Motion Attention Deep Transfer Network for Cross-Database Micro-Expression Recognition

Wanchuang Xia[1], Wenming Zheng[2](✉), Yuan Zong[2], and Xingxun Jiang[2]

[1] School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China
xiawanchuag@seu.edu.cn
[2] School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China
{wenming_zheng, xhzongyuan, jiangxingxun}@seu.edu.cn

Abstract Cross-database micro-expression recognition is a great challenging problem due to the short duration and low intensity of micro-expressions from different collection conditions. In this paper, we present a Motion Attention Deep Transfer Network (MADTN) that can focus on the most discriminative movement regions of the face and reduce the database bias. Specifically, we firstly combine the motion information and facial appearance information to obtain the discriminative representation by merging the optical flow fields between three key-frames (the onset frame, the middle frame, the offset frame)and the facial appearance of the middle frame. Then, the deep network architecture extracts cross-domain feature with the superiority of the maximum mean discrepancy(MMD) loss so that the source and target domains have a similar distribution. Results on benchmark cross-database micro-expression experiments demonstrate that the MADTN achieves remarkable performance in many micro-expression transfer tasks and exceed the state-of-the-art results, which show the robustness and superiority of our approach.

Keywords: Micro-expression recognition · Deep learning · Optical flow · Transfer learning.

1 Introduction

Micro-expressions can reveal true information in social life, which are unconscious and spontaneous facial movements during the time a person emerge emotion but intentionally or involuntarily tries to hide genuine emotion [8]. Therefore, micro-expression recognition has great value in different fields, including lie detection [34], clinical diagnosis [35], business negotiation [36]. This has attracted increasing researchers to analyze micro-expression. Nevertheless, compared to ordinary facial expressions, the duration of a micro-expression is usually very short which is between one twenty-fifth to one half of a second [7]. Moreover, the muscle movements of micro-expressions also have locality and low intensity characteristics [34]. So micro-expressions recognition lack discriminative feature

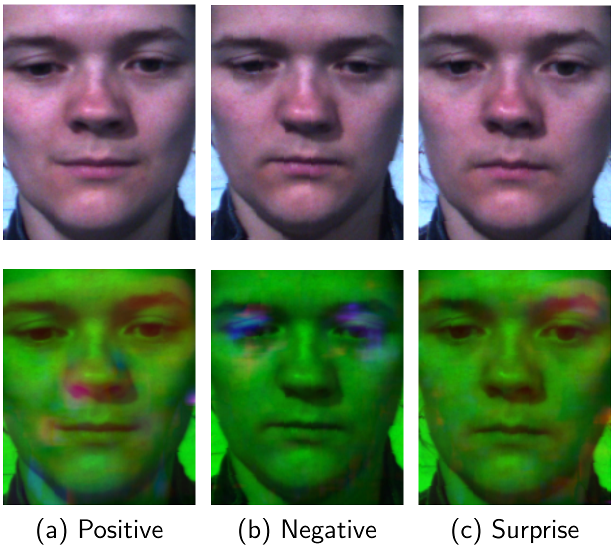


Figure 1. Examples of the middle frame and the motion attention representation in the SMIC-HS database. The first row are the middle frames of a micro-expression sequence.The second row are RGB images that R channel is the optical flow fields between the onset frame and the middle frame of video clip, G channel denote gray image of the middle frame, and B channel is the optical flow fields between the middle frame and the offset frame.

representations. The low intensity and short duration of micro-expressions make its recognition a very difficult task for people, even with professional training [11]. Thus, it is necessary to build an automatic micro-expression recognition system by using machine learning and computer vision techniques.

In recent years, many researchers have proposed a series of algorithms based on the characteristics of micro-expressions. In the work of [31], Park et al. proposed using the Eulerian Motion Magnification (EMM) [38] to exaggerate subtle change in the micro-expression video. Temporal interpolation method (TIM) [43] and Sparsity-Promoting Dynamic Mode Decomposition (DMDSP) [18] was employed to solve asymmetrical length of micro-expression video in [24], [22], [23]. Due to local binary pattern on three orthogonal planes (LBP-TOP)[42] could encode spatio-temporal variations, Pfister et al. chosen it to extract micro-expression representation [33]. Subsequently, Wang et al. [37] proposed local Binary Pattern with Six Interception Points (LBP-SIP) by reducing redundant information in LBP-TOP. Later on, lots of spatio-temporal descriptors were employed in micro-expression recognition, such as Spatio-temporal Completed Local Quantized Patterns (STCLQP) [16], Spatio-temporal LBP with Integration Projection (STLBP-IP) [15], Histogram of Oriented Gradient-TOP (HOG-TOP) [24]. Optical Flow (OF) [14] can readily portray motion information displayed in the micro-expression video. Several works were proposed based on OF, for example, Main Directional Mean Optical Flow (MDMO) [27], Facial Dynamics Map

(FDM) [39], Bi-Weighted Oriented Optical Flow (Bi-WOOF) [26]. Meanwhile, with deep learning widely prevailing in the visual task, it is obstacle for micro-expressions recognition by deep learning that lack large spontaneous micro-expressions database. Current popular spontaneous micro-expressions datasets are small, such as SMIC[25], CASME II[40]. Due to characteristics of micro-expressions, it is very hard to collect many spontaneous micro-expressions sample. Some works [20,32,19] explored to utilize neural networks in micro-expression recognition task.

Nevertheless, the above work mainly assumes that training(source) set and test(target) set satisfy the same distribution. This assumption is hard to conform with practical situations because samples recorded with different equipment under diverse backgrounds, illumination, angle. Without taking this into account, model trained on the source domain may fail to generalize well to the sample in the target domain. To alleviate this problem, transfer learning leverages the source domain with label information, and transfers the knowledge of the source domain to the unlabeled target domain [6]. Therefore, Zong et al. firstly investigate cross-database micro-expression recognition in [44,46,45] to alleviate distribution shift across domains.

In order to construct more discriminative and robustness features, we propose Motion Attention Deep Transfer Network (MADTN) for cross-database micro-expression recognition in this paper. Intuitively, people perceive micro-expressions by observing facial muscle movements in a video instead of only an image. Inspired by the intuition, MADTN perceives facial movement and pays attention mainly to the variational facial regions. Fig.1 illustrates the synthetic image of optical flow fields and facial appearance information. Optical Flow can obviously display the variational regions of the face. As can be shown in Fig.1.(a), movements occur in the corners of the mouth and the tip of the nose. Thus MADTN is able to focus on the discriminative regions of facial image. Firstly, we estimate facial deformations between the middle frame and the onset or offset frame in a micro-expression sequence. Then, the motion information is weighted to different regions of the face by Convolution Neural Network (CNN). Finally, we are able to reduce the feature distribution gap between domains by inserting the MMD loss into CNN. In this paper, our main contributions are summarized as follows:

- We propose Motion Attention Deep Transfer Network (MADTN) to conduct cross-database micro-expression recognition. MADTN can perceive the motion regions of the face and reduce the distribution shift between source and target domain.
- Visualized results show that the optical flow algorithm is effective in depicting facial muscle movement. With the integration of optical flow fields and facial appearance information, it can generate a discriminative feature representation, see Fig.1. We only select three frames from the video clips to cut down redundant information in sequence, and optical flow with them display larger degrees of motion information that contribute to the representation more discriminating.

- Experiment results demonstrate the superiority and robustness of the proposed MADTN over other state-of-the-art methods on two benchmark tasks.

2 RELATED WORK

Motion information can effectively improve the performance of micro-expression recognition. Several approaches were proposed based on Optical Flow(OF), which can readily portray motion information. MDMO [27] calculate the main direction of OF in each region of interest (ROI), including local statistics and spatial location information. FDM [39] extracts the motion information of micro-expression in a different granularity that iteratively calculates the principal OF direction of the local facial dynamic. BI-WOOF [26] was weight the Histogram of the Oriented Optical Flow (HOOF) [3] by multiplying with OF magnitude and optical strain magnitude of each ROI. These approaches utilize OF based the ROI level, thus it is important to choose appropriate ROI. However, improper ROI which include different motion direction may damage motion information that only considers the single direction of OF in ROI. In addition, these methods only consider motion information but neglect facial information. Combining motion information and facial information can accurately indicate the facial region where the movement occurs. In computer vision community, the attention mechanism[17] is proposed to weight different ROI and highlight the representations of task-related location. Compare with existing attention model that added attention module in the network, our approach adopts optical flow to produce the attention maps. In this paper, we weight facial information by OF information at the pixel level to generate more discriminative representation.

Deep learning has been shown to be effective in extracting features but is fairly new to this community. Because the lack of micro-expression samples limits the development of deep learning on micro-expression recognition. Kim et al. [20] attempt to utilize CNN and Long Short-Term Memory (LSTM) encoding micro-expression sequence and the network was designed relatively shallower. Peng et al. adapt micro-expression video clips and its OF information to train a 3D-CNN model that named Dual Temporal Scale Convolutional Neural Network(DTSCNN) [32] Enriched Long-term Recurrent Convolutional Network (ELRCN) [19] stack video frame, OF and optical strain with adjacent frames based channel level and feature level. Inspired from the above idea, this paper utilizes CNN to learning discriminative representation that combines facial image and OF information. A larger pixel's movement could contribute to more discriminative representation, thus we calculate OF between the onset or offset frame and the middle frame. The duration of a micro-expression video is usually very short which less than one half of a second, so the peak of micro-expression are more easily captured by high-speed cameras. On the one hand, OF in adjacent frame is too subtle to discriminative. On the other hand, it may not be robust enough to noise so that not accurately depict facial muscle movements.

To satisfy the practice application, cross-database micro-expression recognition is worthy to investigate that mitigate the domain shift between source data

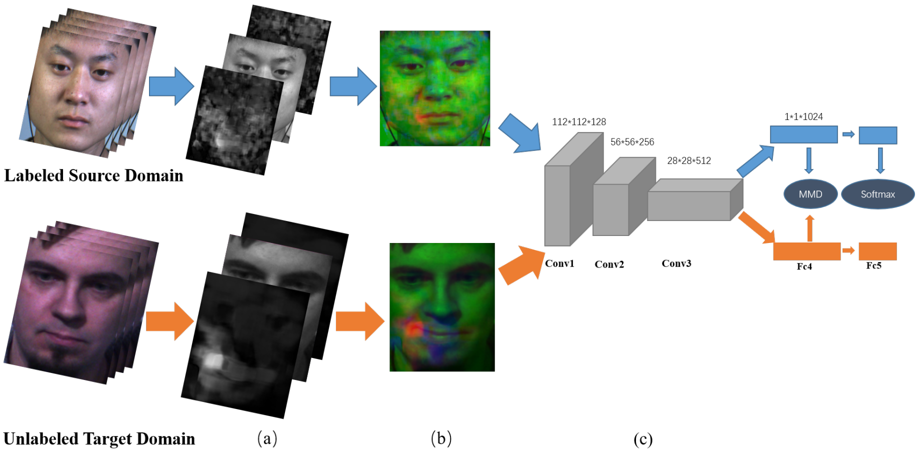


Figure 2. Overview of the proposed Motion Attention Deep Transfer Network (MADTN) for unsupervised cross-domain micro-expression recognition. We select three frames(the onset frame, the middle frame, the offset frame) from a micro-expression sequence to produce more discriminative representation. Our method leverages the representation by backpropagating the MMD loss between features in addition to cross entropy loss. (a) two optical flow fields between the onset or offset frame and the middle frame, the gray image of the middle facial frame. (b) a synthesis RGB image consist of three gray images in (a). (c) Deep Transfer Network. The blue and orange arrows denote the source domain and target domain, respectively.

and target data. Zong et al. [44] proposed Target Sample Re-Generator (TSRG) to regenerate samples that have the same or similar distribution. In the work of [45], Zong et al. attempt to bridge the feature distribution shift by proposing the auxiliary set selection model(ASSM) and transductive transfer regression model (TTRM). MMD can measure the feature distribution distances, thus we utilize MMD to minimizing the distribution distance between the source domain and target domain.

3 METHODOLOGY

In this section, we introduce our proposed Motion Attention Deep Transfer Network(MADTN) for cross-database micro-expression recognition, which utilizes CNN to learn the discriminative representation that combines motion information and facial appearance information. Due to the discrepancy between databases, we embed MMD in the deep convolution network to learning domains-invariant features.

3.1 Motion Attention Representation

Optical Flow could effectively encode the spatio-temporal displacement in the micro-expression video. In this paper, we detect facial location by MTCNN [41] in the onset frame of a micro-expression video. In order to better preserve the micro-expression related information, we crop face region from the onset frame which up to the top of the forehead, down to the bottom of the chin. According to the common facial bounding box in the onset frame, we crop a facial image from other frames of the video so that not hamper the motion information. On the one hand, micro-expressions have a very short duration, which makes sure that the faces of other frames exist in the relative facial position of the onset frames. On the other hand, if every frame was detected facial location, different facial bounding box will result in face displacement. So optical flow does not represent facial muscle movement, it would have a negative effect on optical flow approximation. The optical flow estimation algorithm [9] infers the motion of an object by tracking the displacement of mass points in a sequence. Given two frames in a video clip, the corresponding points on them satisfy the following equation:

$$I(x,y,t) = I(x+\mathit{\Delta} x, y+\mathit{\Delta} y, t+\mathit{\Delta} t) \tag{1}$$

where $I(x,y,t)$ is intensity at pixel point (x,y) in frame at time t, after $\mathit{\Delta} t$ time, the piont move $(\mathit{\Delta} x, \mathit{\Delta} y)$ that exist in another frame. According to Eq.(1), we could define the optical flow constraint equation:

$$I_x V_x + I_y V_y + I_t = 0 \tag{2}$$

where I_x, I_y, I_t are the partial derivatives of the intensity function. V_x and V_y are the horizontal and vertical components of the optical flow, which are defined as follow:

$$V = \left[V_x = \frac{\mathrm{d}x}{\mathrm{d}t}, V_y = \frac{\mathrm{d}y}{\mathrm{d}t} \right]^T \tag{3}$$

Due to obviously discrepancy exists between variations of the onset to the middle and the middle to the offset in a micro-expression sequence, we calculate optical flow between the onset frame and the middle frame, as well as between the middle frame and the offset frame. As can be seen in Fig. 3.(b), movement appears in the left corner of the mouth. The last row is the RGB image in Fig. 3 by concatenating the onset-middle optical flow image, the middle frame and the middle-offset optical flow image. We could observe the variations region of the face with the naked eye in the image. After such processing, we can obtain more discriminative features. As has been mentioned before, we treat a video clip as an image so that reduces redundant information in sequence.

3.2 Deep Transfer Network

Due to the powerful feature extraction capability of Convolutional Neural Network (CNN), we designed MADTN with three convolutional layers and two fully connected layers, see Fig. 2. Max pooling operation is used to reduce dimensionality after every convolutional layer. And after each layer, Leaky Rectified Linear Unit (LReLU) [29] is adopted to increase nonlinearity of model so that improve the fitting ability of the network.

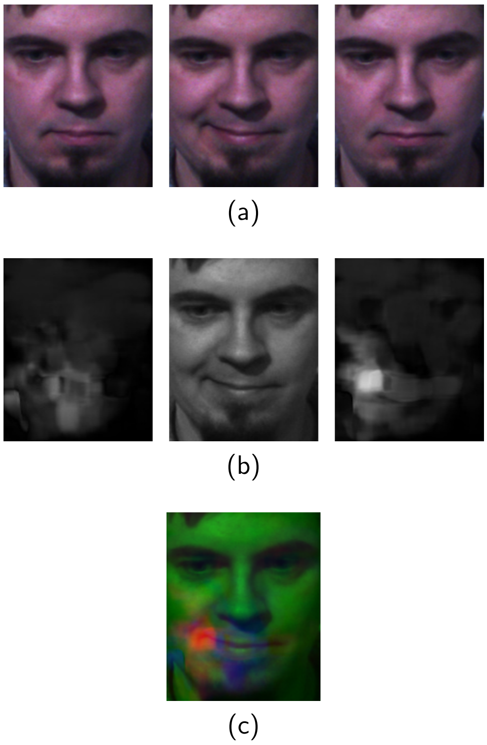


Figure 3. Extracting motion attention representation from a video sequence sample. From left to right, (a) are the onset frame, the middle frame, the offset frame in a video sequence, respectively. (b) are the optical flow fields between the onset frame and the middle frame, the gray image of the middle image, the optical flow fields between the middle frame and the offset frame. (c) is an RGB image that synthesis from three gray images in (b).

Directly utilizing this model trained on source domain to test samples in target domain usually results in poor performance because of database discrepancy. MMD [1] can measure the discrepancy between two domain by calculating the distance based on probability distributions in the reproducing kernel Hilbert space(RKHS). Therefore, we embedded MMD in the first fully connected layer of the network so as to produce domain-invariant features. The MMD can be defined as:

$$\mathrm{MMD}\left[D_s, D_t, \mathcal{F}\right] := \sup_{f \in \mathcal{F}} (\mathbf{E}_{D_s}\left[f(x^s)\right] - \mathbf{E}_{D_t}\left[f(x^t)\right]) \tag{4}$$

where $\mathbf{E}_{D_s}$ and $\mathbf{E}_{D_t}$ denote the expectations of source domain D_s and target domain D_t, respectively. If their distribution is similar, $\mathrm{MMD}\left[D_s, D_t\right]$ would close to zero. To satisfy practice calculation, MMD also is expressed as:

$$\mathrm{MMD}\left[D_s, D_t, \mathcal{H}\right] = \| \frac{1}{N_s} \sum_{i=1}^{N_s} \kappa(d_i^s) - \frac{1}{N_t} \sum_{i=1}^{N_t} \kappa(d_i^t) \|_{\mathcal{H}} \tag{5}$$

Because RKHS is often a high-dimensional or even infinite-dimensional space, the corresponding kernel chooses the Gaussian kernel,

$$\kappa(d^s, d^t) = \exp(-\frac{\|d^s - d^t\|^2}{2\delta^2}) \tag{6}$$

According to the above assumptions, the unbiased estimator of $\mathrm{MMD}^2\,[D_s, D_t, \mathcal{H}]$ was proposed:

$$\begin{aligned}\mathrm{MMD}^2\,[D_s, D_t, \mathcal{H}] =& \frac{1}{N_s(N_s-1)} \sum_{i \neq j}^{N_s} \kappa(d_i^s, d_j^s) \\ &+ \frac{1}{N_t(N_t-1)} \sum_{i \neq j}^{N_t} \kappa(d_i^t, d_j^t) \\ &- \frac{2}{N_s N_t} \sum_{i,j=1}^{N_s, N_t} \kappa(d_i^s, d_j^t)\end{aligned} \tag{7}$$

By adding the MMD loss into the network, the total loss function of MADTN becomes as follows:

$$L_{\mathrm{all}} = L_s(\varPhi(x^s), y^s) + \lambda \mathrm{MMD}^2\,[D_s, D_t, \mathcal{H}] \tag{8}$$

where $L_s(\varPhi(x^s), y^s)$ denotes cross-entropy loss function of the source domain, and λ is hyper-parameter to trade off these two loss functions. Through joint training the loss functions, it not only can learn discriminative representation about micro-expression, but also can reduce the difference between the source and the target domains. The experiment details are presented in the next section.

Table 1. Results(mean F1-score/accuracy) based on the TYPE-I experiments, which a series of transfer tasks between three subsets of the SMIC database. For short, H = SMIC-HS, V = SMIC-VIS, N = SMIC-NIR. The bold elements correspond to the best results

Method	$H \rightarrow V$	$V \rightarrow H$	$H \rightarrow N$	$N \rightarrow H$	$V \rightarrow N$	$N \rightarrow V$	*Average*
Baseline[2]	0.8002/80.28	0.5421/54.27	0.5455/53.52	0.4878/54.88	0.6186/63.38	0.6078/63.38	0.6003/61.62
IW-SVM[13]	0.8868/88.73	0.5852/58.54	0.7469/74.65	0.5427/54.27	0.6620/69.01	0.7228/73.24	0.6911/68.07
TCA[30]	0.8269/83.10	0.5477/54.88	0.5828/59.15	0.5443/57.32	0.5810/61.97	0.6598/67.61	0.6238/64.01
GFK[12]	0.8448/84.51	0.5957/59.15	0.6977/70.42	0.6197/**62.80**	0.7619/76.06	0.8142/81.69	0.7223/72.44
SA[10]	0.8037/80.28	0.5955/59.15	0.7465/74.65	0.5644/56.10	0.7004/71.83	0.7394/74.65	0.6917/69.44
STM[4,5]	0.8253/83.10	0.5059/51.22	0.6628/66.20	0.5351/56.10	0.6427/67.61	0.6922/70.42	0.6440/65.78
TKL[28]	0.7742/77.46	0.5738/57.32	0.7051/70.42	0.6116/62.60	0.7558/76.06	0.7579/76.06	0.6964/69.92
TSRG[44]	**0.8869/88.73**	0.5652/56.71	0.6484/64.79	0.5770/57.93	0.7056/70.42	0.8116/81.69	0.6991/70.05
DRFS-T[47]	0.8643/85.92	0.5767/57.32	0.7179/71.83	0.6163/61.59	0.7286/73.24	0.7732/77.46	0.7128/71.23
DRLS[47]	0.8604/85.92	0.6120/60.98	0.6599/66.20	0.5599/55.49	0.6620/69.01	0.5771/61.97	0.6552/66.60
RSTR[46]	0.8721/87.32	0.6401/64.02	0.7466/74.65	0.5765/57.32	0.7506/76.06	0.8428/84.51	0.7381/73.98
MADTN(ours)	0.8302/83.11	**0.6704/66.46**	**0.7641/77.44**	**0.6252**/62.21	**0.8435/84.52**	**0.8732/87.30**	**0.7678/76.84**

Table 2. Results(mean F1-score/accuracy) based on the TYPE-II experiments, which a series of transfer tasks between the CASME II database and one subsets of the SMIC(HS, VIS, NIR) database. For short, C = CASME II, H = SMIC-HS, V = SMIC-VIS, N = SMIC-NIR. The bold elements correspond to the best results

Method	$C \rightarrow H$	$H \rightarrow C$	$C \rightarrow V$	$V \rightarrow C$	$C \rightarrow N$	$N \rightarrow C$	Average
Baseline[2]	0.3697/45.12	0.3245/48.46	0.4701/50.70	0.5367/53.08	0.5295/52.11	0.2368/23.85	0.4112/45.55
IW-SVM[13]	0.3541/41.46	0.5829/62.31	0.5778/59.15	0.5537/54.62	0.5117/50.70	0.3456/36.15	0.4876/50.73
TCA[30]	0.4637/46.34	0.4870/53.08	0.6834/69.01	0.5789/59.23	0.4992/50.70	0.3937/42.31	0.5177/53.45
GFK[12]	0.4126/46.95	0.4776/50.77	0.6361/66.20	0.6056/61.50	0.5180/53.52	0.4469/46.92	0.5161/54.31
SA[10]	0.4302/47.56	0.5447/62.31	0.5939/59.15	0.5243/51.54	0.4738/47.89	0.3592/36.92	0.4877/50.90
STM[4,5]	0.3604/43.90	0.6115/63.85	0.4015/52.11	0.2715/30.00	0.3523/42.25	0.3850/41.54	0.3982/45.61
TKL[28]	0.3829/44.51	0.4661/54.62	0.6042/60.56	0.5378/53.08	0.5392/54.93	0.4248/43.85	0.4925/51.93
TSRG[44]	0.5042/51.83	0.5171/60.77	0.5935/59.15	0.6208/63.08	0.5624/56.34	0.4105/46.15	0.5348/56.22
DRFS-T[47]	0.4524/46.95	0.5460/60.00	0.6217/63.38	0.6762/68.46	0.5369/56.34	0.4653/50.77	0.5498/57.65
DRLS[47]	0.4924/53.05	0.5267/59.23	0.5757/57.75	0.5942/60.00	0.4885/49.83	0.3838/42.37	0.5102/53.71
RSTR[46]	0.5297/54.27	0.5622/60.77	0.5882/59.15	0.7021/70.77	0.5009/50.70	0.4693/50.77	0.5587/57.74
MADTN(ours)	**0.6100/62.79**	**0.7486/77.54**	**0.7056/70.41**	**0.7304/72.85**	**0.7305/73.24**	**0.7403/75.98**	**0.7109/72.14**

4 EXPERIMENTS

We evaluate the proposed Motion Attention Deep Transfer Network by cross-domain micro-expression recognition tasks on SMIC [25] , CASME II [40]. And we keep the same experiment protocols with [46] to guarantee fair comparison.

4.1 Databases

SMIC[25] has three subsets SMIC-HS, SMIC-VIS and SMIC-NIR, which collected from three distinct cameras: a high-speed camera with 100 fps, a normal visual camera with 25 fps, a near-infrared camera with 25 fps, respectively. All samples were divided into three categories which are Positive, Negative and Surprise. SMIC-HS contains 164 samples from 16 subjects, while SMIC-VIS and SMIC-NIR have 71 samples belonging to the last eight subjects from all subjects.

CASME II[40] consists of 257 samples video sequences of 26 subjects belonging to seven classes: Happy, Disgust, Repression, Sad, Fear, Surprise and Others. The frame rate of all sample videos is up to 200 fps. In cross-database classification task, the different databases should have same labels. According to the definition of the label in SMIC, we relabel the samples from CASME II, Happy samples are relabeled to the Positive(32 samples), Disgust, Sad and Fear are given the Negative(73 samples), and the Surprise(25 samples) stays unchanged.

4.2 Implementation details

To produce discriminative representations, we synthesize a RGB image by utilizing optical flow fields and appearance information. Optical flow fields have the horizontal components V_x and the vertical components V_y which expressed in the Cartesian coordinate system. Optical Flow indicates the direction and intensity of frame pixel movement, so we transform the Cartesian coordinate

$V = (V_x, V_y)$ into the Polar coordinate $V = (r, \theta)$, where r and θ are the amplitude and oriention of the optical flow, respectively. In the HSV representation model, Hue H typically measured in degrees $[0°, 360°]$, as well as Saturation S and Value V measured on the range $[0, 1]$. In order to form an image, we set the value of V to 1 and assigned r, θ to S, H, respectively. Then we convert the image to a grayscale image, see Fig. 3.(b).

In order to avoid overfitting problems, we select the 20% frames in the middle of a training(source) set video clip as the middle frame but the test(target) set only selected the most middle frame in a video clip. This increases the diversity of sample due to the different degrees of facial muscle movement. In addition, each sample randomly rotated between the angles $[-30°, 30°]$. And our network was designed relatively shallow. All images were resized to 112×112 pixels before inputting to the network, as well as in [46]. We optimize our model by Adam[21] solver with learning rate 2×10^{-4}. The batch size is set to 32 for each domain. We set the MMD penalty parameter λ to equal 2 in every experiment. Zong et. al[46] establish a benchmark cross-database micro-expression recognition(CDMER) experimental evaluation protocol, which contains two kinds of CDMER tasks: TYPE-I, TYPE-II. TYPE-I denote experiments between three subset of SMIC(SMIC-HS(H), SMIC-VIS(V), SMIC-NIR(N)), i.e., $H \to V$, $V \to H$, $H \to N$, $N \to H$, $V \to N$, $N \to V$. TYPE-II indicate experiments between the selected CASME II(C) and SMIC including $C \to H$, $H \to C$, $C \to V$, $V \to C$, $C \to N$, $N \to C$. Experiments are measured using mean F1-score and Accuracy. Mean F1-Score is the F1-score of each class divided by the number of classes without consideration of every class size, which provides a reasonable metric in the class imbalanced data.

4.3 Results

The results on TYPE-I, TYPE-II experiments and comparisons with other methods are reported in Table 1, Table 2. To fair comparison, the results of other methods directly reported from [46] on TYPE-I, TYPE-II. Our proposed MADTN model has state-of-the-art overall performance than all the comparison methods in average mean F1-score and average accuracy. Especially in the TYPE-II experiment, our method is superior to the highest average mean F1-score/accuracy of [46] by 0.1522/14.4%. Furthermore, MADTN substantially outperforms the comparison methods on most of the experiments, and with larger rooms of improvement. In addition, some comparison methods outperform MADTN at $H \to V$. SMIC-HS and SMIC-VIS are very similar that samples was collected from same environment, and the frame rate is the main difference between them. Hence other methods can achieve a relatively good result at $H \to V$. While motion information in some samples of SMIC-VIS is not obvious enough, probably because of the low frame rate, see Fig.4. It may not capture the peak and valley of micro-expression in normal-speed(24fps) camera. It make MADTN to have a relatively poor performance. Distinct motion information is very critical to train our approach for a remarkable performance.

Figure 4. Samples in the SMIC-VIS database

It is worth noting that three subsets of SMIC are very similar in many respects, their difference is recorded by different cameras. Nevertheless, SMIC and CASME II are much more different due to collected by different researchers. Hence, the TYPE-II task is more difficult than the TYPE-I. As shown in Table 1, Table 2, It can demonstrate that the result of all methods in the TYPE-II and TYPE-III task are much lower than those in the TYPE-I task. We notice that the performance of other method drop sharply from the TYPE-I task to the TYPE-II task, while the result of MADTN has only a small drop. These results demonstrate the strong robustness of our proposed method. This could be caused by enhancing feature more discriminating and alleviating distribution discrepancy.

4.4 Ablation Analysis

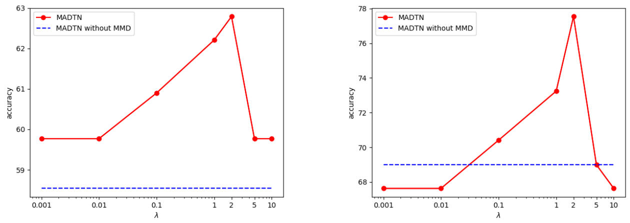
To look more deeply into our model, we conduct an extensive ablation experiment to study how components of MADTN affect performance. Firstly, we evaluate these variant ingredients of Motion Attention Feature on $C \rightarrow H$, $C \rightarrow V$ experiments following the same setting. In order to keep invariable network architecture, we set it to equal zero if an ingredient has not existed. The results are shown in Table 3, that the *Face* denotes a grayscale image of the middle frame in a video clip, the *onOF* represent optical flow fields between the onset frame and the middle frame, the *offOF* signify optical flow fields between the middle frame and the offset frame. We compared *Face* and *onOF-offOF* to verify benefit of motion representation. The promotions of *onOF-offOF* suggest that facial encoded representation has fallen behind in reflecting micro-expression compared with motion representation. Comparing the *onOF-Face* and the *onOF-Face-offOF*, transformation during micro-expression vanishing may benefits the performance. It may be because larger muscle movement occurs in the last half of the video clip. With the help of facial and motion representations, the *onOF-Face-offOF* achieves the best performance than others. This is because the optical flow enables the model to attention movement-related facial regions. Facial information can help to reduce the impact of motion bias on different faces.

Then, we validate the effectiveness of the MMD loss on an ablation experiment that eliminates the effect of the MMD loss. Specifically, we demonstrate particular cross-dataset results on $C \rightarrow H$ and $C \rightarrow V$ in terms of different

Table 3. Experimental result(mean F1-score/accuracy) of our method with different input features on $C \rightarrow H$, $C \rightarrow V$ and $H \rightarrow N$. The *Face* denotes facial appearance information of the middle frame, the *onOF* denotes optical flow fields between the onset frame and the middle frame, the *offOF* denotes optical flow fields between the middle frame and the offset frame

Input	$C \rightarrow H$	$C \rightarrow V$	$H \rightarrow N$
Face	0.4115/42.07	0.4715/52.10	0.4673/52.10
onOF-Face	0.5598/56.69	0.6207/61.96	0.5713/57.76
onOF-offOF	0.5858/59.13	0.6520/64.79	0.7352/73.24
onOF-Face-offOF	0.6100/62.79	0.7056/70.41	0.7641/77.44

value λ, see Fig.5. We can observe that the accuracy reach the maximum at $\lambda = 2$ and then fall off. Fig.5 display the promoting effect of the supervision of the cross-entropy loss and the MMD loss. It demonstrates that the MMD loss can help MADTN to alleviate the distribution shift between source and target domains and to enhance performances of our method.



(a) $C \rightarrow H$ (b) $C \rightarrow V$

Figure 5. Performances w.r.t λ on $C \rightarrow H$ and $C \rightarrow V$

5 Conclusion

In this paper, Motion Attention Deep Transfer Network (MADTN) has been presented to conduct unsupervised cross-database micro-expression recognition. We select three frames into micro-expression sequences such that can reduce redundant information, then combine their motion information and facial appearance information to pay more attention to facial regions that occur muscle movement. What's more, deep transfer network has proposed to bridge the distribution discrepancy between source and target domains which increase the robustness of our method. Experiments on two benchmark tasks show that the MADTN achieves remarkable performance in many transfer tasks and outperforms all other counterparts, demonstrating the robustness and superiority of our approach.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1305200, in part by the National Natural Science Foundation of China under Grant 61921004, Grant 61902064, Grant 81971282, Grant U2003207, and Grant 62076064, and in part by the Fundamental Research Funds for the Central Universities under Grant 2242018K3DN01.

References

1. Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B., Smola, A.J.: Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics **22**(14), e49–e57 (2006)
2. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST) **2**(3), 1–27 (2011)
3. Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1932–1939. IEEE (2009)
4. Chu, W.S., De la Torre, F., Cohn, J.F.: Selective transfer machine for personalized facial action unit detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3515–3522 (2013)
5. Chu, W.S., De la Torre, F., Cohn, J.F.: Selective transfer machine for personalized facial expression analysis. IEEE transactions on pattern analysis and machine intelligence **39**(3), 529–545 (2016)
6. Csurka, G.: Domain adaptation in computer vision applications, vol. 8. Springer (2017)
7. Ekman, P.: Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition). WW Norton & Company (2009)
8. Ekman, P., Friesen, W.V.: Nonverbal leakage and clues to deception. Psychiatry **32**(1), 88–106 (1969)
9. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Scandinavian conference on Image analysis. pp. 363–370. Springer (2003)
10. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Unsupervised visual domain adaptation using subspace alignment. In: Proceedings of the IEEE international conference on computer vision. pp. 2960–2967 (2013)
11. Frank, M., Herbasz, M., Sinuk, K., Keller, A., Nolan, C.: I see how you feel: Training laypeople and professionals to recognize fleeting emotions. In: The Annual Meeting of the International Communication Association. Sheraton New York, New York City (2009)
12. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2066–2073. IEEE (2012)
13. Hassan, A., Damper, R., Niranjan, M.: On acoustic emotion recognition: compensating for covariate shift. IEEE Transactions on Audio, Speech, and Language Processing **21**(7), 1458–1468 (2013)
14. Horn, B.K., Schunck, B.G.: Determining optical flow. In: Techniques and Applications of Image Understanding. vol. 281, pp. 319–331. International Society for Optics and Photonics (1981)

15. Huang, X., Wang, S.J., Zhao, G., Piteikainen, M.: Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 1–9 (2015)
16. Huang, X., Zhao, G., Hong, X., Zheng, W., Pietikäinen, M.: Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. Neurocomputing **175**, 564–578 (2016)
17. Itti, L., Koch, C.: Computational modelling of visual attention. Nature reviews neuroscience **2**(3), 194–203 (2001)
18. Jovanović, M.R., Schmid, P.J., Nichols, J.W.: Sparsity-promoting dynamic mode decomposition. Physics of Fluids **26**(2), 024103 (2014)
19. Khor, H.Q., See, J., Phan, R.C.W., Lin, W.: Enriched long-term recurrent convolutional network for facial micro-expression recognition. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 667–674. IEEE (2018)
20. Kim, D.H., Baddar, W.J., Ro, Y.M.: Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 382–386 (2016)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Le Ngo, A.C., Liong, S.T., See, J., Phan, R.C.W.: Are subtle expressions too sparse to recognize? In: 2015 IEEE International Conference on Digital Signal Processing (DSP). pp. 1246–1250. IEEE (2015)
23. Le Ngo, A.C., Oh, Y.H., Phan, R.C.W., See, J.: Eulerian emotion magnification for subtle expression recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1243–1247. IEEE (2016)
24. Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., Pietikäinen, M.: Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. IEEE Transactions on Affective Computing **9**(4), 563–577 (2017)
25. Li, X., Pfister, T., Huang, X., Zhao, G., Pietikäinen, M.: A spontaneous micro-expression database: Inducement, collection and baseline. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). pp. 1–6. IEEE (2013)
26. Liong, S.T., See, J., Wong, K., Phan, R.C.W.: Less is more: Micro-expression recognition from video using apex frame. Signal Processing: Image Communication **62**, 82–92 (2018)
27. Liu, Y.J., Zhang, J.K., Yan, W.J., Wang, S.J., Zhao, G., Fu, X.: A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Transactions on Affective Computing **7**(4), 299–310 (2015)
28. Long, M., Wang, J., Sun, J., Philip, S.Y.: Domain invariant transfer kernel learning. IEEE Transactions on Knowledge and Data Engineering **27**(6), 1519–1532 (2014)
29. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml. vol. 30, p. 3 (2013)
30. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks **22**(2), 199–210 (2010)
31. Park, S.Y., Lee, S.H., Ro, Y.M.: Subtle facial expression recognition using adaptive magnification of discriminative facial motion. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 911–914 (2015)

32. Peng, M., Wang, C., Chen, T., Liu, G., Fu, X.: Dual temporal scale convolutional neural network for micro-expression recognition. Frontiers in psychology **8**, 1745 (2017)
33. Pfister, T., Li, X., Zhao, G., Pietikäinen, M.: Recognising spontaneous facial micro-expressions. In: 2011 international conference on computer vision. pp. 1449–1456. IEEE (2011)
34. Porter, S., Ten Brinke, L.: Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. Psychological science **19**(5), 508–514 (2008)
35. Russell, T.A., Chu, E., Phillips, M.L.: A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool. British journal of clinical psychology **45**(4), 579–583 (2006)
36. Salter, F., Grammer, K., Rikowski, A.: Sex differences in negotiating with powerful males. Human Nature **16**(3), 306–321 (2005)
37. Wang, Y., See, J., Phan, R.C.W., Oh, Y.H.: Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. In: Asian conference on computer vision. pp. 525–537. Springer (2014)
38. Wu, H.Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., Freeman, W.: Eulerian video magnification for revealing subtle changes in the world. ACM transactions on graphics (TOG) **31**(4), 1–8 (2012)
39. Xu, F., Zhang, J., Wang, J.Z.: Microexpression identification and categorization using a facial dynamics map. IEEE Transactions on Affective Computing **8**(2), 254–267 (2017)
40. Yan, W.J., Li, X., Wang, S.J., Zhao, G., Liu, Y.J., Chen, Y.H., Fu, X.: Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. PloS one **9**(1) (2014)
41. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
42. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE transactions on pattern analysis and machine intelligence **29**(6), 915–928 (2007)
43. Zhou, Z., Zhao, G., Pietikäinen, M.: Towards a practical lipreading system. In: CVPR 2011. pp. 137–144. IEEE (2011)
44. Zong, Y., Huang, X., Zheng, W., Cui, Z., Zhao, G.: Learning a target sample re-generator for cross-database micro-expression recognition. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 872–880 (2017)
45. Zong, Y., Zheng, W., Cui, Z., Zhao, G., Hu, B.: Toward bridging microexpressions from different domains. IEEE transactions on cybernetics (2019)
46. Zong, Y., Zheng, W., Hong, X., Tang, C., Cui, Z., Zhao, G.: Cross-database micro-expression recognition: A benchmark. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval. pp. 354–363 (2019)
47. Zong, Y., Zheng, W., Huang, X., Shi, J., Cui, Z., Zhao, G.: Domain regeneration for cross-database micro-expression recognition. IEEE Transactions on Image Processing **27**(5), 2484–2498 (2018)