IdenNet: Identity-Aware Facial Action Unit Detection

Cheng-Hao Tu[1], Chih-Yuan Yang[2] and Jane Yung-jen Hsu[2]
Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[1]r04922023@ntu.edu.tw, [2]{yangchihyuan, yjhsu}@csie.ntu.edu.tw

***Abstract*— Facial Action Unit (AU) detection is an important task to enable the emotion recognition from facial movements. In this paper, we propose a novel algorithm which utilizes identity-labeled face images to tackle the identity-based intra-class variation of AU detection that the appearances of the same AU vary significantly among different subjects, which makes existing methods generate low performance under cross-domain scenarios in case that the training and test datasets are dissimilar. The proposed method is based on network cascades consisting of two sub-tasks, face clustering and AU detection. The face clustering network, trained from a large dataset containing numerous identity-annotated face images, is designed to learn a transformation to extract identity-dependent image features, which are used to predict AU labels in the second network. The cascades are jointly trained by AU- and identity-annotated datasets that contain numerous subjects to improve the method's applicability. Experimental results show that the proposed method achieves state-of-the-art AU detection performance on benchmark datasets BP4D, UNBC-McMaster, and DISFA.**

I. INTRODUCTION

Facial expressions are important non-verbal signals conveying people's emotional states and intentions. Psychological studies show that emotion signals only occur at a small number of face regions such as the nose, eyes, and mouth [1]. In order to systematically describe facial expressions, psychologists devised Facial Action Coding Systems [2], providing the labels of taxonomized facial muscle movements. With the advance of computer vision, the system is widely used to describe a person's emotions by observing one's facial appearance.

Based on different descriptive frameworks, existing recognition methods for facial expressions are categorized into two approaches: primary-emotion-based and AU-based. The former [3], [4] aims to recognize the overall emotional expressions shown on subjects' faces, usually categorized into 6 to 8 classes [5], as they are common to observe and simple to describe. As a result, emotion-annotated datasets are usually rich of subjects and images, serving methods in this approach, which require numerous training data to extract effective facial features from face images to directly predict the target emotion labels. However, the number of primary emotions used in those dataset are limited, far less than the overall emotions people can express because emotional expressions are hard to be precisely classified due to their tiny visual changes. If the number increases, the labels of emotional expressions subjectively recognized by different annotators will lead to a higher level of ambiguity and result in low performance of most recognition methods.

Facial action units (AUs) are the enhanced representation of facial expressions in terms of the accuracy and flexibility, in particular for real-world expressions triggered by mixed emotions such as happy and surprised or angry and sad [6], [7]. Numerous AU detection methods have been proposed in the literature. Among them, one challenge lies in compiling AU-annotated facial expression datasets, which requires well-trained annotators because AUs' fine-grained appearances are uneasy to recognize. In general, it takes six months to train a qualified annotator, and the annotator needs at least one hour to annotate a one-minute video clip [8] because multiple AUs often occur simultaneously on a frame.

Compared with primary-emotion-labeled datasets, AU-labeled datasets are smaller in terms of numbers of both subjects and images, which leaves the performance of cross-domain scenarios still an open question. In this paper, we propose an AU detection method achieving good performance under both within- and cross-dataset scenarios.

The proposed method is based on convolutional neural networks (CNNs), which have been shown to achieve state-of-the-art performance on many computer vision problems, such as object recognition [9], pose estimation [10], and face verification [11]. One of CNNs' advantages lies in its nonlinear architecture, which is very successful to extract effective features among highly various images. With sufficient data and efficient machines, CNNs have become a powerful tool to tackle many long-standing recognition and classification problems. For the facial expression recognition (FER) problem, we observe two challenges still unsolved. First, caused by diverse races, genders, and ages, subjects' appearance difference can easily and significantly exceed the change brought by facial expressions. For example, wrinkles on a subject's forehead may mean surprise if the subject is young, but neutral if the subject is old due to their natural existence. Some facial hairstyles may hide lips, chins and jaws, and lead to incorrect expression predictions. Second, compared with large face image datasets such as CelebA [12], the sample numbers and subject diversity of existing AU-annotated datasets are much smaller. For example, the widely used AU-annotated datasets UNBC-McMaster [13], DISFA [14], and BP4D [15] cover only 25, 27, and 41 subjects respectively. With a few subjects,

This research was supported in part by the Ministry of Science and Technology of Taiwan (MOST 108-2633-E-002-001, 107-2218-E-002-009, 107-2811-E-002-575), National Taiwan University, Intel Corporation, and Delta Electronics.

978-1-7281-0089-0/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: National Taiwan University. Downloaded on May 29,2023 at 07:53:18 UTC from IEEE Xplore. Restrictions apply.

it is questionable how effectively an AU-detection algorithm can be trained and widely applied, especially for a cross-domain scenario where test subjects are unlike the training ones because the used model may be insufficiently trained.

For the first challenge, the fine-grained appearance detection, we propose multi-task network cascades with a specialized task of face clustering to reduce individual bias using translation transformation in the feature space [16]. For the second challenge, the shortage of a large number of subjects, we exploit subject-rich non-FER face datasets and integrate them with AU-annotated datasets to improve the performance, especially for cross-domain scenarios.

To sum up, we propose the novel method named IdenNet that achieves state-of-the-art performance under cross-domain scenarios, and its novelty is twofold. First, we propose a novel process to tackle individual differences using auxiliary identity-labeled training images by applying translation transformation in the learned feature space. Second, to realize the idea, we adopt the architecture of CNN cascades containing two separate tasks for identity-dependent feature extraction in one task and identity subtraction along with AU detection in another task.

II. RELATED WORK

The proposed method—exploiting identity-labeled facial images to support AU recognition—is related to several fields of studies in computer vision. Thus we categorize related methods into a few groups by the ways they deal with the problem and discuss the relationship between our and existing methods, especially for the use of face images.

Facial Action Unit Recognition. Facial action unit recognition addresses the limitations of prototypic facial expressions that human emotions and intentions are more often communicated by changes in one or a few discrete facial features [17]. Facial action coding systems decompose prototypic facial expressions into detailed facial muscle movements and result in a more challenging problem to detect tiny changes on faces. Like other studies in computer vision, using effective image features is essential to improve recognition performance. Many state-of-the-art methods [18], [19] use learned CNN features to replace hand-crafted features [20], [21] to generate more distinguishing representation.

In addition to extract distinguishing features, another important issue of AU detection is how to handle the inter- and intra-class variations from extracted features. Based on the observation that human faces are structurally similar and their discriminative characteristics can be extracted through group sparsity, the method JPML (abbreviated from Joint Patch and Multi-label Learning) [21] develops a mapping function through a sparse dictionary to predict AU labels from facial regions. Inspired by the success of deep learning in computer vision, a CNN-based method DRML (Deep Region and Multi-label Learning) [18] replaces the sparse dictionary used in JPML with an end-to-end CNN model and generates better performance due to the high-performance feature extractors embedded in CNNs. DRML introduces a region layer which splits the input face image into a uniform grid and extracts features from individual block regions. However, there is an obvious limitation in its presumption. Due to the wide range of face shapes, faces are unable to be fully aligned, and thus the split grid blocks inherently cover inconsistent facial regions. To mitigate this problem, another CNN-based method ROINet [22] crops regions of interest by detecting facial landmarks at first.

Similar to our method, Zhang et al. [23] also utilizes identity-annotated datasets to enrich face variants in AU detection learning process. They propose an Adversarial Training Framework (ATF) that confuses CNNs in classifying identities to produce subject-invariant features. From the difference perspective, our method incorporate the existing face clustering CNNs with good performance on discriminating identities to construct robust features for AU detection. Compared with ATF, IdenNet is easier to improve by replacing with a better face clustering model.

All the four methods belongs to inductive learning and learn a mapping function from an image to AU labels, which is intuitive and straightforward. However, they only report their performance under the within-dataset scenario, i.e. learn from and test on the same dataset, and there is a question about their performance under the cross-dataset scenario. We address this question in this paper, and find the proposed method outperforms existing methods under both scenarios because the proposed method exploits a large number of identity-labeled images to reduce subjects' attribute bias.

Other than inductive learning, several existing methods take another learning approach called transductive learning to recognize facial AUs, by which an test image's AU labels are inferred directly from examples, rather than from a pre-trained mapping function. Selective Transfer Machine [24] trains personalized SVM (Support Vector Machine) classifiers for individual subjects. Given a test image, the method selects a group of example images based on the similarity to the test image, and creates a personal classifier using the weighted example images. Since this method minimizes the mismatch between training and test samples by disregarding dissimilar training samples, it outperforms the baseline of a single generic classifier evenly learned from all training data. However, the method has to re-weigh all training data for every test sample and its optimization process is formulated in an iterative manner, which is computationally expensive for a large training set.

To improve its speed, SVTPT (Support Vector-based Transductive Parameter Transfer) [25] is proposed to train multiple SVM classifiers, one for a subject, along with an additional SVM regressor, to generate a new SVM classifier for a test subject by predicting the new classifier's model parameters from the previously trained ones through regression. However, SVM regression is sensitive to subtle changes if there are only limited training data, which may lead to incorrect AU labels from mutable facial expressions. To overcome the problem, CPM (Confidence Preserving Machine) [26] is proposed to isolate difficult, confusing training samples at first, and use the remaining data to train a

Authorized licensed use limited to: National Taiwan University. Downloaded on May 29,2023 at 07:53:18 UTC from IEEE Xplore. Restrictions apply.

pair of classifiers to separate assuredly positive and negative samples, and finally propagate the predictions from the easy samples to the isolated hard samples through spatio-temporal constraints. Because methods in this category require more computation in the test phase, they are generally slower than inductive-learning methods in terms of real-time response.

Because both facial unit action detection and primary-emotion-labeled FER aim to generate labels from facial images, they are two problems highly related each other and their shared information is exploited by a few recently published methods. The abundant amount of primary-emotion-labeled datasets are transferred to offer probabilistic priors as extra information to improve the accuracy of an AU detector [27]. To alleviate the intensive effort to generate AU labels, weakly supervised learning methods are proposed to take advantage of unlabeled training data [19], [28], [29]. Unlike facial images captured in labs where subject's poses are highly controlled, facial images in the wild contain diverse poses. To address the issue, pose information is extracted in terms of facial landmarks and combined with appearance features to recognize facial action units [30].

Facial Expression Recognition. Facial expression recognition is a research topic highly close to facial AU recognition and differs primarily in annotations—expression labels are created based on subjectively emotional expressions while AU labels are coded through a system taxonomizing facial muscle movements. Numerous methods and datasets have been proposed in the literature to recognize several primary emotions including anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral [31], [32], [33]. To the best of our knowledge, no existing AU recognition method exploits identity subtraction in the same way as the proposed IdenNet, but a few existing facial expression recognition methods partially share similar components of IdenNet.

CNN-based feature extractors are used by CCNET+CDA [34], IACNN [35], 2B(N+M) [36], and IdenNet, but the CNN architecture of CCNET+CDA is very shallow, i.e. only a convolutional layer and a max-pooling layer. Unlike the other three methods which learn convolutional kernels using back propagation, CCNET+CDA learns the kernels through contractive discriminative analysis (CDA) and presumes that expressions are the only discriminative factor. Therefore, CCNET+CDA does not use any identity information.

IACNN differs from 2B(N+M) and IdenNet in the way of using identity labels. IACNN simply treats identity labels as another type of expression labels so its architecture is simply expanded from its simplified baseline Exp-Net. This approach works well for datasets of which images are captured in well controlled situations where image variation only comes from either expressions or subjects. On the contrary, IdenNet is motivated by the observation that images in the wild may be highly noisy and inconsistent. Along with identities, several issues such as lighting conditions, camera models, distance, head poses, and noise will all increase the intra-class variation. Thus IdenNet first exploits a noise-resistant face feature extractor, then refines extracted features by identity labels, and finally employs identity subtraction to improve recognition accuracy.

2B(N+M) and IdenNet are similar in terms of the CNN-based architecture containing two branches, one for facial label classification and another for metric learning using identity information. While 2B(N+M) primarily validates that its proposed 2B(N+M)-tuplet cluster loss along with the two-branch architecture is more effective than other tuplet losses for metric learning, IdenNet is proposed to generate robust performance efficiently. Thus IdenNet adopts LightCNN as the feature extractor instead of 2B(N+M)'s Inception-style convolutional groups because LightCNN is explicitly designed for face feature extraction. In addition, to integrate feature vectors generated by the two branches, IdenNet utilizes identity subtraction instead of concatenation used by 2B(N+M).

Neutral-subtracted features are effective to represent expressive appearances by reducing the influence of background pixels and skin tones [37]. Although both Haber et al.'s method [32] and IdenNet utilize subtraction to generate effective features, the targets to be subtracted are different. Haber et al.'s target is an average neutral face because a neutral face is most different from expressive ones. Thus, Haber et al. train a SVM to explicitly classify faces into expressive and neutral ones, and then generate the average neutral face feature for subtraction. In contrast, IdenNet does not require an external binary classifier. Its triplet loss of the clustering task implicitly drives CNN weights to distinguish neutral and expressive faces because the samples of the triplet loss are collected by hard-negative mining, which prefers significantly different training examples, i.e. a pair of neutral and expressive faces.

Face Space. Face space is a theoretical idea in psychology where recognizable faces are stored [38], and there are many psychological properties contained in the space as such race bias and recognition distinctiveness effect [39], [40]. In recent years, new findings are discovered that invariant features of faces can be used to facilitate facial expressions [41] and facial motor information is sufficient for identity recognition [42]. As images are digitalized, they can be interpreted as a group of data in a high-dimensional space accessible by computers, and face space is a special subspace contained. These psychology studies provide theoretical supports for the proposed method to detected AU patterns. From the respective of face space, facial AU detection is equivalent to localizing numerous subspaces where AU-labeled faces are stored.

Face-based Identity Recognition. Face recognition is a long-established computer vision problem and numerous datasets and methods are repeatedly proposed to address the problem. Advanced by large datasets and CNNs' sophisticated models, promising results have been demonstrated on many 2D image datasets [43], [11], [44]. Due to its long research history and successful high performance, there are ample opportunities to utilize developed face recognition

Authorized licensed use limited to: National Taiwan University. Downloaded on May 29,2023 at 07:53:18 UTC from IEEE Xplore. Restrictions apply.

methods along with their training data to address other face-related problems. The proposed method takes the approach and fully utilizes existing face recognition methods [44], [45] by adopting their feature extracting architecture and CNN weights pre-trained by a large face recognition dataset [12].

III. PROPOSED METHOD

IdenNet aims to tackle a problem of facial AU detection: the identity-caused appearance change may exceed the one generated by expressions. If a classifier is trained without taking the problem into consideration, the learned model may fail to predict correct AUs for new subjects. To handle this issue, the theoretically best approach is to expand the training dataset, to ensure that the dataset covers sufficiently diverse subjects so that any attribute can be sampled. However, due to the considerable effort to annotate AUs, all existing AU-labeled datasets are all compiled with limited subjects. To overcome the difficulty, we propose multi-task network cascades containing two sub-tasks—face clustering and AU detection—with shared convolutional layers to extract low-level features, as shown in Fig. 1. The incorporated face clustering task is trained using identity-annotated (ID-annotated) datasets by exploiting their abundant amount of face images and wide range of subject attributes. It is designed to extract features from face images in a way that the feature vectors will be close to each other in the feature space if they are from the same subject, and far apart if they are not. The extracted features are used as parameters for translation transformation in the feature space so that the difference caused by subjects will be significantly reduced. To illustrate the idea, we show an example in Fig. 2 using the t-SNE [46] to visualize a set of real-world images. The individual differences among subjects make their feature vectors (represented by face images) scatter at different locations in the feature space. We aim to learn subject-specific transformations for all subjects to eliminate the bias caused by their identities so that AUs will become easier to be recognized. At the end, we recognize the AUs of an input image using its transformed facial features through an AU detection task.

IdenNet's cascades structure is inspired by [47] to solve sub-tasks sequentially by exploiting causal dependency among them. To accommodate a large amount of training data with efficient computation, we utilize a compact CNN model named LightCNN [45] as the shared layers between our two sub-tasks.

Face Clustering. Let $\mathbf{x}_i$ be a face image and $\mathbf{y}_i \in \{-1,1\}^c$ be its label where c is the number of AU classes, X and Y be the training dataset containing N sample pairs, and $S(\mathbf{x}_i)$ be a feature vector extracted by the shared layers presented in Table I. Note that the MFM (Max-Feature-Map) operation in LightCNN is an alternative of the widely used ReLU (Rectified Linear Unit) activation, and makes this model light and robust since it only suppresses a small number of neurons. Our face clustering task learns a mapping function $F: X \rightarrow \mathbb{R}^d$ that encodes a face image $\mathbf{x}_i$ into a d-dimension feature vector $F(\mathbf{x}_i)$ and preserves identity (ID) consistency, Specifically, we minimize $||F(\mathbf{x}_i) - F(\mathbf{x}_j)||$ if $ID(\mathbf{x}_i) = ID(\mathbf{x}_j)$ and maximize $||F(\mathbf{x}_i) - F(\mathbf{x}_j)||$ otherwise. We propose

$$F(\mathbf{x}) = NORM(FC5(S(\mathbf{x}))), \tag{1}$$

where $FC5$ is a fully connected layer and $NORM$ is an l_2 normalization layer, inspired by FaceNet [44], to achieve higher recognition performance. To train $F(\cdot)$, we use the triplet loss, which defines a image triplet $(\mathbf{x}_i^a, \mathbf{x}_i^p, \mathbf{x}_i^n)$ that $\mathbf{x}_i^a$ is an image serving as an anchor and $\mathbf{x}_i^p$ and $\mathbf{x}_i^n$ are two images with a positive and negative relationship in terms of identities between $\mathbf{x}_i^a$, i.e. $\mathbf{x}_i^a \neq \mathbf{x}_i^p$ and $ID(\mathbf{x}_i^a) = ID(\mathbf{x}_i^p)$, but $ID(\mathbf{x}_i^a) \neq ID(\mathbf{x}_i^n)$. Because the amount of triplets available from X is in the cubic order of $|X|$, we adopt a scheme of hard negative mining to select the most effective triplets to train the CNN model through back-propagation. As illustrated in Fig. 3, we use a hyper-parameter r to control the ratio of hard samples. For each identity in our training set, we randomly select a pair of $\mathbf{x}_i^a$ and $\mathbf{x}_i^p$, then generate the feature vector $F(\mathbf{x}_i^a)$ to find the hard samples by navigating all negative images and picking the portion which generates the large losses, and randomly choose negative images from the remaining portion. By setting up a proper hyper-parameter r as the ratio of the hard negative portion to a training batch, the scheme ensures the balance between training cost and generality because the most decisive samples will be first used and all other samples will still be picked with an even probability. The triplet loss is defined as

$$\begin{aligned} L_{ID}(X) = \sum_{(\mathbf{x}_i^a, \mathbf{x}_i^p, \mathbf{x}_i^n) \in T} & max(0, k + ||F(\mathbf{x}_i^a) - F(\mathbf{x}_i^p)||_2^2 \\ & - ||F(\mathbf{x}_i^a) - F(\mathbf{x}_i^n)||_2^2), \end{aligned} \tag{2}$$

where k is a given margin value to reduce over-sensitivity, and T is the set of triplets selected by hard negative mining.

AU Detection. We use the feature vectors extracted from the face clustering task as the offsets for translation transformation in the AU-detection task, as illustrated in Fig. 1. The insight under this design is that the change on facial appearance is highly structured as it is caused by a specific AU. Therefore, we detect AUs in the domain of transformed features rather than the domain of features on its own. We compute the transformed features through identity-specific translation transformation that subtracts the identity-dependent features. Let $G: X \rightarrow \mathbb{R}^c$ be the mapping function intended for the AU detection task. In order to train the CNN parameters in G, we use the sigmoid cross entropy loss

$$\begin{aligned} L_{AU}(X,Y) = & \frac{-1}{N \times c} \sum_{i=1}^{N} \sum_{j=1}^{c} [\mathbf{y}_i^j = 1] \log(G(\mathbf{x}_i)^j) \\ & + [\mathbf{y}_i^j = 0] \log(1 - G(\mathbf{x}_i)^j), \end{aligned} \tag{3}$$

where $[x]$ is an indicator function returning 1 if the statement x is true, and 0 otherwise, and j indicates the element index of the feature vectors $\mathbf{y}_i$ and $G(\mathbf{x}_i)$.

To subtract a vector from the vector generated from the face clustering task $F(\cdot)$, their dimension need to be matched

Authorized licensed use limited to: National Taiwan University. Downloaded on May 29,2023 at 07:53:18 UTC from IEEE Xplore. Restrictions apply.

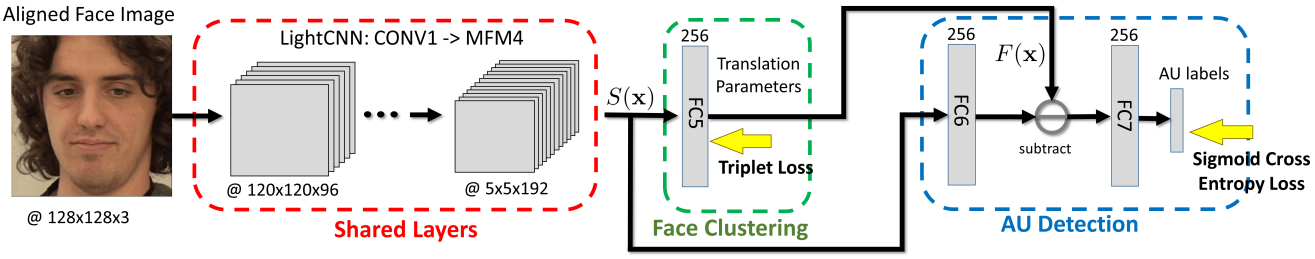
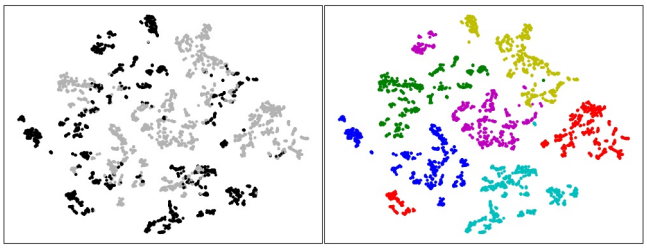
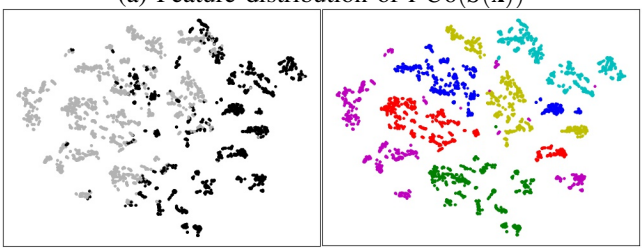


Fig. 1. Overview of the proposed IdenNet. The method is implemented in an architecture of multi-task network cascades where the two sub-tasks, face clustering and AU detection, share a common network and own their specific CNN layers.



(a) Feature distribution of $FC6(S(\mathbf{x}))$



(b) Feature distribution of $FC6(S(\mathbf{x})) - F(\mathbf{x})$

Fig. 2. Illustration of identity subtraction for AU detection. We use t-SNE to visualize the feature distribution of 6000 images (1000 images per subject). The colors on the left and right sides indicate AU12 labels and subjects' identities. (a) Because of the face clustering task, $FC6(S(\mathbf{x}))$ features cluster together primarily by their identities and secondarily by their AU12 labels. (b) The $FC6(S(\mathbf{x})) - F(\mathbf{x})$ features benefit from identity subtraction and organize those images into two groups based on the AU12 labels. Therefore, the features better separate the existence of AUs.

so we apply a fully connected layer on the feature extracted from the shared layers $S(\cdot)$ as the manner for $F(\cdot)$, to reduce its dimension to d. Thereafter, to predict AUs from the translated feature vector, we apply another fully connected layer and a sigmoid layer in the same way as ROINet to formulate

$$G(\mathbf{x}) = \sigma\Big(FC7\big(FC6(S(\mathbf{x})) - F(\mathbf{x})\big)\Big), \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid function. We use a single fully connected layer for classification because we want to keep our model simple to evaluate the performance of the idea of exploiting identity information. It is possible to use a complicated model to improve the classification accuracy, but also likely to make the impact of our idea less clear.

Optimization. To train IdenNet, we combine the two loss functions by

$$L(X,Y) = \alpha L_{ID}(X) + (1-\alpha)L_{AU}(X,Y), \tag{5}$$

where α is a weight balancing the losses of the two sub-tasks, and we repeatedly adjust the weight upon the two

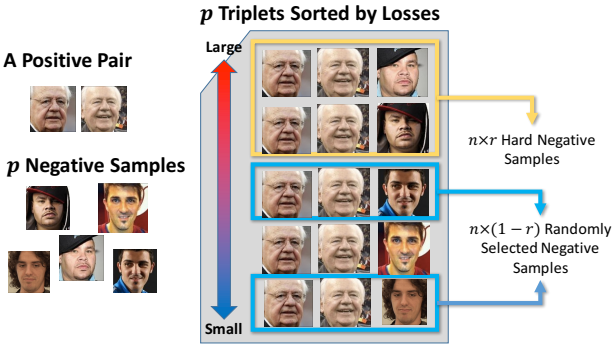


Fig. 3. Illustration of hard negative mining. Given two images from the same subject to form a positive pair and p images from subjects different from the one of the positive pair, let n be the number of negative images used by a positive pair to form triplets in a training batch, and r be the hyper-parameter of the ratio of hard samples. We first pick the leading $n \times r$ out of the p negative samples which generate large losses, and then randomly select $n \times (1-r)$ instances from the remaining negative samples.

different types of training batches we used to generate an effective cascades model because we use both types of AU- and ID-annotated images.

Our training process contains two stages: First, we train $S(\cdot)$ and $F(\cdot)$ on an ID-annotated dataset using $\alpha = 1.0$ because AU labels are unavailable. Second, we train the whole cascades in an end-to-end manner using both AU- and ID-annotated datasets. We set α as 0.5 for batches which are generated from images of an AU-annotated dataset because both identity and AU labels are available, and α as 1 for batches from an ID-annotated dataset.

IV. EXPERIMENTS

We evaluate our method on the widely used BP4D, UNBC-McMaster and DISFA datasets and use CelebA as the ID-annotated dataset. To investigate the capability of generalization, we conduct experiments for cross-dataset scenarios where models are trained on BP4D and applied on UNBC-McMaster and DISFA directly without additional optimization processes. For a fair comparison, we use the same F1-frame as existing methods [25], [18], [22], which is a frame-based F1 score, defined as the harmonic mean of precision and recall and averaged by all frames. In the preprocessing stage, we align images at eyes and resize the resolution to 128×128 pixels. IdenNet's architecture is shown in Table I. We initialize $S(\cdot)$ and $F(\cdot)$ (CONV1 to NORM) using LightCNN's pre-trained weights model1, and

Authorized licensed use limited to: National Taiwan University. Downloaded on May 29,2023 at 07:53:18 UTC from IEEE Xplore. Restrictions apply.

TABLE I

ARCHITECTURE OF THE PROPOSED MULTI-TASK NETWORK CASCADES

Layer Type	Filter Size /Stride	Output Size	#Params
Shared Layers (Input: Image $128 \times 128 \times 3$)			
CONV1	9×9 / 1	$120 \times 120 \times 96$	23.3K
POOL1	2×2 / 2	$60 \times 60 \times 96$	-
MFM1	-	$60 \times 60 \times 48$	-
CONV2	5×5 / 1	$56 \times 56 \times 192$	230.4K
POOL2	2×2 / 2	$28 \times 28 \times 192$	-
MFM2	-	$28 \times 28 \times 96$	-
CONV3	5×5 / 1	$24 \times 24 \times 256$	614.4K
POOL3	2×2 / 2	$12 \times 12 \times 256$	-
MFM3	-	$12 \times 12 \times 128$	-
CONV4	4×4 / 1	$9 \times 9 \times 384$	786.4K
POOL4	2×2 / 2	$5 \times 5 \times 384$	-
MFM4	-	$5 \times 5 \times 192$	-
Face Clustering (Input: MFM4)			
FC5	-	512	2,457.6K
MFM5	-	256	-
NORM	-	256	-
AU Detection (Input: MFM4)			
FC6	-	512	2,457.6K
MFM6	-	256	-
SUBTRACT (MFM6-NORM)	-	256	-
FC7	-	12	3K
Total	-	-	6,572.8K

TABLE II

COMPARISON ON MODEL SIZE AND SPEED

Method	#Params	Frames per second
DRML	56,855.1K	82.52
ROINet	>12,351.1K	N.A.
E-Net+ATF	>12,351.1K	N.A.
SVTPT	—	2.00
IdenNet	**6,572.8K**	**155.06**

layers from FC6 to FC7 using random numbers. IdenNet's size and execution time are compared in Table II.

Because the code of ROINet and E-Net+ATF is not released, their execution time is unable to report and their model sizes are conservatively calculated from their feature extractors. SVTPT is not a CNN-based method so that we do not measure its number of parameters. We empirically set the hyperparameters k, n, r, and d as 0.5, 4, 0.5 and 256, and optimize the model using Adam optimizer [48] with a learning rate 0.0001 and batch size 64. We generate a set of 64 image triplets as a training batch by randomly selecting 16 identities at first. Then we make 16 positive pairs by randomly selecting the anchor and positive images $\mathbf{x}_i^a$ and $\mathbf{x}_i^p$. For each positive pair, we compute the feature vector $F(\mathbf{x}_i^a)$ and feature vectors of all negative samples $F(\mathbf{x}_i^n)$ to do the hard negative mining. Because we set n as 4 and r as 0.5, 2 negative images are hard sampled and the other 2 negative images are randomly selected. We use Caffe [49] to conduct all experiments on a machine with an Intel i7 3.4 GHz CPU, 32GB memory, and an NVidia GPU Titan X. Our code is available at `https://github.com/andytu455176/IdenNet`.

Within-dataset Scenarios. In this scenario, we train and apply IdenNet using the same AU-annotated dataset. For a fair comparison, we use the same splits used by ROINet on BP4D in our experiments and the released code of DRML and SVTPT from their project websites. On DISFA and UNBC-McMaster, we randomly split the datasets into 3 folds which contain mutually exclusive sets of subjects. Although the compared method AlexNet [9] is not originally designed for AU detection, it is a representative method for object recognition, working as a baseline in our experiments. Please note that there are two variants of ROINet, named R-T1 and R-T2 [22], which generate higher f1-frames over ROINet on the BP4D dataset. However, they require a different format of input data—video—because their architecture contains LSTM [50] which exploits temporal consistency but cannot work for still images. For ATF, we compare IdenNet with the best model in [23], E-Net+ATF, and E-Net denotes the CNN with enhancing layers proposed in [51]. Since we use LightCNN's structure as our shared layers and its pre-trained model as our initial weights, we conduct the comparison against it as a baseline.

TABLE III

F1-FRAME ON THE BP4D DATASET WITH 3-FOLD RANDOM SPLITS.

AU	AlexNet	DRML	SVTPT	LightCNN	E-Net+ATF	ROINet	IdenNet
1	40.4	36.4	39.3	44.0	39.2	36.2	**50.5**
2	26.0	**41.8**	34.9	32.9	35.2	31.6	35.9
4	40.8	43.0	37.5	49.8	45.9	43.4	**50.6**
6	69.0	55.0	64.7	74.8	71.6	77.1	**77.2**
7	66.0	67.0	72.4	72.9	71.9	73.7	**74.2**
10	77.8	66.3	75.0	80.0	79.0	**85.0**	82.9
12	81.7	65.8	79.6	83.6	83.7	**87.0**	85.1
14	51.8	54.1	48.2	58.2	**65.5**	62.6	63.0
15	23.5	33.2	39.2	29.9	33.8	**45.7**	42.2
17	51.8	48.0	57.7	55.2	60.0	58.0	**60.8**
23	25.6	31.7	33.0	31.3	37.3	38.3	**42.1**
24	34.2	30.0	40.4	36.2	41.8	37.4	**46.5**
Ave	49.1	48.3	51.8	54.1	55.4	56.4	**59.3**

TABLE IV

F1-FRAME ON THE DISFA DATASET WITH 3-FOLD RANDOM SPLITS.

AU	AlexNet	DRML	SVTPT	LightCNN	E-Net+ATF	ROINet	IdenNet
1	6.9	17.3	14.9	19.4	**45.2**	41.5	25.5
2	6.6	17.7	19.1	16.5	**39.7**	26.4	34.8
4	39.4	37.4	41.4	45.8	47.1	**66.4**	64.5
6	38.4	29.0	43.2	40.2	48.6	**50.7**	45.2
9	28.9	10.7	25.1	28.5	32.0	8.5	**44.6**
12	54.4	37.7	67.9	70.9	55.0	**89.3**	70.7
25	58.2	38.5	64.6	74.4	86.4	**88.9**	81.0
26	40.3	20.1	35.0	51.3	39.2	15.6	**55.0**
Ave	35.4	26.7	38.9	43.4	49.2	48.5	**52.6**

As shown in Tables III and V, IdenNet generates the best F1 scores on 9 out of 12 AUs on BP4D, 4 out of 6 on UNBC-McMaster, and the best average F1 scores on the three datasets as shown in Table IV. The improvement over LightCNN on the three datasets shows the effectiveness of the proposed face clustering task for AU detection and we attribute it to two factors. First, images with similar facial appearances are likely to share similar AU responses, and we exploit this property through the face clustering task. Second, although CelebA and the three AU-annotated datasets are compiled to serve different purposes, all of their images belong to the same domain, i.e. human faces, and IdenNet takes advantage of the commonalty to train its CNN model.

Cross-dataset Scenarios. To verify IdenNet's generalization

Authorized licensed use limited to: National Taiwan University. Downloaded on May 29,2023 at 07:53:18 UTC from IEEE Xplore. Restrictions apply.

TABLE V

F1-FRAME ON UNBC-MCMASTER WITH 3-FOLD RANDOM SPLITS.

AU	AlexNet	DRML	SVTPT	LightCNN	IdenNet
6	27.0	29.3	10.3	36.3	**53.9**
7	8.0	15.3	1.4	18.0	**22.3**
12	32.6	35.5	14.1	34.0	**48.5**
25	1.7	6.4	**9.0**	7.1	4.7
26	4.7	0.8	4.7	**5.7**	3.5
43	9.5	17.9	7.4	27.8	**37.8**
Ave	13.9	17.5	7.8	21.5	**28.5**

TABLE VI

F1-FRAME ON DISFA USING MODELS TRAINED ON BP4D.

AU	AlexNet	DRML	SVTPT	LightCNN	IdenNet
1	13.4	11.6	12.4	13.0	**20.1**
2	10.2	4.7	11.2	8.2	**25.5**
4	27.5	32.5	13.1	36.5	**37.3**
6	33.7	32.8	25.9	41.3	**49.6**
12	47.4	49.7	44.3	53.7	**66.1**
Ave	26.4	26.3	21.4	30.6	**39.7**

ability, we conduct experiments under two cross-dataset scenarios: to train a model on the BP4D dataset and apply it to the UNBC-McMaster and DISFA datasets without any fine-tuning. Unlike cross-dataset experiments reported in previous work [18], [22] that require additional optimization processes on new datasets, ours are more challenging practical because we do not exploit any prior knowledge of the test datasets. Since the AUs labeled in the three datasets are different, we evaluate as many AUs in common as possible.

As shown in Tables VI and VII, IdenNet achieves the best performance, leaving a margin over existing methods. The reason is the dissimilarity between the training and test datasets. All subjects in the BP4D dataset are young adults at the ages less than 30, but those in UNBC-McMaster are the middle-aged who suffer from shoulder pain. The dissimilarity makes DRML and LightCNN, which adopt the direct inductive-learning approach, generate low performance by applying learned rules on new data. On the other hand, SVTPT, the method belonging to the transductive learning approach by adopting predefined kernels to measure subject similarity, makes mistakes because the test samples are highly different from the training samples. In contrast, IdenNet contains a sub-network specialized in identity subtraction and trained from a large dataset containing numerous subjects at various ages, so IdenNet extracts effective features and generates robust predictions.

For the DISFA dataset, although its subjects are similar to BP4D in terms of age, its environment setting is definitely dissimilar, especially for lighting. The light appearance is cool white in DISFA but warm yellow in BP4D, and the light type is spotted in DISFA but ambient in BP4D. In addition, a significant light source is set up on top of subjects only in DISFA, which results in shiny hairs and shadowy cheeks.

Because images of BP4D and DISFA are highly dissimilar, the cross-dataset experimental F1 scores shown in Table VI are significantly worse than those generated under the within-dataset scenario shown in Table IV in respect to the same AUs, i.e. units 1, 2, 4, 6, and 12. To measure the performance between within- and cross-dataset scenarios, we define the drop rate as $\frac{1}{K}\sum_{i=1}^{K}\frac{f_i^o - f_i^n}{f_i^o}$, where f is the F1-frame, o and n denotes the original (training) and new (test) datasets, respectively, and i is the index of K AUs shared by o and n. IdenNet generates the smallest drop rate of 33%, while other methods generate drop rates greater than 46% for the case that o is BP4D and n is DISFA. For another case that o is BP4D and n is UNBC-McMaster, the drop is more severe because the subjects of the two datasets are highly different. IdenNet still generates the smallest drop rate of 72%, while other methods generate drop rates greater than 74%.

TABLE VII

F1-FRAME ON UNBC-MCMASTER USING MODELS TRAINED ON BP4D.

AU	AlexNet	DRML	SVTPT	LightCNN	IdenNet
4	3.7	4.6	6.1	6.3	**9.7**
6	25.4	24.6	20.6	28.1	**33.0**
7	14.8	13.4	11.6	12.4	**12.8**
10	4.5	2.7	2.0	**2.8**	**2.8**
12	27.9	31.0	25.4	29.0	**42.1**
Ave	15.3	15.3	13.1	15.7	**20.1**

Ablation Study on the Normalization Layer. We conduct an ablation study on the *NORM* layer in Eq. 1. Without the layer, IdneNet's F1-frame on the BP4D, DISFA and UNBC-McMaster within-dataset scenario drops from 59.3 to 58.0, 52.6 to 47.8, and 28.5 to 23.9, respectively. The reason is that LightCNN adopts cosine similarity as the metric to evaluate the similarity among faces. IdenNet uses LightCNN's pre-trained weights, but adopts Euclidean distance as the metric instead of cosine similarity. Therefore, the normalization layer reorganizes the output features and helps improve the face clustering performance.

V. CONCLUSIONS AND FUTURE STUDY

In this paper, we propose IdenNet, a novel method for AU detection by exploiting identity information and a large number of ID-annotated training images. IdenNet extracts identity-dependent features in the face clustering task and normalize them in the AU detection network. After reducing the bias caused by individual subjects, the identity-subtracted features better present the differences generated by AUs. Experiments conducted under both within- and cross-dataset scenarios on benchmark datasets validate the effectiveness and robustness of the proposed method.

We utilize the simple translation transformation for identity subtraction and find that it is effective to reduce individual bias. It remains an open question whether a complicated transformation such as the affine transformation will make the network more powerful. In addition, the approach reported in [22] to improve ROINet's performance by adopting LSTM to exploit temporal consistency is worthy of comparison because IdenNet is capable of incorporating LSTM to reduce the differences of temporal AU changes among subjects.

Authorized licensed use limited to: National Taiwan University. Downloaded on May 29,2023 at 07:53:18 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992.

[2] Paul Ekman, Richard J. Davidson, and Wallace V. Friesen. The Duchenne smile: Emotional expression and brain physiology. *Journal of personality and social psychology*, 58(2):342–353, 1990.

[3] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N. Metaxas. Learning active facial patches for expression analysis. In *CVPR*, 2012.

[4] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *CVPR*, 2014.

[5] Paul Ekman. Universals and cultural differences in facial expressions of emotions. *Nebraska Symposium on Motivation*, 19:207–283, 1972.

[6] Sudha Velusamy, Hariprasad Kannan, Balasubramanian Anand, Anshul Sharma, and Bilva Navathe. A method to infer emotions from facial action units. In *ICASSP*, 2011.

[7] Mehdi Ghayoumi and Arvind K Bansal. Unifying geometric features and facial action units for improved performance of facial expression analysis. *arXiv preprint*, 2016.

[8] Paul Ekman. Methods for measuring facial action. In *Handbook of methods in nonverbal behavior research*, chapter 2, pages 45–135. Cambridge University Press, 1982.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[10] Alexander Toshev and Christian Szegedy. DeepPose: Human pose estimation via deep neural networks. In *CVPR*, 2014.

[11] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, 2015.

[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[13] Patrick Lucey, Jeffrey F. Cohn, Kenneth M. Prkachin, Patricia E. Solomon, and Iain Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *FG*, 2011.

[14] Seyed Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. DISFA: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[15] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing*, 10(32):692–706, 2014.

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*, 2015.

[17] Ying-li Tian, Takeo Kanade, and Jefferey F. Cohn. Recognizing action units for facial expression analysis. *PAMI*, 23(2):97–115, 2001.

[18] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *CVPR*, 2016.

[19] Kaili Zhao, Wen-Sheng Chu, and Aleix M. Martinez. Learning facial action units from web images with scalable weakly supervised clustering. In *CVPR*, 2018.

[20] Ying-li Tian, Takeo Kanade, and Jefferey F. Cohn. Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *FG*, 2009.

[21] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F. Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit and holistic expression recognition. *TIP*, 25(8):3931–3946, 2016.

[22] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *CVPR*, 2017.

[23] Zheng Zhang, Shuangfei Zhai, and Lijun Yin. Identity-based adversarial training of deep cnns for facial action unit recognition. In *BMVC*, 2018.

[24] Wen-Sheng Chu, Fernando De la Torre, and Jeffery F. Cohn. Selective transfer machine for personalized facial action unit detection. In *CVPR*, 2013.

[25] Gloria Zen, Enver Sangineto, Elisa Ricci, and Nicu Sebe. Unsupervised domain adaptation for personalized facial emotion recognition. In *ICMI*, 2014.

[26] Jiabei Zeng, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F. Cohn, and Zhang Xiong. Confidence preserving machine for facial action unit detection. In *ICCV*, 2015.

[27] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Classifier learning with prior probabilities for facial action unit recognition. In *CVPR*, 2018.

[28] Guozhu Peng and Shangfei Wang. Weakly supervised facial action unit recognition through adversarial training. In *CVPR*, 2018.

[29] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *CVPR*, 2018.

[30] C. Fabian Benitez-Quiroz, Yan Wang, and Aleix M. Martinez. Recognition of action units in the wild with deep nets and a new global-local loss. In *ICCV*, 2017.

[31] Simon Lucey Abhinav Dhall, Roland Goecke and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. *BEFIT Workshop at ICCV*, 2011.

[32] Nick Haber, Catalin Voss, Azar Fazel, Terry Winograd, and Dennis P. Wall. A practical approach to real-time neutral feature subtraction for facial expression recognition. In *WACV*, 2016.

[33] Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 4DFAB: A large scale 4d database for facial expression analysis and biometric applications. In *CVPR*, 2018.

[34] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. Disentangling factors of variation for facial expression recognition. In *ECCV*, 2012.

[35] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-aware convolutional neural network for facial expression recognition. In *FG*, 2017.

[36] Xiaofeng Liu, B. V. K. Vijaya Kumar, Jane You, and Ping Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *CVPRW*, 2017.

[37] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, 2010.

[38] Tim Valentine. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology*, 43(2):161–204, 1991.

[39] Tim Valentine and M. Endo. Towards an exemplar model of face processing: The effects of race and distinctiveness. *The Quarterly Journal of Experimental Psychology*, 44(4):671–703, 1992.

[40] Michael B. Lewis. Face-space-R: Towards a unified account of face recognition. *Visual Cognition*, 11(1), 2004.

[41] Jonathan Vitale, Mary-Anne Williams, and Benjamin Johnston. The face-space duality hypothesis: a computational model. In *CogSci*, 2016.

[42] Jonathan Vitale, Benjamin Johnston, and Mary-Anne Williams. Facial motor information is sufficient for identity recognition. In *CogSci*, 2017.

[43] Changxing Ding and Dacheng Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on Intelligent Systems and Technology*, 7(3):37:1–37:42, 2016.

[44] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[45] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light CNN for deep face representation with noisy labels. *arXiv preprint*, 2015.

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

[47] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016.

[48] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[49] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014.

[50] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[51] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *FG*, 2017.

Authorized licensed use limited to: National Taiwan University. Downloaded on May 29,2023 at 07:53:18 UTC from IEEE Xplore. Restrictions apply.