Neurocomputing 333 (2019) 231–239

Contents lists available at ScienceDirect

Neurocomputing

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

Cross-domain facial expression recognition via an intra-category common feature and inter-category Distinction feature fusion network

Yanli Ji, Yuhan Hu, Yang Yang, Fumin Shen, Heng Tao Shen*

Center for Future Media, School of Computer Science Engineering, University of Electronic Science and Technology of China, Chengdu, China

ARTICLE INFO

Article history:
Received 4 January 2018
Revised 29 November 2018
Accepted 3 December 2018
Available online 30 December 2018

Communicated by Dr. Qingshan Liu

Keywords:
Facial expression recognition
Cross-domain recognition
Deep learning
ICID fusion network

ABSTRACT

Facial expression recognition is crucial for various human-robot interaction applications, which requires facial expression analysis having a broad generalization. However, existing researches focus on the recognition in databases containing a limited number of samples. In this paper, we propose a novel feature fusion network for facial expression recognition in a cross-domain manner in order to realize the facial expression recognition in extensive scenarios. The proposed network consists of an Intra-category Common feature representation (IC) channel and an Inter-category Distinction feature representation (ID) channel for facial expression representation, and finally combine learned features of the two channels for facial expression recognition in cross databases. The IC channel learns the common features of intra-category facial expressions, and the ID channel learns the characteristic features of different categories. We evaluate the proposed approach in various experiment settings for cross-domain recognition, and achieves the state-of-the-art performances. We also evaluate the proposed approach for expression recognition in single databases, and also obtains the outstanding performance in the CK+, MMI, SFEW and RAF databases.

1. Introduction

With the development of robot technology, the human-robot interaction plays a crucial role in applications of service robots. For a natural and fluency interaction between robots and the human, emotion recognition and interaction become more important. Facial expression is the major route to express emotion feelings for both of the human and robots. Therefore, facial expression recognition is a basic but important research topic for various HRI applications. In the social interactions, the human is able to perceive emotions of any person in various conditions. Extending to the HRI, approaches of facial expression recognition should have a broad generalization.

Many research endeavors have been dedicated to solving the facial expression recognition. Most of the methods for facial expression recognition are evaluated in single databases [1,2]. However, frequently used databases always consist of the same property types of samples and a limited number of samples *e.g.*CK+ [3], MMI [4,5]. Obviously, it is far from the requirement of generalization in real applications. Furthermore, approaches of cross-domain recognition were presented where classification models were trained using samples in one database and tested samples in other databases [6–8]. To improve the effectiveness of cross-database recognition, approaches fusing facial expression features in multiple databases were also presented [9]. Meanwhile, cross-domain recognitions extent generality of facial expression analysis. However, it is still a challenging problem to design effective models for feature learning and category classification in a cross-domain manner because of the complexity and variation of various facial expressions. Wang et al. [10] presented a constrained asymmetric multi-task discriminant component analysis (cAMT-DCA) approach to solving cross-view person reidentification problems. The approach maximized the data discrepancy on the shared component in different scenarios, maximized the local inter-class variation and minimized local intra-class variation in all scenarios. Inspired by it, we tread facial expressions in different databases as expressions that captured in different scenarios, and propose a feature fusion network for cross-domain recognition.

Existing facial expression databases involve lab-setting database and database containing wild facial expression images collected from the internet. For instance, databases JAFFE [11], Cohn–Kanade [12], Cohn-Kanade+ [3], MMI [4,5], DISFA [13], multi-PIE [14], Bosphorous [15] and CASME [16] and CAS(ME)^2 [17] are captured in a lab-controlled surrounding, while databases of SFEW [18], EmotioNet [19] and RAF Database [20] are real-world facial expression images collected from the internet. Cross-domain recognition had been evaluated on the CK+ and the RAF database to explore the difference between expressions of lab-controlled

* Corresponding author.
E-mail addresses: yanliji@uestc.edu.cn (Y. Ji), shenhengtao@hotmail.com (H.T. Shen).

https://doi.org/10.1016/j.neucom.2018.12.037
0925-2312/

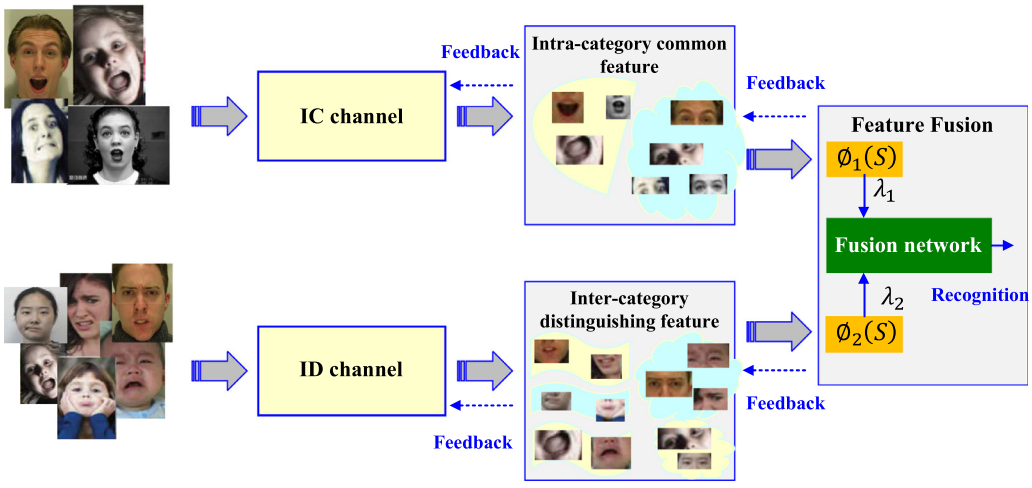


Fig. 1. Diagram of the proposed ICID Network. The IC channel learns common features of intra-category facial expressions for the common representation, and the ID channel learns characteristic features of different categories for the distinction representation of facial expressions. Finally, a fusion network combines two-channel features for facial expression recognition.

posed face and the real-world affective face [20]. Comparing with differences, it is even more important to extract common representations of expressions from various different data sets to extend the generalization of recognition. Various hand-craft features were designed to represent facial expressions in the earlier research stage [21–26], appearance features and geometry features. Appearance features include Gabor features [27,28], the Local Binary Pattern(LBP) [29], the Local Phase Quantization (LPQ) [30], 2D-DCT [31], and BoW Hist features [32]. Geometry features include landmark positions [33,34], HOG [35], SIFT [36] *etc.*. Moreover, deep networks are frequently used for facial expression recognition due to the excellent performance in feature learning [26,37–39]. To achieve a broad generalization, we design a deep network to learn common representations of expressions from multiple data sources, simultaneously, we emphasis representation differences of different categories to achieve a higher recognition result.

In this paper, we propose a novel feature fusion network to learn a common representation of expressions in different databases for facial expression recognition in a cross-domain manner. The proposed network consists of an Intra-category Common feature representation channel (IC) and an Inter-category Distinction feature representation channel (ID) for facial expression representation, and a fusion network combines two channel features for facial expression recognition in cross databases. The proposed approach is named as ICID fusion network. The diagram of our proposed approach is shown in the Fig. 1. The IC channel learns common features of intra-category facial expressions for the common representation, and the ID channel learns Distinction features of different categories for the representation of facial expressions. Such representation enhances the discrimination of facial expressions. It improves recognition effectiveness by maximizing the inter-class variation and minimizing intra-class variation in multiple databases. The proposed approach aims at recognizing facial expressions in extensive scenarios, and realizing the real-world application of facial expression recognition. In the end, we evaluate the proposed approach on multiple databases which were captured in lab surroundings or collected from the internet. We design a set of cross-domain validation experiments to certify the effectiveness and the generalization of our proposed approach.

2. Related works

In this section, we review existing approaches of the cross-domain facial expression recognition. According to the number of databases that used in training and test processing, existing approaches are separated to single-to-single cross-domain recognition and multiple-to-single cross-domain recognition. Here, single-to-single cross-domain recognition refers to training a model in one database and test in the other database, while multiple-to-single cross-domain recognition is that a model is trained using multiple databases and tested in another database.

2.1. Single-to-single cross-domain recognition

For facial expression recognition, manually designed features are still widely adopted. Tong et al. [6] employed the Dynamic Bayesian Network (DBN) to represent probabilistic relationships among different AUs and to modulate the temporal facial activities. AU relationships and AU dynamics were integrated to recognize AUs. They trained the recognition model using the CK database and performed the test in the MMI database for a cross-validation. Koelstra et al. [40] adopted dynamic texture to represent facial Action Units for AU recognition and temporal models determination. Then frame-based GentleBoost was combined with Hidden Markov Models for facial expression recognition. The approach was designed for single-database recognition, and it was evaluated in cross databases. Valstar et al. [6] used a Gabor-feature based facial point detector to localize facial fiducial points, and tracked these points in facial sequences to model temporal facial activations for facial expression recognition. Cross-validation was performed in the MMI and the CK database. Liu et al. [41] presented a novel Boosted Deep Belief Network (BDBN) to learn facial features which characterized expression-related facial appearance/shape variations. It trained a boosted classifier for facial expression recognition. The approach was evaluated in cross databases. Zhang et al. [42] combined multiple types of facial features via multiple kernel learning (MKL) in multiclass support vector machines (SVM) for facial expression recognition. The approach was evaluated in single databases and cross databases. Based on the Deep Belief Network, Liu and Chen [43] adopted a bottom-up unsupervised feature learning (BU-UFL) process that learned hierarchical features and a boosted top-down supervised feature strengthen (BTDSFS) process to do fine-tune in a supervised manner.

Recent years, deep learning based approaches illustrated that deep learning is a crucial component in facial expression recognition. Levi and Hassner [37] firstly transferred image pixels of facial expressions to a 3D metric space, and input the 3D metric data

to a Convolutional Neural Network (CNN) model to train a classification model using limited labeled training samples. Jung et al. [44] used one deep network to extract features from image facial sequences, while used the other deep network to learn temporal geometry features from temporal facial landmark points. Finally, the two networks were combined through a fine-tuning operation to boost recognition performance. Kim et al. [45] adopted multiple deep convolutional neural networks to make a primary decision for facial expression samples, then they constructed a hierarchical architecture for facial expression classification which fused exponentially-weighted decisions in multiple levels. Hasani and Mahoor [46] used a DNN-based architecture to extract features of facial expressions, and employed a linear chain Conditional Random Field (CRF) module to realize recognition in facial videos. Cross-data validation was performed on CK+, MMI and FERA databases. Lopes et al. [8] adopted pre-processing techniques to extract specific features of facial expression images. Then the CNN was used to extract features for facial representation and recognition. The proposed approach was evaluated for cross-domain recognition, *e.g.*. training in the CK+ database, test in the JAFFE database. Though the above approaches can also be used for cross-domain recognition, they were not specially designed for the target. To obtain a better performance, we propose a more effective approach which breaks boundaries among different databases to realize cross-domain recognition.

2.2. Multiple-to-single cross-domain recognition

To improve the generalization of facial expression recognition, approaches that fused features of multiple databases were presented. Shan et al. [47] normalized faces to a fixed distance between the two eyes in order to decrease sample difference in different databases. It relied on the preprocessing to extend the applicability of facial recognition in various scenarios. Kahou et al. [48] combined multiple deep neural networks to solve the problem of multi-model fusion for emotion recognition in the 2013 Emotion Recognition Wild Challenge. Their approach outperformed other teams and won the challenge. Ruiz et al. [9] presented a Hidden-Task Learning framework to recognize facial AUs. The HTL adopted easily obtained training samples of facial expressions to learn a set of AUs with rare training samples. They even extended the HTL to Semi-Hidden Task Learning where AU samples were available. They trained the proposed model in three databases and performed testing in another database. The two approaches fused features of different databases to improve recognition accuracy in cross databases. Aiming at realizing person reidentification in multiple scenarios, Wang et al. [10] presented a constrained asymmetric multi-task discriminant component analysis (cAMT-DCA) approach, which maximized the data discrepancy on the shared component in different scenarios, maximized the local inter-class variation and minimized local intra-class variation in all scenarios. Considering that facial expressions in different databases and scenarios should have common features, we proposed an approach to learn the common representation, and to weaken the representation difference in various different scenarios.

3. ICID fusion network for expression recognition

In this section, we introduce the ICID fusion network to recognize facial expressions in cross databases. Due to the excellent performance of feature learning in object detections, the DarkNet-19 network [49] is employed to build our ICID fusion network. Suppose that there are multiple databases, and samples of the mth database are represented by $D_m = \{S_i^c, i = 1, \ldots, N_c^m; c = 1, \ldots, C\}$. Here, S_i^c refers to the ith sample of category c in D_m. There are total C expression categories, and there are N_c^m samples in the category c. Using data samples in multiple databases, we train the ICID network to recognize facial expressions in cross databases.

3.1. Intra-category Common feature learning channel

For arbitrary multiple databases $D_m, m \in \{1, 2, \ldots, M\}$, we learn the common features of facial expressions in intra-categories via the Intra-category Common feature learning channel (IC channel). The IC channel is constructed based on the pre-trained DarkNet-19 network [49], and its architecture is shown in the Fig. 2. The IC channel keeps the preceding 23-layer of the DarkNet-19 network for feature extraction, and further consists of one full connected layer, one average pooling layer following the 23-layer pre-trained layers. For training of the IC channel, a determination layer is set to learn common features.

Using multiple databases $D_m, m \in \{1, 2, \ldots, M\}$ for the training of the IC channel, the process of common feature extraction of one expression category is defined as a function $\phi_1()$. With an input sample of facial expression S, the extracted feature is represented by $\phi_1(S)$. We train the IC channel network by setting the Eq. (1) as the objective function. In the equation, S_i^c and S_j^c are facial images of the same category c in different databases. Here, $\phi_1(S_i^c)$ and $\phi_1(S_j^c)$ refer to extracted common features of S_i^c and S_j^c. Then features $\phi_1(S_i^c)$ and $\phi_1(S_j^c)$ are input to the determination layer. The training processing seeks a minimum distance between $\phi_1(S_i^c)$ and $\phi_1(S_j^c)$ in the determination layer. In other words, it aims to seek a common representation for facial expressions in the same category. Therefore, the IC channel provides a common representation for facial expressions of the same category in various databases. Obviously, the training can be extended to involve multiple databases, and also be used for feature extraction in single databases. Through training, the IC channel learns common features of intra-categories that enhances the representation of facial expressions belonging to the same category.

$$\min_{i \in D_m, j \in D_n, m,n \in \{1,\ldots,M\}} \| w_1^T \phi_1(S_i^c) - w_1^T \phi_1(S_j^c) \|_F^2; \quad (1)$$

3.2. Inter-category Distinction feature learning channel

The classification of facial expressions is to maximize distances of samples in different categories, and a classifier model is designed to distinguish them. In this section, we use a deep network to realize the Distinction feature extraction in inter-categories.

We design an Inter-category Distinction feature learning channel (ID Channel) for feature extraction. The architecture of the ID Channel is shown in Fig. 3. The ID channel keeps the preceding 23-layer of the DarkNet-19 network for facial feature extraction, and consists of a softmax classifier following the 23-layer network. Using facial samples in D_m, we pre-train the ID channel by setting an objective function for category determination following the Eq. (2). Here, $Y = (y_1, y_2, \ldots, y_C)$ refers to a vector of facial expression categories, and S_i is one facial expression sample in databases $\{D_m\}$. Here, $\{D_m\}$ refers to any one database or multiple databases. The $\phi_2(S_i)$ represents extracted feature of the facial image sample S_i in the ID channel. Since the objective function is set to distinguish different categories, learned features via the ID channel are Distinction features of facial expressions in different categories.

$$\min_{i \in \{D_m\}} \| Y - w_0^T \phi_2(S_i) \|_F^2; \quad (2)$$

In training step, training samples of all categories in multiple databases are input to the channel to train the ID channel network. The trained ID channel can be regarded as a feature learning function $\phi_2(S)$, which generates Distinction features for facial expression recognition. Since we train the ID channel using training samples in multiple databases, the network has good generalization in cross-domain recognition.

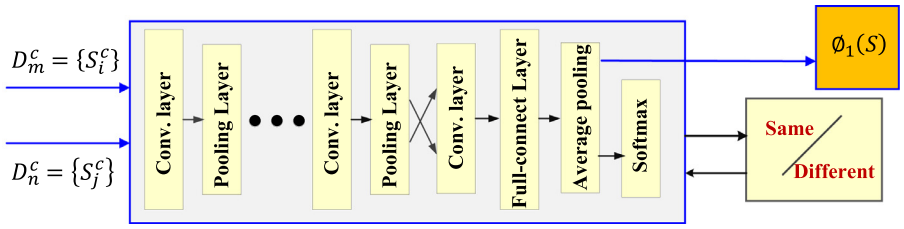


Fig. 2. Architecture of the IC channel for intra-category common feature learning. The structure of the IC channel follows the DarkNet-19 network, and it consists of 19 convolution layers, 5 max-pooling layers, and a SoftMax layer for category determination.

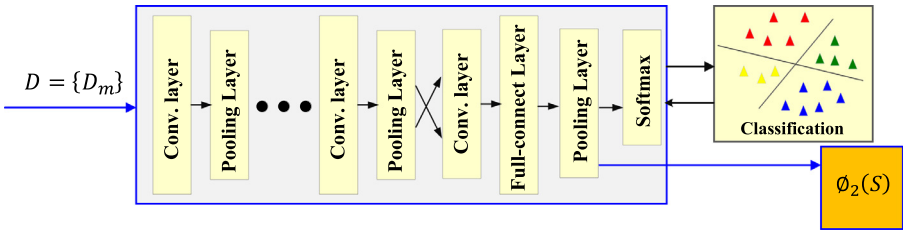


Fig. 3. Architecture of the ID channel for inter-category Distinction feature learning. The ID channel keeps the preceding 23-layer of the DarkNet-19 network for facial feature extraction, and consists of a SoftMax classifier following the 23-layer network.

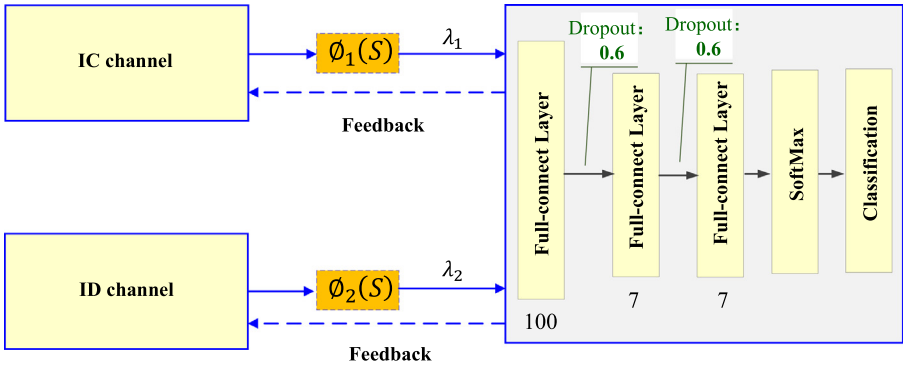


Fig. 4. The fusion network fusing the intra-category common feature and the inter-category Distinction feature for facial expression recognition. It mainly consists of 3 full-connected layers for feature fusion and feature extraction, and a ratio of 0.6 is set for dropout processing.

3.3. Feature fusion for facial expression recognition

Both of common features and Distinction features are crucial for facial expression recognition. We design a fusion network which connects the two feature learning channels, and fuses feature learned from the two channels for final facial expression recognition. The fusion network consists of 3 full connection layers, with 100, 7, and 7 neurons respectively, and a SoftMax layer determines the final expression category. The structure of the fusion network is shown in the Fig. 4. In order to automatically assign a balance between the two channels, we add two weight parameters. The two weights λ_1, λ_2 are automatically trained in the fusion network by setting a loss function, Eq. (3), to train the fusion network. Here, $Y = (y_1, y_2, \ldots, y_C)$, refers to the category vector. $X(S_k)$ refers to the fused feature. Like the Fig. 4 shows, the fusion network fuses weighted features, $\phi_1(S_k)$, $\phi_2(S_k)$, and the SoftMax classifier classifies facial expressions into corresponding categories.

Obviously, the approach can be extended to multiple-to-single and single-to-single cross-domain recognition. Moreover, it can also be applied for traditional expression recognition in a single database. We evaluate its performance of possible settings in our experiments.

$$\min_{k\in\{D_m\}} \| Y - w_2^T X(S_k) \|_F^2; \tag{3}$$

$$X(S_k) = \lambda_1\phi_1(S_k) + \lambda_2\phi_2(S_k);\quad 0 < \lambda_1, \lambda_2 < 1; \lambda_1 + \lambda_2 = 1. \tag{4}$$

3.4. Inference

As the Fig. 4 shows, we connect the IC and the ID channel using a fusion network to construct an end-to-end ICID fusion network for facial expression recognition. The complete objective function for the ICID fusion network is defined in the Eq. (5). The network training is performed in three steps. First of all, we perform pre-training on the two feature learning channels. For the IC channel, we input samples of the same category but belonging to multiple databases to the network and adopt the Eq. (1) as the loss function to train the network. Then we train the ID channel using training samples of all categories, and use the Eq. (2) as loss function to extract Distinction features $\phi_2(S_k)$. Finally, we train the ICID fusion network again to obtain weight values of λ_1, λ_2 and fine-tune the whole fusion network. During fine-tuning, the final

recognition results of the fusion network provide feedback to the IC and the ID channel for fine-tuning network parameters.

When testing a facial sample S_t, it is input to the two feature learning channels to obtain the common feature $\phi_1(S_t)$ and the Distinction feature $\phi_2(S_t)$. Then two features are fused in the fusion network, and the fusion network provides the final determination of the category for the sample S_t.

$$\min_{k,i,j\in\{D_m\},m\in\{1,\cdots,M\}} \| Y - w_2^T X(S_k) \|_F^2 + \| Y - w_0^T \phi_2(S_k) \|_F^2 + \| w_1^T \phi_1(S_i^c) - w_1^T \phi_1(S_j^c) \|_F^2 . \quad (5)$$

4. Experiments and discussions

We perform experiments on four public databases, the CK+ database [3], the MMI database [5], the SFEW database [18] and the RAF database [20], also the EmotioNet database [19]. Experiment evaluation is performed for cross-domain recognition and single-database recognition.

4.1. Databases

CK+ database is a database that captured in a lab surrounding. It consists of 1079 facial behavior sequences of 210 adults. Participants are 18 to 50 years of age. The database includes sequences for frontal views and 30-degree views, and the image resolution is either 640×490 or 640×480 pixel arrays. Image sequences are with 8-bit grayscale or 24-bit color values. There are 7 basic emotion categories: Anger, Contempt, Disgust, Fear, Happy, Sadness and Surprise. In our experiments, we clip facial behavior sequences to images for recognition.

MMI database is also a lab-captured databases. It contains more than 1500 samples of both static images and image sequences from 19 male and female participants. There consists of samples captured in the front view and the profile view. For a detailed analysis of facial expressions, the database provides single AU and multiple AU activation, and AU temporal activation patterns (onset- apex-offset). In the database, static facial-expression images are all true color (24bit) images with a resolution of 720×576 pixels, and sequences are of variable length, lasting between 40 and 520 frames. There are 6 facial expression categories.

RAF database is a large scale database containing total $29,672$ real-world facial images. Different with other databases where each image being assigned with one label, the database obtains a multi-label annotation result for each image, i.e., a seven-dimensional vector that each dimension corresponds to the votes of relevant emotion. The RAF database includes 6-class basic emotions and 12-class compound emotions. In the database, the basic attributes (gender, age, and race) of all facial images are manually annotated. In our experiment, the basic emotions are used for experimental evaluation.

SFEW database is a static facial expression database consisting of facial images extracted from movies. The database contains 7 expression categories (Anger, Disgust, Fear, Neutral, Happy, Sad, Surprise), which are acted by 95 subjects. There are total 700 images, separated into the training set and the test set.

EmotioNet database totally includes 975,000 images of facial expressions in the wild, and 25,000 images are assigned manual annotations of AUs. Expressions in the database have 6 basic facial emotion categories and 17 compound emotion categories. Because the cross-domain recognition requires the same expression categories in different databases, only part of facial samples in the EmotioNet database are chosen in our experiments.

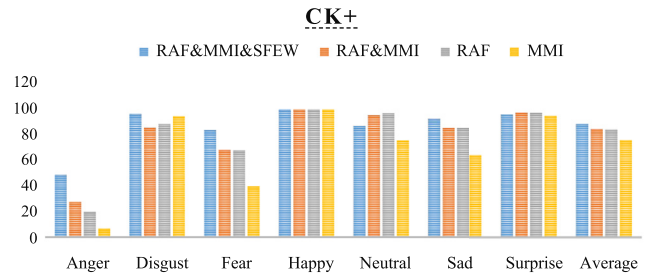
Table 1
Experiment results of cross-domain recognition using the ICID fusion network. The average recognition accuracies of 4 databases rank in descending order are as follows CK+, MMI, SFEW, RAF. Multiple database fusion effectively improves the result in cross-domain recognition.

Training Dat.	Test Dat.	Precision(%)
CK+ RAF SFEW	MMI	69.4
MMI RAF SFEW	CK+	88.7
CK+ RAF MMI	SFEW	49.4
CK+ SFEW MMI	RAF	43.8
MMI RAF	CK+	84.5
MMI SFEW	CK+	78.1
CK+ RAF	MMI	66.3
CK+ RAF	SFEW	47.1
CK+ SFEW	RAF	42.8
CK+	MMI	57.2
MMI	CK+	76.1
RAF	CK+	84.5
RAF	MMI	64.7
RAF	SFEW	45.7
CK+	SFEW	31.8
CK+	RAF	40.1

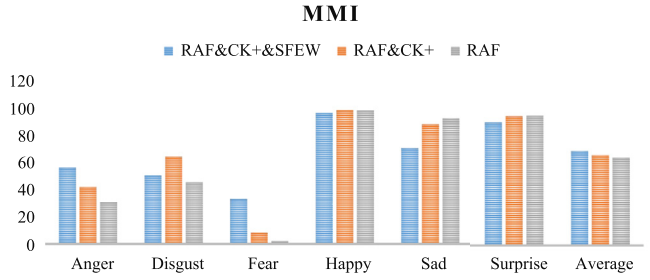
4.2. Experiments for cross-domain recognition

We evaluate the performance of our proposed approach in two experiment settings for cross-domain recognition, multiple-to-single and single-to-single cross-domain recognition. In the multiple-to-single cross-domain recognition, we train the ICID fusion network using facial samples of multiple databases, and evaluate its recognition performance in another database. In the single-to-single cross-domain recognition, the ICID fusion network is trained using facial samples in one database, and test in the other database.

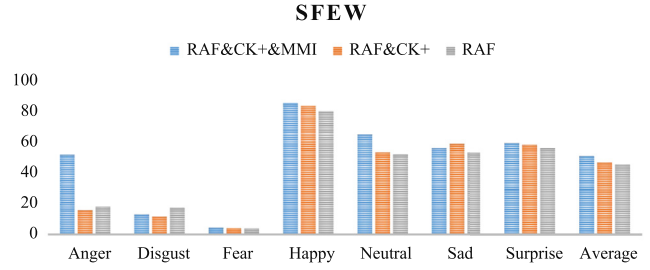
We evaluate the proposed approach for cross-domain recognition in two experiment settings, and list results in the Table 1. In our experiments, we fuse facial samples of two or three databases for training and test the trained model in another database, *e.g.*. training on CK+ and RAF and SFEW, test on the MMI, or training on CK+ and RAF, test on the MMI. In addition, we evaluate the proposed approach for single-to-single cross-domain recognition, *e.g.*. training using the CK+ database and test on the MMI database. Since the RAF database consists more data samples than other databases, it is used to train the ICID network for the test of other databases. Comparing recognition precisions of four databases, it can be seen that the CK+ always has higher precision values than the MMI, RAF and the SFEW due to the lower complexity of the CK+ database, and The average recognition accuracies of 4 databases rank in descending order are as follows $CK+ > MMI > SFEW > RAF$. We also compare recognition precisions of the CK+ when we use different databases in training. Using the MMI for training, we get a precision of 76.1%, while 84.1% is achieved using the RAF for training. In addition, when we use three databases for training, *e.g.*. MMI, RAF, and SFEW, we get a precision of 88.7%. The similar situation also occurs in the RAF database. When we train a model using CK+, SFEW, and MMI three databases, we get a test result of 43.8% in the RAF database. In addition, using CK+ and SFEW for training, a precision of 42.8% is obtained, while only a precision of 40.1% is achieved using the CK+ only for training. The reason is that the model training uses a large number of samples, including wild and lab-setting facial expressions, which improves the generalization of the trained classification model for distinguishing facial expressions. Therefore, larger and more complex samples benefit network training, and further benefits effective facial expression recognition. Experimental results indicate that multiple database fusion is able to effectively improve the accuracy of



(a) Cross-domain recognition results of CK+



(b) Cross-domain recognition results of MMI



(c) Cross-domain recognition results of SFEW

Fig. 5. Cross-domain recognition results obtained by fusing multiple databases. These figures show that higher recognition precisions are achieved when we train the ICID fusion network using more databases.

cross-domain recognition. Using more databases for training equals to improving the generalization of classifiers.

In Fig. 5, we compare experimental results of cross-domain recognition in three databases, the CK+, the MMI and the SFEW, where multiple databases are fused in the training step. According to the Fig. 5(a), the recognition precision of the CK+ obtained by training networks in three databases (the RAF, the MMI, and the SFEW) is higher than precisions obtained by training models using two (RAF and MMI) or one (MMI) database. The average recognition precisions of fusing three databases(RAF, MMI, and SFEW), two databases (RAF and MMI) and one database (MMI) for the test of CK+ database are 88.7%, 84.5%, and 76.1%, respectively. The same phenomena can also be observed from the result exhibition of the database MMI and SFEW in Fig. 5(b) and 5(c). In the Fig. 5(b), the average recognition precisions by fusing three databases(RAF, CK+, and SFEW), two databases (RAF and CK+) and one database (RAF) for the test of MMI database are 69.4%, 66.3%, and 64.7%, respectively. Obviously, it is able to effectively improve recognition precisions to fuse multiple databases in training for cross-domain facial expression recognition. And the proposed ICID fusion network always performs better with more databases for training.

We compare the performance of the proposed approach for cross-domain facial expression recognition with the state-of-the-art in the Table 2. Currently, the cross-domain recognition of facial expressions is generally performed by training a classification model in one database and test in the other database. We compare with some approaches which presented single-to-single cross-domain recognition in Table 2. Compared with Li's approach [20], our approach improves the recognition precision by 22.5% when the RAF is used for training and the CK+ is tested. When we train the fusion ICID network using samples in the MMI and test the CK+, we get at least 12% higher recognition precision than existing approaches. In paper [43], an AUDN approach was evaluated for the cross-domain recognition, *e.g.*. training CK+, test MMI, and training MMI, test CK+, which obtained recognition accuracies of 72.20% and 93.46%, respectively. The two results are very high compared with current approaches. However, the AUDN approach had much worse performance when used the SFEW as the test database. Compared with the AUDN approach [43], our proposed approach has much reliable improvement on all databases. We further perform an experiment which trains the ICID fusion network in the CK+ and the RAF databases, and tests the CK+ database. We obtain a 97.7% recognition precision, which is higher than all existing approaches. The comparison indicates the effectiveness of our proposed approach for cross-domain recognition.

Table 2
Comparison with the state-of-the-art approaches for cross-domain recognition. Our approach improves the recognition precision by 22.5% when the RAF is used for training and the CK+ is tested. When using the MMI to train the ICID network, and test the CK+, our approach obtains at least 12% higher precision than existing approaches.

Training Dat.	Testing Dat.	Precision
CK+	SFEW	**31.8**[Ours]; 29.43 [43];
MMI	SFEW	**27.6**[Ours]; 25 [43];
MMI	CK+	**76.1**[Ours]; 60.8 [50]; 64.2 [51];
CK+	MMI	**57.2**[Ours]; 53.2 [50]; 55.6 [51];
RAF	CK+	**84.5**[Ours]; 62 [20];
CK+	RAF	**40.1**[Ours]; 38.8 [20];

Table 3
Experiment results of single-database recognition using the ICID fusion network. Categories of Happy, Sad and Surprise are easily distinguished from other expressions, and the category Fear is always confused with the category Disgust because these two expressions sometimes have no significant difference.

Database	Ang.	Dis.	Fear	Hap.	Nat.	Sad.	Surp.	Ave.
CK+	87.9	93.4	83.4	100	100	100	93.7	95.1
MMI	61.1	72.6	44.6	83.4	–	89.8	87.4	76.3
SFEW	51.9	13.1	4.3	86.1	65.1	56.2	31.6	51.2
RAF	79.1	56.3	50.0	87.7	83.8	86.4	84.8	75.4

4.3. Evaluation of recognition in single databases

We also evaluate the proposed ICID fusion network for facial expression recognition in single databases. For that, we separate one database to the training set and the test set and further separate the training set into two parts for the training of ICID fusion network. Recognition results are compared with related approaches.

Recognition precisions of four databases, *i.e.* CK+, MMI, SFEW, and RAF, obtained by the proposed ICID fusion network are listed in the Table 3. The table shows recognition precisions of all 7 facial expression categories. As the table shows, the average recognition precision of the database CK+ is higher than other databases. Both of the CK+ and the MMI database are captured in a lab-guided setting. The CK+ database consists of a larger quantity of samples, thus it always obtains better experiment performance than the MMI database. Among all facial expression categories, categories of Happy, Sad and Surprise are easily distinguished from other expressions according to the recognition results in the Table 3.

Table 4
Comparison with related approaches for single-database recognition. The proposed ICID fusion network is also effective for facial expression recognition in single databases.

Database	Precision
CK+	**95.1**[Ours]; 91.4 [29]; 96.4 [52,53]; 93.7 [43]; 93.2 [51]; 95.75 [8]; 93.04 [46]; 95.78 [20]
MMI	**76.3**[Ours]; 77.6 [51]; 75.8 [43];
SFEW	**51.2**[Ours]; 51.05 [20]; 30.14 [43];
RAF	**75.4**[Ours]; 74.2 [20];

Table 5
Experiment results obtained using the ResNet and the Darknet-19 for feature learning in the ICID fusion network. These results certified that the Darknet-19 performs better for feature learning in our proposed ICID fusion network.

Training Dat.	Test Dat.	Precision(%)
MMI RAF SFEW	CK+	88.7(DarkNet), 84.7(ResNet)
CK+ RAF MMI	SFEW	49.4(DarkNet), 46.34(ResNet)
MMI RAF	CK+	84.5(DarkNet), 82.2(ResNet)
MMI SFEW	CK+	78.1(DarkNet), 76.0(ResNet)
MMI	CK+	76.1(DarkNet), 73.4(ResNet)
CK+	SFEW	31.8(DarkNet), 30.3(ResNet)
CK+	CK+	95.1(DarkNet), 94.3(ResNet)
MMI	MMI	76.3(DarkNet), 70.8(ResNet)
SFEW	SFEW	51.2(DarkNet), 47.0(ResNet)
RAF	RAF	75.4(DarkNet), 70.8(ResNet)

Moreover, the category Fear is always confused with the category Disgust because these two expressions sometimes have no significant difference.

In Table 4, we compare the ICID fusion network with related approaches. As shown, the ICID fusion network also achieves higher precisions for facial expression recognition in single databases although the approach is designed for cross-domain recognition. In the RAF database, the ICID fusion network obtains a precision which is 1.2% higher than the state-of-the-art result. Therefore, the proposed ICID fusion network is also effective for facial expression recognition in single databases. It is mainly due to the excellent feature learning ability in the IC and the ID channels. The approach is also able to solve other classification problems, *e.g.* the action recognition, image classification, *etc.*

4.4. Experiments using the ResNet for feature learning

In order to evaluate the performance of different deep networks on feature learning in our model, we further use the ResNet101 [54] as a feature learning network to replace the DarkNet in our ICID fusion network, and evaluate its performance for single-database and cross-domain facial recognition.

All experiments have the same experimental setting with experiments using the DarkNet. In this experiment, we modify the ResNet101 by setting 7 convolution kernels in the last convolution layer for the classification of 7 basic expression categories, and use the modified ResNet to replace the DarkNet part in the IC and ID channels for feature learning. Experiment results of the cross-domain and the single-database recognition are shown in the Table 5. Here, we mark these results with (ResNet) which are obtained using the ResNet for feature learning, and results obtained using the Darknet-19 for feature learning are marked with (DarkNet). As shown in the Table 5, the ResNet has worse performance in both the cross-domain recognition and the single-database recognition. Basic network units of both the ResNet and the DarkNet-19 have three convolution layers, but have different number of convolution kernels. The reason may be because the network structure of DarkNet-19 is compact that it is suitable for facial feature extraction. Therefore, we choose the DarkNet-19 to learn facial features in our proposed ICID network.

Table 6
Evaluation of cross-domain recognition on the EmotioNet database. Training is performed using facial samples in other one or more databases, and evaluation is realized in the EmotioNet database.

Training Dat.	Precision(%)
MMI RAF SFEW	61.4
CK+ RAF MMI	62.3
CK+ SFEW MMI	58.0
MMI RAF	60.7
MMI CK+	49.6
CK+ RAF	54.8
CK+ SFEW	44.7
CK+	38.1
MMI	46.5
RAF	52.6

4.5. Evaluation of cross-domain recognition on the EmotioNet database

Since only part of facial samples are assigned expression labels in the EmotioNet database, we use these 1136 facial images which are provided expression labels of 6 basic categories in this cross-domain evaluation. Other samples with Action Units (AU) definitions are not suitable for our experiment. Because there are less than 100 samples in each category, we train the ICID network using facial samples in other one or more databases, and perform evaluation on selected samples in the EmotioNet database. Evaluation results are illustrated in the Table 6.

Because the MMI database contains 6 basic expression categories, and other databases contain 7 basic categories, there are three experiments that we train the recognition model using 6 expression categories, including experiments using the MMI database, the RAF and the MMI database, the CK+ and the MMI database for training. Since the EmotioNet also has 6 basic expression categories, thus we get higher accuracies in these three experiments. In other experiments, samples of 7 categories are used for training, and the trained model is used to recognize 6 categories in the EmotioNet. As illustrated in the Table 6, obviously the combination of multiple databases contributes a higher recognition accuracy. The reason is not only due to more training samples, but also because the mixture of samples in multiple databases enhances the ability of the ICID network to extract distinguishable features. The experiment of cross-domain recognition in the EmotioNet database certifies the effectiveness in wild databases.

5. Conclusion

In this paper, we proposed a novel ICID fusion network to recognize facial expressions in cross databases. The ICID fusion network consisted of an Intra-category Common feature representation channel (IC) and an Inter-category Distinction feature representation channel (ID) for facial expression representation, and a fusion network combined features of two channels for facial expression recognition. The proposed approach was evaluated for cross-domain recognition and single-database recognition. In the cross-domain recognition, we designed two experiment settings to evaluate the performance of the ICID fusion network, multiple-to-single cross-domain recognition, and single-to-single cross-domain recognition. According to experiments of the cross-domain recognition, multiple database fusion is able to effectively improve the accuracy of cross-domain recognition, which equals to improving

the generalization of classifiers. Compared with state-of-the-art results, the proposed ICID fusion network got a significant improvement, for example, 22.5% improved precision in the experiment of training the RAF and testing the CK+. For single-database recognition, the ICID fusion network also achieved state-of-the-art performances.

Acknowledgments

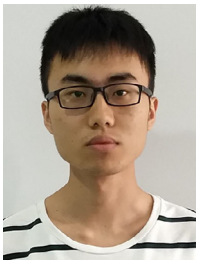
This research is supported by the Natural Science Foundation of China (NSFC) under grant no. 61673088 and grant no. 61305043. This work was partly supported by the 111 Project No. B17008.

References

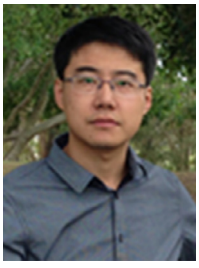
[1] P. Liu, S. Han, Z. Meng, Y. Tong, Facial expression recognition via a boosted deep belief network, in: Proceedings of the CVPR, 2014.
[2] M. Liu, S. Shan, R. Wang, X. Chen, Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition, in: Proceedings of the CVPR, 2014.
[3] 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, in: CVPR, 2010.
[4] M. Pantic, M.F. Valstar, R. Rademaker, L. Maat, Web-based database for facial expression analysis, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2005.
[5] M.F. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition to the mmi facial expression database, in: Proceedings of the International Conference Language Resources and Evaluation, Workshop on EMOTION, 2010.
[6] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007) 1683–1699.
[7] M.F. Valstar, M. Pantic, Fully automatic recognition of the temporal phases of facial actions, IEEE Trans. Syst. Man Cybern. Part B 42 (1) (2012) 28–43.
[8] A.T. Lopes, E. 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 Recognit. 61 (2017) 610–628.
[9] A. Ruiz, J. Weijer, X. Binefa, From emotions to action units with hidden and semi-hidden-task learning, in: Proceedings of the ICCV, 2015.
[10] X. Wang, W. Zheng, X. Li, J. Zhang, Cross-scenario transfer person reidentification, IEEE Trans. Circuits Syst. Video Technol. 26 (6) (2016) 1447–1460.
[11] M.J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, Coding facial expressions with gabor wavelets, in: Proceedings of the International Conference on Face and Gesture Recognition, 1998.
[12] T. Kanade, Y. Tian, J.F. Cohn, Comprehensive database for facial expression analysis, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2000.
[13] S.M. Mavadati, M.H. Mahoor, K. Bartlett, P. Trinh, Automatic detection of non–posed facial action units, in: Proceedings of the IEEE International Conference on Image Processing, 2012.
[14] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2008.
[15] A. Savran, N. Alyuz, H. Dibeklioglu, O. Celiktutan, B. Gokberk, B. Sankur, L. Akarun, Bosphorus database for 3d face analysis, in: Proceedings of the Biometrics and Identity Management, 2008.
[16] W. Yan, Q. Wu, Y. Liu, S. Wang, X. Fu, Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces, in: Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013.
[17] F. Qu, S. Wang, W. Yan, X. Fu, Cas(me)2: A database for spontaneous macro–expression and micro-expression spotting and recognition, IEEE Trans. Affect. Comput. 99 (2017). 1–1
[18] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark, in: Proceedings of the ICCV, 2011.
[19] C.F. Benitez-Quiroz, R. Srinivasan, A.M. Martinez, Emotionet: an accurate, real–time algorithm for the automatic annotation of a million facial expressions in the wild, in: Proceedings of the CVPR, 2016.
[20] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: Proceedings of the CVPR, 2017.
[21] S.A.M. Al-Sumaidaee, M.A.M. Abdullah, R.R.O. Al-Nima, S.S. Dlay, J.A. Chambers, Multi-gradient features and elongated quinary pattern encoding for image-based facial expression recognition, Pattern Recognit. 71 (2017) 249–263.
[22] Y. Liu, Y. Li, X. Ma, R. Song, Facial expression recognition with fusion features extracted from salient facial areas, Sensors 17(4) (2017).
[23] X. Fan, T. Tjahjadi, A dynamic framework based on local Zernike moment and motion history image for facial expression recognition, Pattern Recognit. 64 (2017) 399–406.
[24] B. Ryu, A.R. Rivera, J. Kim, O. Chae, Local directional ternary pattern for facial expression recognition, IEEE Trans. Image Process. 26(12) (2017) 6006–6018.
[25] K. Sikka, G. Sharma, M. Bartlett, Lomo: latent ordinal model for facial analysis in videos, in: Proceedings of the CVPR, 2016.
[26] M. Wollmer, M. Kaiser, F. Eyben, B. Schuller, G. Rigoll, Lstm-modeling of continuous emotions in an audiovisual affect recognition framework, Image Vis. Comput. 31 (2) (2013) 153–163.
[27] M.J. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Trans. Pattern Anal. Mach. Intell. 21 (12) (1999) 1357–1362.
[28] W. Gu, C. Xiang, Y. Venkatesh, D. Huang, H. Lin, Facial expression recognition using radial encoding of local Gabor features and classifier synthesis, Pattern Recogn. 45 (1) (2012) 80–91.
[29] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.
[30] A. Dhall, A. Asthana, R. Goecke, T. Gedeon, Emotion recognition using PHOG and LPQ features, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2011.
[31] Y. Koda, Y. Yoshitomi, M. Nakano, M. Tabuse, A facial expression recognition for a speaker of a phoneme of vowel using thermal image processing and a speech recognition system, in: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication, 2009.
[32] P. Liu, L. Yin, Spontaneous facial expression analysis based on temperature changes and head motions, in: Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 2015.
[33] R. Walecki, O. Rudovic, V. Pavlovic, M. Pantic, Variable-state latent conditional random fields for facial expression recognition and action unit detection, in: Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 2015.
[34] A. Kacem, M. Daoudi, B.B. Amor, J.C. Alvarez-Paiva, A novel space-time representation on the positive semidefinite cone for facial expression recognition, in: Proceedings of the ICCV, 2017.
[35] P. Lemaire, M. Ardabilian, L. Chen, M. Daoudi, Fully automatic 3D facial expression recognition using differential mean curvature maps and histograms of oriented gradients, in: Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 2013.
[36] S. Berretti, B.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.
[37] G. Levi, T. Hassner, Emotion recognition in the wild via convolutional neural networks and mapped binary patterns, in: Proceedings of the ACM on International Conference on Multimodal Interaction, 2015.
[38] W. Sun, H. Zhao, Z. Jin, An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks, Neurocomputing 267 (2017) 385–395.
[39] P. Barros, G. I.Parisi, C. Weber, S. Wermter, Emotion-modulated attention improves expression recognition: A deep learning model, Neurocomputing 253 (2017) 104–114.
[40] S. Koelstra, M. Pantic, I. Patras, A dynamic texture-based approach to recognition of facial actions and their temporal models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 1940–1954.
[41] P. Liu, S. Han, Z. Meng, Y. Tong, Facial expression recognition via a boosted deep belief network, in: Proceedings of the CVPR, 2014.
[42] X. Zhang, M.H. Mahoor, S.M. Mavadati, Facial expression recognition using lp-norm Mkl multiclass-svm, Mach. Vis. Appl. 26 (4) (2015) 467–483.
[43] S.S. M. Liu S. Li, X. Chen, Au-inspired deep networks for facial expression feature learning, Neurocomputing 159 (2015) 126–136.
[44] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: Proceedings of the ICCV, 2015.
[45] B.K. Kim, J. Roh, S.Y. Dong, S.Y. Lee, Hierarchical committee of deep convolutional neural networks for robust facial expression recognition, J. Multimod. User Interfaces 10 (2) (2016) 173–189.
[46] B. Hassani, M.H. Mahoor, Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields, in: Proceedings of the CoRR, abs/1703.06995 (2017).
[47] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.
[48] S.E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C. Gulcehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R.C. Ferrari, M. Mirza, S. Jean, P.L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, K.R. Konda, Z. Wu, Combining modality specific deep neural networks for emotion recognition in video, in: Proceedings of the International Conference on Multimodal Interaction, 2013.
[49] J. Redmon, A. Farhadi, Yolo9000: better, faster, stronger, in: Proceedings of the CVPR, 2017.
[50] C. Mayer, M. Eggers, B. Radig, Cross-database evaluation for facial expression recognition, Pattern Recognit. Image Anal. 24 (1) (2014) 124–132.
[51] A. Mollahosseini, D. Chan, M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2016.
[52] F.D. laTorre, W.S. Chu, X. Xiong, F. Vicente, X. Ding, J. Cohn, Intraface, in: Proceedings of the IEEE International Conference and Workshops on Automatic Faceand Gesture Recognition(FG), 2015.

[53] A. Dapogny, K. Bailly, S. Dubuisson, Pairwise conditional random forests for facial expression recognition, in: Proceedings of the ICCV, 2015.
[54] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the CVPR, 2016.

Yanli Ji is currently an Associate Professor in the University of Electronic Science and Technology of China (UESTC). She obtained her Ph.D degree from Department of Advanced Information Technology, Kyushu University, Japan at Sep. 2012. Her research interests include Human Robot Interaction related topics, e.g. human activity recognition, emotion analysis, hand gesture recognition and facial expression recognition.



Yuhan Hu is currently a Master student in the University of Electronic Science and Technology of China (UESTC).



Yang Yang is currently with University of Electronic Science and Technology of China. He was a Research Fellow under the supervision of Prof. Tat-Seng Chua in National University of Singapore during 2012-2014. He was conferred his Ph.D. Degree (2012) from The University of Queensland, Australia. During the PhD study, Yang Yang was supervised by Prof. Heng Tao Shen and Prof. Xiaofang Zhou. He obtained Master Degree (2009) and Bachelor Degree (2006) from Peking University and Jilin University, respectively. His research interests include multimedia content analysis, computer vision and social media analytics.

Fumin Shen received his Bachelor degree at 2007 and Ph.D degree at 2014 from Shandong University and Nanjing University of Science and Technology, China, respectively. Now he is an Associate Professor with Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China. His major research interests include computer vision and machine learning. He was the recipient of the Best Paper Award Honorable Mention at ACM SIGIR 2016 and the World's FIRST 10K Best Paper Award - Platinum Award at IEEE ICME 2017.

Heng Tao Shen is currently a Professor of National "Thousand Talents Plan", the Dean of School of Computer Science and Engineering, and the Director of Center for Future Media at the University of Electronic Science and Technology of China. He is also an Honorary Professor at the University of Queensland. He obtained his BSc with 1st class Honours and PhD from Department of Computer Science, National University of Singapore in 2000 and 2004 respectively. He then joined the University of Queensland as a Lecturer, Senior Lecturer, Reader, and became a Professor in late 2011. His research interests mainly include Multimedia Search, Computer Vision, Artificial Intelligence, and Big Data Management. He has published 200+ peer-reviewed papers, most of which appeared in top ranked publication venues, such as ACM Multimedia, CVPR, ICCV, AAAI, IJCAI, SIGMOD, VLDB, ICDE, TOIS, TIP, TPAMI, TKDE, VLDB Journal, etc. He has received 6 Best Paper Awards from international conferences, including the Best Paper Award from ACM Multimedia 2017 and Best Paper Award-Honorable Mention from ACM SIGIR 2017. He has served as a PC Co-Chair for ACM Multimedia 2015 and currently is an Associate Editor of IEEE Transactions on Knowledge and Data Engineering.