learning model that could simultaneously learn several similar tasks related to language processing and computer vision. Rothe et al. [31] considered an age regression task as a deep classification task and utilized multi-task learning to achieve image classification. Subsequently, Rajeev et al. [29] developed a multi-task learning algorithm for face classification, pose estimation, and gender recognition.

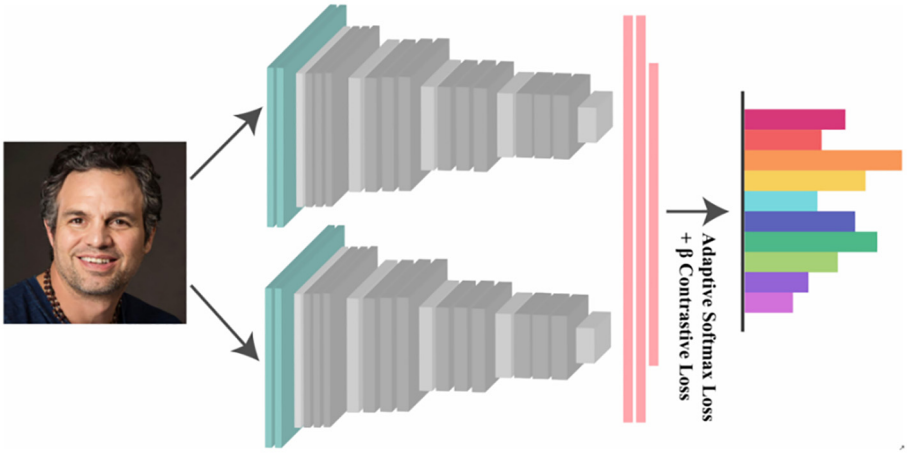

Fig. 1. A flowchart to illustrate DDMTL.

2.2. Convolutional neural networks

With the development of deep learning techniques, multiple Convolutional Neural Networks (CNNs) have proven their superior power in many research areas, especially in image processing and video processing. The widely accepted CNN model was developed by LeCun et al. [20]. They proposed an end-to-end training of CNN models by using gradient based optimization. Comparing with various conventional handwritten character recognition methods, LeCun et al. [20] demonstrated that combining convolutional neural networks with either search mechanism or inference mechanism had the ability to simulate interdependent complicated outputs with minimal pre-processing.

Alex et al. [2] developed a large deep CNN model called AlexNet. It had achieved the state-of-the-art performance at that time by classifying 1.3 million high-resolution images into 1000 different classes. Furthermore, some other convolutional neural networks such as VGGNet [35], GoogLeNet [38], and ResNet [19] were developed and achieved good performance (such as high classification accuracy) by increasing the depth and the width of CNNs. Specifically, VGGNet [35] has achieved significant improvement to the prior art configurations by increasing the network depth to 16–19 weighted layers. Szegedy et al. [38] developed GoogLeNet by increasing both the depth and the width of CNNs. They increased the number of convolutional layers to 22. ResNet was developed by He et al. [19], in which the number of convolutional layers was set to 152.

In addition, DenseNet [17] achieved better performance by shortening the length of connections between the layer close to the input and the layer close to the output. This is also able to enhance feature propagation, mitigate the vanishing gradient problem, encourage feature reuse, and decrease the number of parameters.

2.3. Siamese network

The siamese network consists of two duplicate sub-networks using shared parameters. The siamese network was first developed in [5] for signature verification. Currently, siamese networks have been applied to many areas such as speech feature classification [10] and text classification [45].

Further, Chopra et al. [12] developed a siamese network for fully supervised face verification. The developed siamese network using a discriminative learning framework for energy based models [39] to calculate loss function. Mobahi et al. [27] proposed a similar siamese network for large-scale object recognition in videos. The proposed network is capable of labelling pairs of input videos frames as similar (when they are close to each other in time) or dissimilar (when they are far away from each other in time). Liu et al. [23] proposed a parameterized probabilistic siamese network for learning representation, in which learning was formulated as maximizing the likelihood of binary similarity labels for pairs of input images. In addition, Bertinetto et al. [4] developed an end-to-end fully convolutional siamese network for video object detection.

3. Proposed method

In this section, we detail how discriminative information is added to the softmax loss function of the deep multi-task learning algorithm. Then we present how to design a reweighting module and calculate the contrastive loss. Finally, the optimization of DDMTL is illustrated.

3.1. Discriminative deep multi-task learning (DDMTL)

In the existing DMTL based methods, the discriminative information from the spatial distribution of samples has not been considered. Hence, samples may be classified inaccurately in some cases by using these methods. In order to address this issue, we propose a discriminative DMTL based method. In DDMTL, the discrimination is employed with softmax loss using an adaptive reweighting module to obtain the local spatial distribution information from samples. Furthermore, in order to combine class label information with constraints loss, we design a siamese network with two types of loss functions: (i) a discriminative softmax loss, which uses the information of class labels and the information of local spatial distribution of samples, and (ii) a contrastive loss, which contributes to effectively preventing the overfitting problem. Suppose that given a dataset X consisting of N samples from M different classes, then the discriminative softmax loss can be formulated as follows:

$$SLoss(x^i, \Theta, y_i) = -\sum_{t=1}^{M} 1\{y_i == t\} log\widehat{p_t} = -log\widehat{p_{y_i}} = -log\frac{e^{\Theta_{y_i}^T x^i}}{\sum_{t=1}^{M} e^{\Theta_{y_i}^T x^i}}, \tag{1}$$

where x^i presents the network output corresponding to a training sample x^i, and $1\{y_i == t\}$ is an indicator function; if $y_i == t$ is true, then the result is 1; otherwise, the result is 0. Θ is the parameters of network layers; Θ_t is the weight of the t^{th} network output; $t = 1,\ldots, M$. $\widehat{p_{y_i}}$ is the predicted probability. The goal of DDMTL is to utilize discriminative information from training samples. Thus, we develop an adaptive reweighting module W_k that adds local spatial distribution information to the proposed siamese network. In addition, the reweighting module, which assigns weights to each sample based on the confidence between sample pairs, can both minimize the distance between similar samples and maximize the distance between dissimilar samples. Then the discriminative softmax loss of the proposed method can be defined as:

$$SL(x^i, x^j, \Theta_1, \Theta_2, y_i, y_j) = \sum_{i,j,k}^{N} W_k\left(SLoss(x^i, \Theta_1, y_i) + SLoss(x^j, \Theta_2, y_j)\right), \tag{2}$$

where x^i and x^j are the output of the network corresponding to training sample x^i and x^j; y_i and y_j are the class labels corresponding to x^i and x^j; Θ_1 and Θ_2 are the parameters of the two softmax losses. The adaptive reweighting module W_k improves classification performance by providing specific local distribution information.

3.2. Adaptive reweighting module W_k

In order to design an adaptive reweighting module W_k, it is considered that samples from the same class can be represented with high-level representations, while samples from different class cannot. Thus, $pCON(x^i, x^j)$ can be defined as follows:

$$pCON(x_i, x_j) = \sum_{x_j \in N_k(x_i)} \frac{max(k - dist(x_i), d(x_i, x_j))}{max(k - dist(x_i), d(x_i, x_i))}, \tag{3}$$

where $N_k(x_i)$ is the set of k nearest neighbors of x_i, and $\max\{k\text{-}dist(x_i), d(x_i,x_j)\}$ is the reachability distance from x_i to x_j. In other words, if x_i and x_j are sufficiently close, then the reachability distance is $k\text{-}dist(x_i)$; If x_i is far away from x_j, then the reachability distance is $d(x_i,x_j)$. Since it is a probabilistic algorithm, the adaptive reweighting module W_k can be defined as:

$$W_k = G\left(\sum_{t=1}^{T} pCON(x_i, x_j)\right), \tag{4}$$

where T presents the number of iterations, and G is the local Gaussian statistics transformation that can be used to scale the probabilistic value. Since DDMTL works iteratively, a complicated procedure is not essential. Furthermore, the discriminative representation can converge to similar results with sufficiently long iterations.

Fig. 2 illustrates the reachability distance when $k = 5$. If object x_j is far away from x_i (such as object x_{j2} in Fig. 2), the reachability distance between the two objects (i.e., object x_j and object x_{j2}) is their actual distance; if object x_j is sufficiently close to x_i (such as object x_{j1} in Fig. 2), the actual distance will be replaced by the k-distance of x_i when calculating the reachability distance between the two objects (i.e., object x_j and object x_{j1}). This is due to the statistical fluctuation of $d(x_i, x_j)$ for all x_j close to x_i can be significantly reduced. Thus, parameter k can control the strength of the smoothing effect, which will lead to achieving a similar reachability distance for objects in proximity.

3.3. Contrastive loss constraints

The role of local distribution constraint is to learn the non-linear mapping of similar input vectors to nearby points as well as the non-linear mapping of dissimilar input vectors to distant points on the output manifold. In this paper, we utilize a contrastive loss to impose the constraint. Suppose that, in a d-dimensional feature space R^d, x_i and x_j are two input training samples, then the similarity indicator $Y_{i,j}$ can be defined as:

$$Y_{i,j} = \begin{cases} 1, & \text{if } x_i, x_j \text{ are dissimilar} \\ 0, & \text{if } x_i, x_j \text{ are similar} \end{cases}. \tag{5}$$

In a deep representation space, the Euclidean distance between x_i and x_j is denoted as follows:

$$D_k(x_i, x_j, \varphi) = \|f^k(x_i|\varphi) - f^k(x_j|\varphi)\|_2, \tag{6}$$

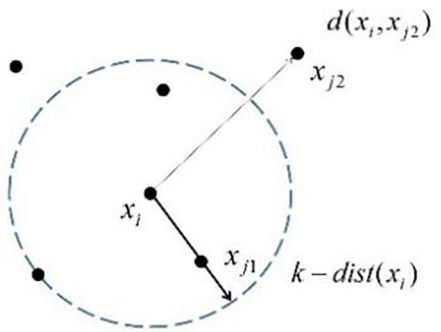

Fig. 2. $d(x_i, x_{j2})$ and $k\text{-}dist(x_i)$.

where $f^k(x_i|\varphi)$ presents the k^{th} layer output of the network $f(\bullet|\varphi)$ with the parameter φ. If two samples are from the same class, then they can be considered as "similar" (i.e., $Y_{i,j} = 0$); if they are from different classes, then they can be considered as "dissimilar" (i.e., $Y_{i,j} = 1$). Thus, the contrastive loss of discriminative feature learning can be formulated as:

$$CL(x_i, x_j, \varphi, k) = 0.5(1 - Y_{i,j})\left(D_k(x_i, x_j, \varphi)\right)^2 + 0.5Y_{i,j}\{max(0, m - D_k(x_i, x_j, \varphi))\}, \tag{7}$$

where $m > 0$ is a margin that formulates how far away that two dissimilar samples should be pulled from each other.

Since the contrastive loss can be seen as a regularization of the proposed network, DDMTL can effectively prevent the overfitting problem and obtain acceptable results even in the case of a small number of training samples.

3.4. Optimization

In this section, the network architecture that combines classification with contrastive loss is described. Our proposed method is jointly optimized by two loss terms:

$$L = SL(x^i, x^j, \Theta_1, \Theta_2, y_1, y_2) + \beta CL(x_i, x_j, \varphi, k), \tag{8}$$

where SL denotes the discriminative softmax loss; CL denotes the contrastive loss, and β denotes a trade-off parameter. We update the parameters of the proposed network using a stochastic gradient descent method. Each iteration traverses through all batch blocks of the training set. When a batch block is traversed, the network parameters will be updated, then all network parameters can be updated after the batch is completed. The updating formula is:

$$\Theta^{i+1} = \Theta^i - \eta\nabla\Theta, \tag{9}$$

where i denotes the iteration number; Θ presents the parameters of network layers, and η presents the learning rate. Therefore,

$$\Theta_1 = \Theta_1 - \eta\nabla\Theta_1, \Theta_2 = \Theta_2 - \eta\nabla\Theta_2, \tag{10}$$

where

$$\begin{aligned} \nabla\Theta_1 &= \partial SLoss(x^i, x^j, \Theta_1, \Theta_2, y_1, y_2)/\partial\Theta_1 \\ \nabla\Theta_2 &= \partial SLoss(x^i, x^j, \Theta_1, \Theta_2, y_1, y_2)/\partial\Theta_2 \end{aligned}. \tag{11}$$

Since the partial derivative of the k^{th} layer output of the network $f(\bullet|\varphi)$ with parameters φ can be calculated as follow:

$$\begin{aligned} \nabla f^k(x_i|\varphi_k) &= \frac{\partial CL(x_i, x_j, \varphi, k)}{\partial f^k(x_i|\varphi_k)} + \frac{\partial SLoss(x^i, x^j, \Theta_1, \Theta_2, y_i, y_j)}{\partial f^k(x_i|\varphi_k)} \\ \nabla f^k(x_j|\varphi_k) &= \frac{\partial CL(x_i, x_j, \varphi, k)}{\partial f^k(x_j|\varphi_k)} + \frac{\partial SLoss(x^i, x^j, \Theta_1, \Theta_2, y_i, y_j)}{\partial f^k(x_j|\varphi_k)} \end{aligned}, \tag{12}$$

the partial derivative of φ_k can be calculated as:

$$\nabla\varphi_k = \nabla f^k(x_i|\varphi_k) \times \partial f^k(x_i|\varphi_k)/\partial\varphi_k + \nabla f^k(x_j|\varphi_k) \times \partial f^k(x_j|\varphi_k)/\partial\varphi_k. \tag{13}$$

Algorithm 1 illustrates the process of DDMTL.

Algorithm 1 Discriminative Deep Multi-Task Learning to Recognize Facial Expression

(i): **Input:** $\{(x_i, y_i)\}, i = 1, 2, \cdots, N$.
(ii): **Initialization:** initializing the parameter of the network, parameter learning m, learning rate η
(iii): **While** not converged **do**
(iv): **input** (x_i, y_i) **and** $\left(x_j, y_j\right) i, j = 1, 2, \cdots, N$;
(v): $\nabla\Theta_1 = \partial SLoss(x^i, x^j, \Theta_1, \Theta_2, y_1, y_2)/\partial\Theta_1$;
(vi): $\nabla\Theta_2 = \partial SLoss(x^i, x^j, \Theta_1, \Theta_2, y_1, y_2)/\partial\Theta_2$;
(vii): $\nabla f^k(x_i|\varphi_k) = \frac{\partial CL(x_i, x_j, \varphi, k)}{\partial f^k(x_i|\varphi_k)} + \frac{\partial SLoss(x^i, x^j, \Theta_1, \Theta_2, y_i, y_j)}{\partial f^k(x_i|\varphi_k)}$;
(viii): $\nabla f^k(x_j|\varphi_k) = \frac{\partial CL(x_i, x_j, \varphi, k)}{\partial f^k(x_j|\varphi_k)} + \frac{\partial SLoss(x^i, x^j, \Theta_1, \Theta_2, y_i, y_j)}{\partial f^k(x_j|\varphi_k)}$;
(ix): $\nabla\varphi_k = \nabla f^k(x_i|\varphi_k) \times \partial f^k(x_i|\varphi_k)/\partial\varphi_k + \nabla f^k(x_j|\varphi_k) \times \partial f^k(x_j|\varphi_k)/\partial\varphi_k$;
Update $\Theta_1 = \Theta_1 - \eta\nabla\Theta_1, \Theta_2 = \Theta_2 - \eta\nabla\Theta_2, \varphi_k = \varphi_k - \eta\nabla\varphi_k$;
End While
(x): **Output:** parameters of the network.

4. Experiments

We implement DDMTL on three datasets: (i) the extended Cohn-Kanade dataset (CK +) [25], (ii) the MMI facial expression dataset (MMI) [41], and (iii) the Static Facial Expressions in the Wild dataset (SFEW) [14]. The direct pre-training of deep neural networks on small datasets may lead to the overfitting problem. Thus, our experimental model of DDMTL consists of 4 convolutional layers, a fully connected layer, and a softmax layer, where each convolutional layer is followed by one max-pooling layer and one local response normalization layer. Features that are extracted from the fully connected layer will be fed into the softmax layer for classification.

We set the initial network learning rate to 0.05, and the learning rate is decayed per epoch. We update the parameters of DDMTL by using stochastic gradient descent in the optimization processing. We compare DDMTL with the state-of-the-art, including AlexNet [2], VGGNet [35], GoogleNet [38], AdaGabor [50], 3D-CNN [7], SJMT [28] and DMTL [48] on the three datasets to validate the efficiency of DDMTL.

4.1. Extended Cohn-Kanade dataset

The extended Cohn-Kanade dataset is a widely used laboratory-controlled dataset for facial expression recognition. It consists of 327 video sequences from 118 subjects that are labelled with 7 expression class labels. The duration of a video sequence ranging from 10 to 60 frames. Each video sequence begins and ends with a neutral expression and can exhibit a shift from a neutral facial expression to a peak facial expression.

In this experiment, all video sequences are firstly pre-processed into 640 $\times$ 480-pixel images with 8-bit precision for grayscale. Then the last 4 frames are selected for each sequence due to maximal expressions (Fig. 3). In order to reduce the variation in face scale or in-plane rotation across different facial images, we align the obtained facial images based on the centres of the eyes and the mouth. Finally, all selected images are cropped to 50 $\times$ 50 pixels. All methods (including DDMTL and other reference methods) are trained and tested on these processed images

We divided the experimental dataset into 8 groups (no subject overlap among groups) and utilize an 8-fold cross-validation strategy (6 groups for training and the remaining 2 groups for evaluation and testing) for validation. All experimental results are reported as the average of the 8 runs.

Table 1 shows the recognition accuracy of AlexNet, VGGNet, GoogleNet, AdaGabor, 3D-CNN, SJMT, DMTL and DDMTL methods on the CK + dataset. Notably, DDMTL outperforms than the other methods; the accuracy of DDMTL is almost 7% higher than that of AlexNet. Also, DDMTL achieves high accuracy rate of 97.63%, with 95.18% for 3D-CNN, 93.40% for AdaGabor, 93.31% for GoogleNet, 92.53% for VGGNet, 95.11% for SJMT, 95.67% for DMTL.

Table 2 shows the confusion matrix of the DDMTL. The proposed DDMTL method has more than 98% recognition accuracy for disgust, happy and surprise expressions, and only 84% for angry and fear expressions, and the lowest recognition accuracy for sad expression. It is because the fear and sad expressions are similar to each other in some cases, they are not distinct in the pixel space, and hence, often confused. Furthermore, an optimal DDMTL is achieved using the local spatial distribution information among the samples. It minimizes the distance between similar samples and maximizes it between dissimilar samples.

To balance the two loss functions, we explore the weight by varying β from 0 to $+\infty$. Fig. 4 demonstrates that within a certain range, the recognition accuracy of DDMTL increases with increasing β. The maximal accuracy is at β = 0.01, then the accuracy begins to drop. According to the above, β is set as 0.01 in our experiments. To demonstrate the effectiveness of DDMTL, we visualize the features by AlexNet, 3D-CNN, and DDMTL on the CK + dataset. In Fig. 5, the dots, diamonds, and stars denote training, evaluation, and testing data, respectively. All sample images are randomly distributed, and the features are clustered according to the provided 6 expression labels. It can be seen that DDMTL (Fig. 5 (iii)) achieves a better separation performance than AlexNet (Fig. 5 (i)) and 3D-CNN (Fig. 5 (ii)).

4.2. MMI facial expression dataset

The MMI dataset is a laboratory-controlled dataset that includes 326 sequences from 32 subjects, in which 213 sequences are labelled with 6 types of expression labels, and 205 sequences are captured in frontal view. Each sequence begins and ends with a neutral expression, reaching the peak of facial expressions near the middle of the sequence. In our experiments, 3 frames near the middle of a sequence are chosen as peak frames and correlated with their expression labels. Fig. 6 shows

Fig. 3. Sample images of one person on the CK + dataset.

Table 1
Recognition accuracy on the CK + dataset.

Method	Validation Accuracy	Test Accuracy
AlexNet [2]	89.41 ± 1%	91.21 ± 3%
VGGNet [35]	90.33 ± 2%	92.53 ± 2%
GoogleNet [38]	91.90 ± 2%	93.31 ± 2%
AdaGabor [50]	91.07 ± 2%	93.40 ± 4%
3D-CNN [7]	93.53 ± 3%	95.18 ± 1%
SJMT [28]	93.39 ± 4%	95.11 ± 2%
DMTL [48]	93.51 ± 3%	95.67 ± 3%
DDMTL	**95.28** ± 2%	**97.63** ± 3%

Table 2
Confusion matrix of facial expression obtained by DDMTL on the CK + dataset.

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**84.28**	1.03	0	0.67	4.33	0
Disgust	3.21	**98.02**	0	0	0	0
Fear	4.89	0	**84.39**	0	10.89	0
Happy	0	0	6.04	**98.11**	0	1.64
Sad	4.57	0	5.21	1.22	**80.25**	0
Surprise	3.05	0.95	4.36	0	4.53	**98.36**

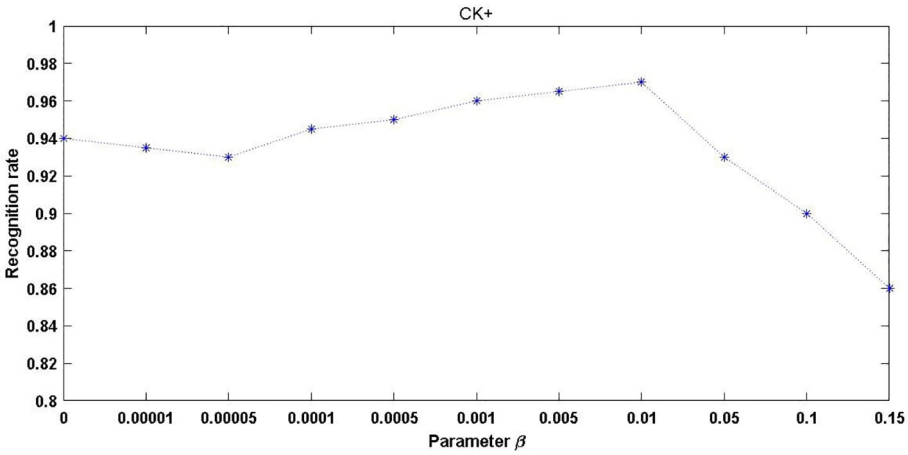

Fig. 4. Recognition accuracy of DDMTL with different parameter β.

some sample images in the MMI dataset. Since the MMI dataset only contains a small number of sample data, which is not sufficient to train a deep neural network, we supplemented additional data of the same expressions from the CK + dataset as experimental data. 8 groups of the MMI dataset are combined with the additional CK + dataset as a training set, and the other 2 groups of the MMI dataset are utilized as the evaluation and testing sets.

Our experiments are based on independent 10-fold cross-validation. We divided the MMI data set into 10 groups, 8 groups of which are utilized for training, and the other 2 groups of which are utilized for evaluation and testing, respectively. Experimental results are reported as the average of the 10 runs.

Table 3 shows the recognition accuracy of DDMTL and other reference methods. DDMTL is superior to the state-of-the-art. Compared to the experimental results on the CK + dataset, the recognition accuracy of the reference methods is significantly reduced, such as AlexNet and VGGNet, while the experimental results of DDMTL still maintain an acceptable recognition accuracy, which is more than 5% higher than the experimental results of those methods. DDMTL achieves better performance than the reference methods since it contains additional discriminative information of samples. This also proves the stability and robustness of DDMTL.

In addition, due to different training conditions, recognition accuracy can be different. For example, DeXpression [6] has reached a recognition accuracy of 98.36% on the MMI dataset, which is higher than the accuracy of our proposed method. It is because DeXpression extracts 20 video frames that represent the most video content for each video in the MMI database. However, in our experiments, only 3 frames near to the middle of a sequence are chosen as peak frames and correlated with their expression labels for experiments. Therefore, we found that facial expression recognition is a challenging task since the

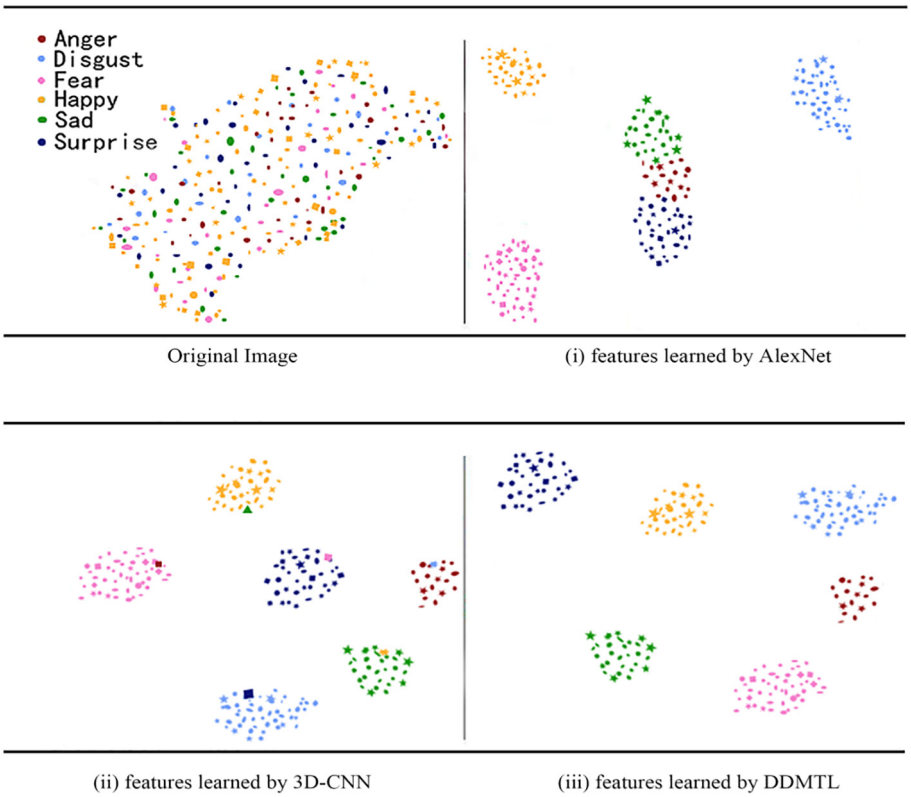

Fig. 5. A visualization of sample images and features learned by (i) AlexNet, (ii) 3D-CNN, and (iii) DDMTL on the CK + dataset. The dots, diamonds, and stars represent the training, validation, and testing data, respectively.

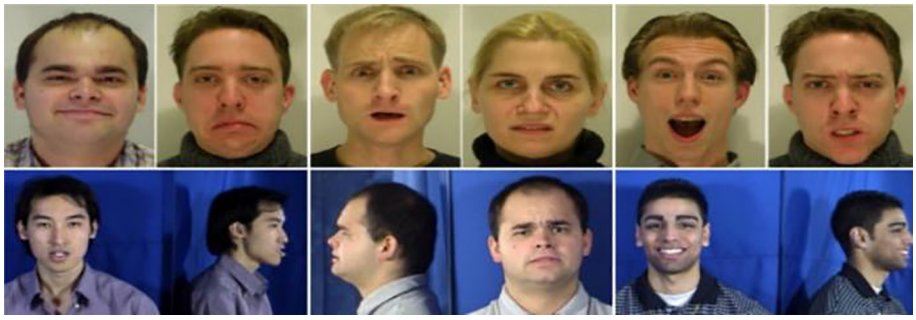

Fig. 6. Sample images in the MMI facial expression dataset.

Table 3
Recognition accuracy on the MMI dataset.

Method	Accuracy
AlexNet [2]	59.81%
VGGNet [35]	61.49%
GoogleNet [38]	62.16%
AdaGabor [50]	62.42%
3D-CNN [7]	64.23%
SJMT [28]	64.01%
DMTL [48]	64.92%
DDMTL	**67.78%**

recognition results are easily affected by the input images, where large inter-personal variations of the facial images can make the process difficult.

Table 4 shows the recognition accuracy of the provided 6 facial expressions on the MMI dataset. Compared with the experimental results on the CK + dataset, the recognition accuracy of all expressions decreased, especially the recognition accuracy of fear expression is only 39.4%, which shows that facial expression recognition is a challenging task.

In addition, to evaluate the ability of DDMTL to prevent overfitting problems, we plotted Fig. 7 to show changes in training losses as the number of iterations increases. Experimental results demonstrate that as the iteration progresses, the training loss decreases rapidly to approximately 0, which indicates that DDMTL can effectively prevent the overfitting problem based on the contrastive loss.

4.3. SFEW facial expression dataset

The experimental dataset is constructed by choosing static frames from the SFEW dataset that computed key frames using facial point clustering. The experimental data on the SFEW dataset is divided into 3 sets: a training set (958 samples), an evaluation set (436 samples), and a testing set (372 samples). Sample images include 7 expression categories: anger, disgust, fear, neutral, happy, surprise, and sad.

Table 5 shows the performance of DDMTL and the reference methods on the SFEW dataset. DDMTL yielded better experimental results than other reference methods. Since the environmental conditions of facial images vary greatly, the facial expression accuracy of all methods is reduced compared to the experimental results on the CK + dataset. However, both the validation accuracy and the test accuracy of DDMTL are at least 5% higher than those of other methods. The confusion matrix on the SFEW dataset is shown in Table 6. DDMTL performs adequately in terms of anger, happiness, and neural, but performs relatively poor in recognizing disgust and fear. This phenomenon may be because disgust and fear are caused by the slight motion of critical facial areas, which are difficult to be captured and thus are difficult to be recognized.

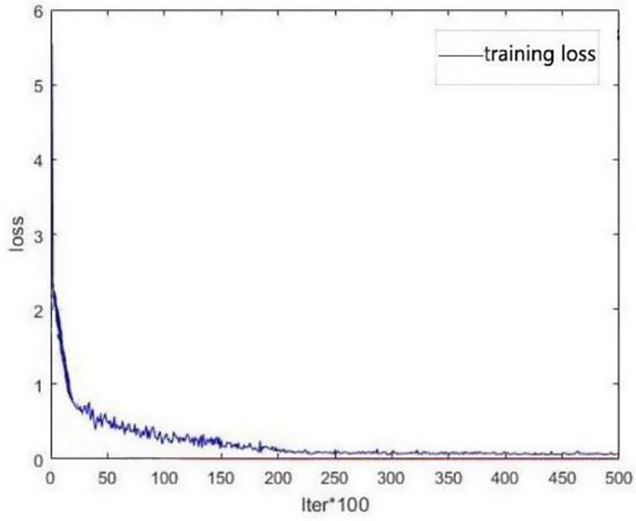

Fig. 7. Illustration of the training loss.

Table 4
Confusion matrix of facial expression obtained by DDMTL on the MMI dataset.

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**79.6**	9.98	5.32	1.21	18.32	7.22
Disgust	2.88	**69.3**	8.76	3.89	7.91	0
Fear	2.91	3.21	**39.4**	0	0	13.69
Happy	1.63	4.90	6.92	**90.7**	1.91	1.05
Sad	11.72	8.91	7.32	3.05	**71.5**	0
Surprise	1.26	3.7	32.28	1.15	0.36	**78.04**

Table 5
Recognition accuracy on the SFEW face expression dataset.

Method	Validation Accuracy	Test Accuracy
AlexNet [2]	47.82 ± 1%	50.32 ± 2%
VGGNet [35]	47.82 ± 3%	50.21 ± 1%
GoogleNet [38]	49.13 ± 2%	52.11 ± 3%
AdaGabor [50]	48.01 ± 2%	51.88 ± 2%
3D-CNN [7]	49.19 ± 3%	52.90 ± 3%
SJMT [28]	49.04 ± 1%	52.88 ± 2%
DMTL [48]	49.22 ± 2%	52.30 ± 2%
DDMTL	**51.21** ± 2%	**54.05** ± 2%

Table 6
Confusion matrix of facial expression obtained by DDMTL on the SFEW dataset.

	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	**78.23**	36.34	33.32	7.9	18.21	22.9	23.89
Disgust	0	**8.05**	0	0	0	0	0
Fear	1.83	0	**9.67**	0	1.84	7.4	2.47
Happy	0	22.18	13.05	**76.3**	3.62	20	8.06
Neutral	8.91	10.55	9.78	4.67	**61.9**	5.89	10.21
Sad	1.25	21.91	7.28	7.35	9.89	**35.8**	7.93
Surprise	9.78	0.97	26.9	3.78	4.54	8.01	**47.44**

5. Conclusion

In this paper, we propose the Discriminative Deep Multi-Task Learning (DDMTL) method for facial expression recognition. DDMTL can handle the loss of local information on spatial distribution by utilising both the information of class labels and the local information of sample spatial distribution simultaneously. Furthermore, DDMTL is based on a siamese network with shared weights designed to measure the local distribution via an adaptive reweighting module, to improve the accuracy of facial expression recognition. Experiments are performed on three widely used facial expression datasets. Experimental results determine that DDMTL performs better than the state-of-the-art methods. Compared with the reference methods, DDMTL is competitive especially when the training samples are small with respect to the local information. Nonetheless, when dealing with combined facial expression datasets (i.e., the training set are combined by several facial expression datasets), the recognition accuracies of the reference methods and DDMTL have all showed slightly decreased although DDMTL still achieved the best performance. How to maintain recognition accuracy when dealing with combined facial expression datasets will be investigated in the future.

CRediT authorship contribution statement

Hao Zheng: Writing - original draft, Invesitgation, Data curation, Funding acquisition. **Ruili Wang:** Conceptulation, Methodology, Supervision, Investigation, Funding acquisition, Writing - review & editing. **Wanting Ji:** Investigation, Writing - review & editing, Validation, Resource, Project adminstration. **Ming Zong:** Resource, Software, Visualization. **Wai Keung Wong:** Writing - original draft, Investigation. **Zhihui Lai:** Writing - original draft, Investigation. **Hexin Lv:** Writing - original draft, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported in part by the Natural Science Foundation of China (No. 61976118, 61806098, 61976145, and 61732011), the Natural Science Foundation of Zhejiang Province (LZ15F020001), the Jiangsu Province Six Talents Peak Project (RJFW-038), the Marsden Fund in New Zealand, and the China Scholarship Council.

References

[1] M. Alam, L.S. Vidyaratne, K.M. Iftekharuddin, Sparse simultaneous recurrent deep learning for robust facial expression recognition, IEEE Trans. Neural Networks Learn. Syst. 29 (10) (2018) 4905–4916.
[2] K. Alex, S. Ilya, and E. Geoffrey, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, pp. 1097-1105. 2012.
[3] S. Berretti, B. Amor, M. Daoudi, A.D. Bimbo, 3D facial expression recognition using SIFT descriptors of automatically detected keypoints, Visual Comput. 27 (11) (2011) 1021–1036.
[4] L. Bertinetto, J. Valmadre, J. Henriques, A. Vedaldi, and H. Philip, "Fully-convolutional siamese networks for object tracking," In European Conference on Computer Vision, pp. 850-865. 2016.
[5] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. "Signature verification using a 'siamese' time delay neural network." In Advances in Neural Information Processing Systems, pp. 737-744. 1994.
[6] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki. "Dexpression: Deep convolutional neural network for expression recognition." arXiv preprint arXiv:1509.05371. 2015.
[7] Y. Byeon, K. Kwak, Facial expression recognition using 3d convolutional neural network, Int. J. Adv. Comput. Sci. Appl. 5 (12) (2014) 107–112.
[8] R. Caruana, Multitask learning, Mach. Learn. 28 (1) (1997) 41–75.
[9] J. Chen, Z. Chen, Z. Chi, H. Fu, Facial expression recognition in video with multiple feature fusion, IEEE Trans. Affective Comput. 9 (01) (2018) 38–50.
[10] K. Chen, and A. Salman. "Extracting speaker-specific information with a regularized siamese deep network." In Advances in Neural Information Processing Systems, pp. 298-306. 2011.
[11] Z. Chen, R. Wang, Z. Zhang, H. Wang, L. Xu, Background foreground interaction for moving object detection in dynamic scenes, Inf. Sci. 483 (2019) 65–81.
[12] S. Chopra. "Learning a similarity metric discriminatively, with application to face verification." In IEEE Conference on Computer Vision and Pattern Recognition, pp. 539-546. 2005.
[13] M. Dahmane and J. Meunier. "Emotion recognition using dynamic grid-based HoG features." In Face and Gesture, pp. 884-888. 2011.
[14] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," In Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 2106-2112. 2011.
[15] P. Ekman, W. Friesen, Facial Action Coding System: Investigators Guide, Consulting Psychologists Press, 1978.
[16] N. Farajzadeh, M. Hashemzadeh, Exemplar-based facial expression recognition, Inf. Sci. 460–461 (2018) 318–330.
[17] H. Gao, Z. Liu, K. Weinberger, and L. Maaten, "Densely connected convolutional networks," arXiv preprint arXiv: 1608.06993. 2016.
[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587. 2014.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778. 2015.
[20] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." In Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324. 1998.
[21] S. Li, W. Deng, Reliable crowdsourcing and deep locality preserving learning for unconstrained facial expression recognition, IEEE Trans. Image Process. 28 (1) (2019) 356–370.
[22] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, M. Pietik-ainen, Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods, IEEE Trans. Affective Comput. 9 (4) (2017) 563–577.
[23] C. Liu, "Probabilistic Siamese Networks for Learning Representations," PhD diss., University of Toronto, 2013.
[24] A.T. Lopes, E. de Aguiar, A.F.D. Souza, T. Oliveira-Santos, Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order, Pattern Recogn. 61 (2017) 610–628.
[25] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression, IEEE Conf. Comput. Vision Pattern Recogn. Workshops (2010) 94–101.
[26] K. Lukaszr, G. Aidan, S. Noam, V. Ashish, P. Niki, J. Llion, and U. Jakob, "One model to learn them all," arXiv preprint arXiv: 1706.05137, 2017.
[27] H. Mobahi, R. Collobert, and J. Weston. "Deep learning from temporal coherence in video." In Proceedings of the 26th Annual International Conference on Machine Learning, pp. 737-744. 2009.
[28] G. Pons and D. Masip, "Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition," arXiv preprint arXiv: 1802.06664. 2018.
[29] R. Rajeev, M. Vishall, and C. Rama, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," arXiv preprint arXiv: 1603.01249, 2016.
[30] R. Ranjan, V.M. Patel, R. Chellappa, Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, IEEE Trans. Pattern Anal. Mach. Intell. 41 (1) (2017) 121–135.
[31] R. Rothe, R. Timofte, and L. V. Gool. "Dex: Deep expectation of apparent age from a single image." In Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 10-15. 2015.
[32] M. Sajjad, M. Nasir, F.U.M. Ullah, K. Muhammad, A.K. Sangaiah, S.W. Baik, Raspberry pi assisted facial expression recognition framework for smart security in law-enforcement services, Inf. Sci. 479 (2019) 416–431.
[33] A. Savran, H. Cao, A. Nenkova, R. Verma, Temporal Bayesian fusion for affect sensing: Combining video, audio, and lexical modalities, IEEE Trans. Cybern. 45 (9) (2014) 1927–1941.
[34] P. Shamsolmoali, J. Zhang, J. Yang, Image super resolution by dilated dense progressive network, Image Vis. Comput. 88 (2019) 9–18.
[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv: 1409.1556. 2014.
[36] B. Sun, L. Li, G. Zhou, X. Wu, J. He, L. Yu, D. Li, and Q. Wei, "Combining multimodal features within a fusion network for emotion recognition in the wild," in Proceedings of the 2015 ACM on International Conference on Multimedia Interaction, pp. 497-502. 2015.
[37] B. Sun, L. Li, G. Zhou, J. He, Facial expression recognition in the wild based on multimodal texture features, J. Electron. Imaging 25 (6) (2016) 061407.
[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: In IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[39] Y.W. Teh, M. Welling, S. Osindero, G.E. Hinton, Energy-based models for sparse overcomplete representations, J. Mach. Learn. Res. 4 (12) (2003) 1235–1260.
[40] Y. Tian, X. Wang, J. Wu, R. Wang, B. Yang, Multi-scale hierarchical residual network for dense captioning, J. Artif. Intell. Res. 64 (2019) 181–196.
[41] M. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition to the MMI facial expression database, in: In Proceedings of the 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research Emotion and Affect, 2010, p. 65.
[42] R. Wang and M. Zong, "Joint self-representation and subspace learning for unsupervised feature selection," World Wide Web, pp. 1-14, 2018.
[43] Y. Yang and T. Hospedales, "Deep multi-task representation learning: A tensor factorisation approach," arXiv preprint arXiv: 1605.06391. 2016.
[45] Z. Yu, G. Liu, Q. Liu, J. Deng, Spatio-temporal convolutional features with nested lstm for facial expression recognition, Neurocomputing 317 (2018) 50–57.
[46] M. Zareapoor, P. Shamsolmoali, and J. Yang. "Learning depth super-resolution by using multi-scale convolutional neural network." Journal of Intelligent & Fuzzy Systems Preprint, pp. 1-11. 2019.
[47] K. Zhang, Y. Huang, Y. Du, L. Wang, Facial expression recognition based on deep evolutional spatial-temporal networks, IEEE Trans. Image Process. 26 (9) (2017) 4193–4203.

[48] R. Rajeev, M. Vishal, M.R. Patel, Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, IEEE Trans. Pattern Anal. Mach. Intell. 41 (1) (2019) 121–135.
[49] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. "Local Gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition." In IEEE International Conference on Computer Vision, pp. 786-791. 2005.
[50] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. Metaxas, "Learning active facial patches for expression analysis," In IEEE Conference on Computer Vision and Pattern Recognition, pp. 2562-2569. 2012.