Adaptive Deep Metric Learning for Identity-Aware Facial Expression Recognition

Xiaofeng Liu[1,2,4*], B.V.K Vijaya Kumar[1], Jane You[3], Ping Jia[2]
[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA
[2] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science
[3] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[4] University of Chinese Academy of Sciences, Beijing, China
liuxiaofeng@cmu.edu, kumar@ece.cmu.edu, csyjia@comp.polyu.edu.hk, jiap@ciomp.ac.cn

Abstract

A key challenge of facial expression recognition (FER) is to develop effective representations to balance the complex distribution of intra- and inter- class variations. The latest deep convolutional networks proposed for FER are trained by penalizing the misclassification of images via the softmax loss. In this paper, we show that better FER performance can be achieved by combining the deep metric loss and softmax loss in a unified two fully connected layer branches framework via joint optimization. A generalized adaptive (N+M)-tuplet clusters loss function together with the identity-aware hard-negative mining and online positive mining scheme are proposed for identity-invariant FER. It reduces the computational burden of deep metric learning, and alleviates the difficulty of threshold validation and anchor selection. Extensive evaluations demonstrate that our method outperforms many state-of-art approaches on the posed as well as spontaneous facial expression databases.

1. Introduction

Facial expression is one of the most expressive nonverbal communication channels for humans to convey their emotional state [5]. Therefore, automatic facial expression recognition (FER) is important in a wide range of applications including human-computer interaction (HCI), digital entertainment, health care and intelligent robot systems [20].

Researchers have achieved great progress in recognizing the posed facial expressions collected under tightly controlled environment. Since the most promising face-related applications occur in more natural conditions, it is our goal to develop a robust system that can operate well in the real word. Despite the significant efforts, FER remains a challenge in the presence of pose and illumination variations as well as inter-subject variations (i.e., identity-specific attributes) [42]. These identity-specific factors degrade the FER performance of new identities unseen in the training data. Since spontaneous expressions only involve subtle facial muscle movements, the extracted expression-related information from different classes can be dominated by the sharp-contrast identity-specific geometric or appearance features which are not useful for FER. As shown in Fig. 1, example x_1 and x_3 are of happy faces whereas x_2 and x_4 are not of happy faces. $f(x_i)$ are the image representations using the extracted features. For FER, we desire that two face images with the same expression label are close to each other in the feature space, while face images with different expressions are farther apart from each other, i.e., the distance D_2 between examples x_1 and x_3 should be smaller than D_1 and D_3, as in Fig. 1(b). However, the learned expression representations may contain irrelevant identity information as illustrated in Fig. 1(a). Due to large inter-identity variations, D_2 usually has a large value while the D_1 and D_3 are relatively small.

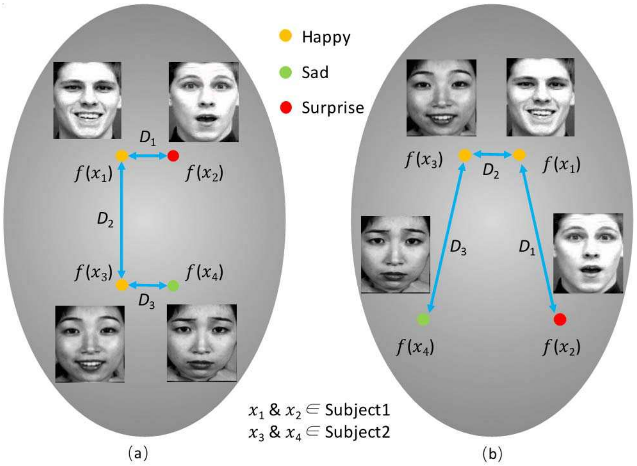

Figure 1. Illustration of representations in feature space learned by (a) existing methods, and (b) the proposed method.

To further improve the discriminating power of the expression feature representations, and address the large intra-subject variation in FER, a potential solution is to incorporate the deep metric learning scheme within a convolutional neural network (CNN) framework. The fundamental philosophy behind the widely-used triplet loss function [7] is to require one positive example closer to the anchor example than one negative example with a fixed gap τ. Thus, during one iteration, the triplet loss ignores the negative examples from the rest of classes. Moreover, one of the two examples from the same class in the triplets can be chosen as

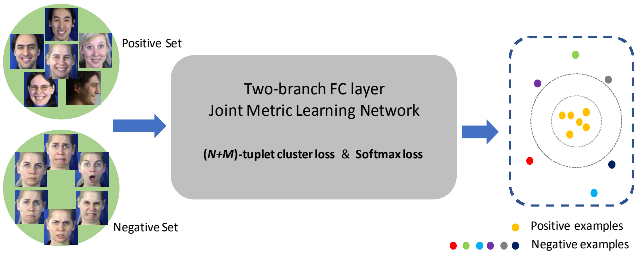
Positive Set
Two-branch FC layer
Joint Metric Learning Network
(*N+M*)-tuplet cluster loss & Softmax loss
Negative Set
● Positive examples
● ● ● ● ● ● Negative examples

Figure 2. Frame work of our facial expression recognition model used for training. The deep convolutional network aims to map the original expression images into a feature space that the images of the same expression tend to form a cluster while other images tend to locate far away.

the anchor point. However, there exist some special cases that the triplet loss function with impropriate anchor may judge falsely, as illustrated in Fig. 3(a). This means the performance is quite sensitive to the anchor selection in the triplets input. We adapted the idea from the (*N*+1)-tuplet loss [37] and coupled clusters loss (CCL) [22] to design a (*N*+*M*)-tuplet clusters loss function which incorporates a negative set with *N* examples and a positive set with *M* examples in a mini-batch. A reference distance *T* is introduced to force the negative examples to move away from the center of positive examples and for the positive examples to simultaneously map into a small cluster around their center c^+. The circles of radius $(T + \frac{\tau}{2})$ and$(T - \frac{\tau}{2})$ centered at the c^+form the boundary of the negative set and positive set respectively, as shown in Fig. 3(d). By doing this, our approach can handle complex distribution of intra- and inter-class variations, and free the anchor selection trouble in conventional deep metric learning methods. Furthermore, the reference distance *T* and the margin τ can be learned adaptively via the propagation in the CNN instead of the manually-set hyper-parameters. We also propose a simple and efficient mini-batch construction scheme that uses different expression images with the same identity as the negative set to avoid the expensive hard-negative example searching, while mining the positive set online. Then, the (*N*+*M*)-tuplet clusters loss guarantees all the discriminating negative samples are efficiently used per update to achieve an identity-invariant FER.

We jointly optimize the softmax loss and (*N*+*M*)-tuplet clusters loss to explore the potential of both the expression labels and identity labels information. Considering the different characteristics of each loss function and their tasks, we design two branches of fully connected (FC) layers, and a connecting layer to balance them. The features extracted by the expression classification branch can be fed to the following metric learning processing. This enables each branch to focus better on their own task without embedding much information of the other. As shown in Fig. 2, the inputs are two facial expression image set: one positive set (images of the same expression from different subjects) and one negative set (images of other expressions with the same identity of the query example). The deep features and distance metrics are learned simultaneously in a network.

The three major contributions in this paper are: 1) We propose a generalized (*N*+*M*)-tuplet clusters loss function with adaptively learned reference threshold which can be seamlessly factorized into a linear-fully connected layer for an end-to-end learning. 2) With the identity-aware negative mining and online positive mining scheme, we learn distance metrics with fewer input passes and distance calculations, without sacrificing the performance for identity-invariant FER. 3) We optimize the softmax loss and (*N*+*M*)-tuplet clusters loss jointly in a unified two-branch FC layer metric learning CNN framework based on their characteristics and tasks. In experiments, we demonstrate that the proposed method achieves promising results not only outperforming several state-of-art approaches in posed facial expression dataset (e.g., CK+, MMI), but also in spontaneous facial expression dataset (namely, SFEW).

2. Related work

FER focus on the classification of seven basic facial expressions which are considered to be common among humans [40]. Much progress has been made on extracting a set of features to represent the facial images [13]. Geometric representations utilize the shape or relationship between facial landmarks. However, they are sensitive to the facial landmark misalignments [35]. On the other hand, appearance features, such as Gabor filters, Scale Invariant Feature Transform (SIFT), Local Binary Patterns (LBP), Local Phase Quantization (LPQ), Histogram of Oriented Gradients (HOG) and the combination of these features via multiple kernel learning are usually used for representing facial textures [3, 15, 49, 51]. Some methods such as active appearance models (AAM) [41] combine the geometric and appearance representations to provide better spatial information. For a comprehensive survey, we refer readers to [34]. Due to the limitations of handcrafted filters, extracting purely expression-related features is difficult.

The developments in deep learning, especially the success of CNN, have made high-accuracy image classification possible in recent years. It has also been shown that carefully designed neural network architectures perform well in FER [29]. Despite its popularity, current softmax loss-based network does not explicitly encourage intra-class compactness and inter-class separation. The emerging deep metric learning methods have been investigated for person recognition and vehicle re-identification problems with large intra-class variations, which suggests that deep metric learning may offer more pertinent representations for FER. Compared to traditional distance metric learning, deep metric learning learns a nonlinear embedding of the data using the deep neural networks. The initial work is to train a Siamese network with contrastive loss function [4]. The pairwise examples are fed into two symmetric sub-networks to predict whether they are from the same class. Without the

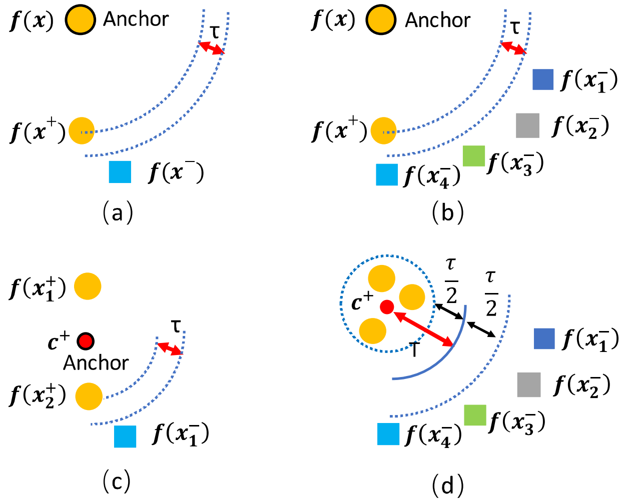

Figure 3. Failed case of (a) triplet loss, (b) (*N*+1)-tuplet loss, and (c) Coupled clusters loss. The proposed (*N*+*M*)-tuplet clusters loss is illustrated in (d).

interactions of positive pairs and negative pairs, the Siamese network may fail to learn effective metrics in the presence of large intra- and inter-class variations. One improvement is the triplet loss approach [7], which achieved promising performance in both re-identification and face recognition problems. The inputs are triplets, each consisting of a query, a positive example and a negative example. Specifically, it forces the difference of the distance from the anchor point to the positive example and from the anchor point to the negative example to be larger than a fixed margin τ. Recently, some of its variations with faster and stable convergence have been developed. The most similar model of our proposed method is the (*N*+1)-tuplet loss [37]. We use $x^+$and $x^-$to denote the positive and negative examples of a query example x, meaning that $x^+$is the same class of x, while $x^-$is not. Considering (*N*+1) tuplet which includes x, x^+ and *N*-1 negative examples $\{x_j^-\}_{j=1}^{N-1}$, the loss is:

$$L\left(x, x^+, \{x_j^-\}_{j=1}^{N-1}; f\right) = \\ log\left(1 + \sum_{j=1}^{N-1} \exp(D(f, f^+) + \tau - D(f, f_j^-))\right) \qquad (1)$$

where $f(\cdot)$ is an embedding kernel defined by the CNN, which takes x and generates an embedding vector $f(x)$. We write it as f for simplicity, with f inheriting all superscripts and subscripts. $D(\cdot,\cdot)$ is defined as the Mahalanobis or Euclidean distance according to different implementations. The philosophy in this paper also shares commonality with the coupled clusters loss [22], in which the positive example center c^+ is set as the anchor. By comparing each example with this center instead of each other mutually, the evaluation times in a mini-batch are largely reduced.

Despite their wide use, the above-mentioned frameworks still suffer from the expensive example mining to provide nontrivial pairs or triplets, and poor local optima. In practice, generating all possible pairs or triplets would result in quadratic and cubic complexity, respectively and the most of these pairs or triplets are less valuable in the training phase. Also, the online or offline traditional mini-batch sample selection is a large additional burden. Moreover, as shown in Fig. 3(a), (b) and (c), all of them are sensitive to the anchor point selection when the intra- and inter-class variations are large. The triplet loss, (*N*+1)-tuplet loss and CCL are 0, since the distances between the anchor and positive examples are indeed smaller than the distance between the anchor and negative examples for a margin τ. This means the loss function will neglect these cases during the back propagation. We need much more input passes with properly selected anchors to correct it. The fixed threshold in the contrastive loss was also proven to be sub-optimal for it failed to adapt to the local structure of data. Li et al. proposed [21] to address this issue by learning a linear SVM in a new feature space. Some works [9, 43] used shrinkage-expansion adaptive constraints for pair-wise input, which optimized by alternating between SVM training and projection on the cone of all positive semidefinite (PSD) matrices, but their mechanism cannot be implemented directly in deep learning.

A recent study presented objective comparisons between the softmax loss and deep metric learning loss and showed that they could be complementary to each other [12]. Therefore, an intuitive approach for improvement is combining the classification and similarity constraints to form a joint CNN learning framework. For example, [39, 47] combining the contrastive and softmax losses together to achieve a better performance, while [52] proposed to combine triplet and softmax loss via joint optimization. These models improve traditional CNN with softmax loss because similarity constraints might augment the information for training the network. The difficult learning objective can also effectively avoid overfitting. However, all these strategies apply the similarity as well as classification constraints directly on the last FC layer, so that harder tasks cannot be assigned to deeper layers, (i.e., more weights) and interactions between constraints are implicit and uncontrollable. Normally, the softmax loss converges much faster than the deep metric learning loss in multi-task networks. This situation has motivated us to construct a unified CNN framework to learn this two loss function simultaneously in a more reasonable way.

3. (*N*+*M*)-tuplet clusters loss

We give a simple description of our intuition to introduce a reference distance *T* to control the relative boundary ($T-\frac{\tau}{2}$) and ($T+\frac{\tau}{2}$) for the positive and negative examples respectively, as shown in Fig. 3(d). We rewrite the (*N*+1)-tuplet loss function in Eq.(1) as follows:

$$L\left(x,x^{+},\{x_j^{-}\}_{j=1}^{N-1};f\right)=$$
$$log\left(1+\sum_{j=1}^{N-1}\exp(D(f,f^{+})+\left(-T+\frac{\tau}{2}+T+\ \frac{\tau}{2}\right)-D(f,f_j^{-}))\right)$$
$$=log\big(1+\sum_{j=1}^{N-1}\exp\big(D(f,f^{+})-T+\ \tfrac{\tau}{2}\big)*\exp\ (T+\ \tfrac{\tau}{2}-D(f,f_j^{-}))\big)\big) \quad (2)$$

Indeed, the $\exp\big(\mathrm{D}(f,f^{+})\text{-T}+\frac{\tau}{2}\big)$ term used to pull the positive example together and the $\exp\big(\text{T-}\frac{\tau}{2}+\mathrm{D}(f,f_j^{-})\big)$ term used to push the negative examples away have an “OR” relationship. The relatively large negative distance will make the loss function ignore the large absolute positive distance. One way to alleviate large intra-class variations is to construct an “AND” function for these two terms.

We also extend the triplet loss to incorporate *N* negative examples and *M* negative examples. Considering a multi-classification problem, the triplet loss and CCL only compare the query example with one negative example, which only guarantees the embedding vector of the query one to be far from a selected negative class instead of every class. The expectation of these methods is that the final distance metrics will be balanced after sufficient number of iterations. However, towards the end of the training, individual iteration may exhibit zero errors due to the lack of discriminative negative examples causing the iterations to be unstable or slow in convergence.

The identity labels in FER database largely facilitate the hard-negative mining to alleviate the effect of the inter-subject variations. In practice, for a query example, we compose its negative set with all the different expression images of the same person. Moreover, randomly choosing one or a group of positive examples is a paradigm of the conventional deep metric methods, but some extremely hard positive examples may distort the manifold and force the model to be over-fitting. In the case of spontaneous FER, the expression label may erroneously be assigned due to the subjectivity or varied expertise of the annotators [2, 50]. Thus, an efficient online mining for *M* randomly-chosen positive examples should be designed for large intra-class variation datasets. We find the nearest negative example and ignore those positive examples with a larger distance. Algorithm 1 shows the detail. In summary, the new loss function is expressed as follows:

$$L\left(\{x_i^{+}\}_{i=1}^{M},\{x_j^{-}\}_{j=1}^{N};f\right)=\frac{1}{M*}\sum_{i=1}^{M*}\max(0,D(f^{+},c^{+})-T+\ \tfrac{\tau}{2})$$
$$+\tfrac{1}{N}\sum_{j=1}^{N}\max(0,T+\ \tfrac{\tau}{2}-D(f_j^{-},c^{+}))) \quad (3)$$

The simplified geometric interpretation is illustrated in Fig. 3(d). Only if the distances from online mined positive examples to the updated c^{+} smaller than $(T-\frac{\tau}{2})$ and the distances to the updated c^{+} than $(T+\frac{\tau}{2})$, the loss can get a zero value. This is much more consistent with the principle used by many data cluster and discriminative analysis methods. One can see that the conventional triplet loss and its variations become the special cases of the (*N*+*M*)-tuplet clusters loss under our framework.

Algorithm 1 Online positive mining

```
Input
  query example and its randomly selected
  positive set $\{x_i^{+}\}_{i=1}^{M}$, and negative set$\{x_j^{-}\}_{j=1}^{N}$
 1. map examples to feature plane with CNN to get:
        $\{f_i^{+}\}_{i=1}^{M}\ and\ \{f_j^{-}\}_{j=1}^{N}$
 2. calculate the positive cluster center $c^{+}=\frac{1}{M}\sum_{i=1}^{M}f_i^{+}$
 3. calculate the distance from $c^{+}$ to each
    positive and negative example $D\big(f_i^{+},c^{+}\big)$, $D\left(f_j^{-},c^{+}\right)$
 4. search for the nearest negative distance:
        $D\ (x_{nearst}^{-},c^{+})$
 5. ignore those positive examples satisfying:
        $D\big(f_i^{+},c^{+}\big)>D\ (x_{nearst}^{-},c^{+})$
 6. update $c^{+}=\frac{1}{M*}\sum_{i=1}^{M*}f_i^{+}$
Output
  Online mined $M*$ positive examples and updated $c^{+}$
```

For a batch consisting of *X* queries, the input passes required to evaluate the necessary embedding feature vectors in our application are *X*, and the total number of distance calculations can be $2(N+M)*X$. Normally, the *N* and *M* are much smaller than *X*. In contrast, triplet loss requires C_X^3 passes and $2\mathrm{C}_X^3$ times calculations, (*N*+1)-tuplet loss requires $(X+1)*X$ passes and $(X+1)*X^2$ times calculations. Even for a dataset with a moderate size, it is intractable to load all possible meaningful triplets into the limited memory for model training.

By assigning different values for T and τ, we define a flexible learning task with adjustable difficulty for the network. However, the two hyper-parameters need manual tuning and validation. In the spirit of adaptive metric learning for SVM [21], we formulate the reference distance to be a function $T(\cdot,\cdot)$ related with each example instead of a constant. Since the Mahalanobis distance matrix **M** in Eq.(4) itself is quadratic, and can be calculated automatically via a linear fully connected layer as in [36], we assume $T(f_1,f_2)$ as a simple quadratic form, i.e., $T(f_1,f_2)=\frac{1}{2}z^t\mathbf{Q}z+\omega^t\mathrm{z}+b$, where $z^t=\left[{f_1}^t{f_2}^t\right]\in\mathbb{R}^{2d}$, $\mathbf{Q}=\begin{bmatrix}\mathbf{Q}_{f_1f_1} & \mathbf{Q}_{f_1f_2}\\ \mathbf{Q}_{f_2f_1} & \mathbf{Q}_{f_2f_2}\end{bmatrix}\in\mathbb{R}^{2d\times2d}$, $\omega^t=\left[\omega_{f_1}^t\omega_{f_2}^t\right]\in\mathbb{R}^{2d}$, $b\in\mathbb{R}$, f_1 and $f_2\in\mathbb{R}^{2d}$ are the representations of two images in the feature space.

$$D(f_1,f_2)=\|f_1-f_2\|_M^2=(f_1-f_2)^T\mathbf{M}(f_1-f_2) \quad (4)$$

Due to the symmetry property with respect to f_1 and f_2, we can rewrite $T(f_1,f_2)$ as follows:

$$T(f_1,f_2)=\tfrac{1}{2}{f_1}^t\widetilde{\mathbf{A}}f_1+\tfrac{1}{2}{f_2}^t\widetilde{\mathbf{A}}n+{f_1}^t\widetilde{\mathbf{B}}f_2+c^t(f_1+f_2)+b \quad (5)$$

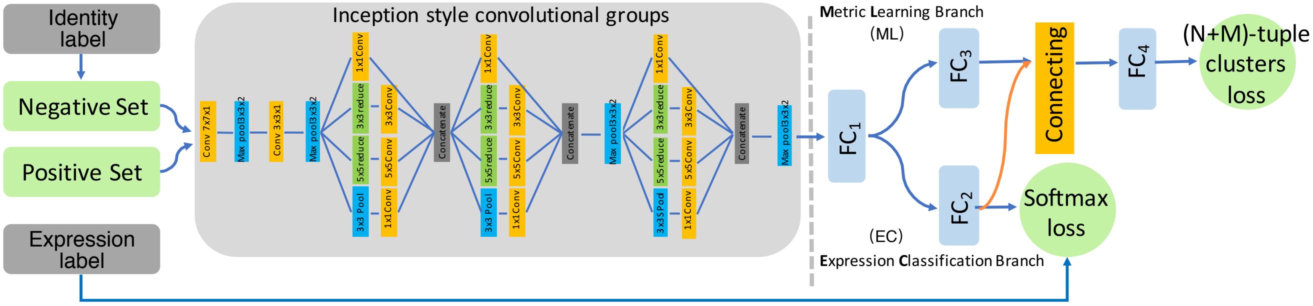

Figure 4. The proposed network structure. In the testing phase, only the convolutional groups and expression classification branch with softmax are used to recognize a single facial expression image.

where $\widetilde{\mathbf{A}}= \mathbf{Q}_{f_1 f_1}= \mathbf{Q}_{f_2 f_2}$ and $\widetilde{\mathbf{B}} = \mathbf{Q}_{f_1 f_2}= \mathbf{Q}_{f_2 f_1}$ are both the $d \times d$ real symmetric matrices (not necessarily positive semi-definite), c = $\omega_{f_1} = \omega_{f_2}$ is a d-dimensional vector, and b is the bias term. Then, a new quadratic formula $H(f_1,f_2)=T(f_1,f_2) - D(f_1,f_2)$ is defined to combine the reference distance function and distance metric function. Substituting Eq.(4) and Eq.(5) to $H(f_1,f_2)$, we get:

$$H(f_1,f_2)=\tfrac{1}{2}{f_1}^t(\widetilde{\mathbf{A}}-2\mathbf{M})f_1+\tfrac{1}{2}{f_2}^t(\widetilde{\mathbf{A}}-2\mathbf{M})f_2 \\ +{f_1}^t(\widetilde{\mathbf{B}}+2\mathbf{M})f_2+c^t(f_1+f_2)+b \qquad (6)$$

$$H(f_1,f_2)=\tfrac{1}{2}{f_1}^t\mathbf{A}f_1+\tfrac{1}{2}{f_2}^t\mathbf{A}f_2+{f_1}^t\mathbf{B}f_2+c^t(f_1+f_2)+b \qquad (7)$$

where $\mathbf{A}=(\widetilde{\mathbf{A}}-2\mathbf{M})$ and $\mathbf{B}=(\widetilde{\mathbf{B}}+2\mathbf{M})$. Suppose $\mathbf{A}$ is positive semi-definite (PSD) and $\mathbf{B}$ is negative semi-definite (NSD), $\mathbf{A}$ and $\mathbf{B}$ can be factorized as $\mathbf{L}_{\mathrm{A}}^T\mathbf{L}_{\mathrm{A}}$ and $\mathbf{L}_{\mathrm{B}}^T\mathbf{L}_{\mathrm{B}}$. Then $H(f_1,f_2)$ can be formulated as follows:

$$H(f_1,f_2)=\tfrac{1}{2}{f_1}^t\mathbf{L}_{\mathrm{A}}^T\mathbf{L}_{\mathrm{A}}f_1+\tfrac{1}{2}n^t\mathbf{L}_{\mathrm{A}}^T\mathbf{L}_{\mathrm{A}}f_2+{f_1}^t\mathbf{L}_{\mathrm{B}}^T\mathbf{L}_{\mathrm{B}}f_2 \\ +c^t(f_1+f_2)+b$$

$$=\tfrac{1}{2}(\mathbf{L}_{\mathrm{A}}f_1)^t(\mathbf{L}_{\mathrm{A}}f_1)+\tfrac{1}{2}(\mathbf{L}_{\mathrm{A}}f_2)^t(\mathbf{L}_{\mathrm{A}}f_2)+(\mathbf{L}_{\mathrm{B}}f_1)^t(\mathbf{L}_{\mathrm{B}}f_2) \\ +c^tf_1+c^tf_2+b \qquad (8)$$

Motivated by the above, we propose a general, computational feasible loss function. Following the notations in the preliminaries and denote $(\mathbf{L}_{\mathrm{A}},\mathbf{L}_{\mathrm{B}},c)^T$ as W, we have:

$$L\left(W,\{x_i^+\}_{i=1}^M,\{x_j^-\}_{j=1}^N;f\right)= \\ \tfrac{1}{M*}\textstyle\sum_{i=1}^{M*}\max(0,H({f_i}^+,c^+)+\tfrac{\tau}{2})+\tfrac{1}{N}\sum_{j=1}^{N}\max(0,H({f_j}^-,c^+)+\tfrac{\tau}{2}) \qquad (9)$$

Given the mined N+M* training examples in a mini-batch, $l(\cdot)$ is a label function. If the example x_k is from the positive set, $l(x_k)=-1$, otherwise, $l(x_k)=1$. Moreover, we simplify the $\frac{\tau}{2}$ to be the constant 1, and changing it to any other positive value results only in the matrices being multiplied by corresponding factors. Our hinge-loss like function is:

$$L\left(W,\{x_i^+\}_{i=1}^M,\{x_j^-\}_{j=1}^N;f\right)= \\ \tfrac{1}{N+M*}\textstyle\sum_{k=1}^{N+M*}\max(0,l(x_k)*H(f_k,c^+)+1) \qquad (10)$$

We optimize Eq.(12) using the standard stochastic gradient descent with momentum. The desired partial derivatives of each example are computed as:

$$\frac{\partial L}{\partial W^l}=\tfrac{1}{N+M*}\textstyle\sum_{k=1}^{N+M*}\frac{\partial L}{\partial X_k^l}\frac{\partial X_k^l}{\partial W^l} \qquad (11)$$

$$\frac{\partial L}{\partial X_k^l}=\frac{\partial L}{\partial X_k^{l+1}}\frac{\partial X_k^{l+1}}{\partial X_k^l} \qquad (12)$$

where X_k^l represents the feature map of the example x_k at the l_{th} layer. Eq.(11) shows that the overall gradient is the sum of the example-based gradients. Eq.(12) shows that the partial derivative of each example with respect to the feature maps can be calculated recursively. So, the gradients of network parameters can be obtained with back propagation algorithm.

In fact, as a straightforward generalization of conventional deep metric learning methods, the (N+M)-tuplet clusters loss can be easily used as a drop-in replacement for the triplet loss and its variations, as well as used in tandem with other performance-boosting approaches and modules, including modified network architectures, pooling functions, data augmentations or activation functions.

4. Network architecture

The proposed two-branch FC layer joint metric learning architecture with softmax loss and (N+M)-tuplet clusters loss, denoted as 2B(N+M)Softmax, is illustrated in Fig. 4. The convolutional groups of our network are based on the inception FER network presented in [28]. We adopt the parametric rectified linear unit (PReLU) to replace the conventional ReLU for its good performance and generalization ability when given limited training data. In addition to providing the sparsity to gain benefits discussed in Arora et al. [1], the inception layer also allows for improved recognition of local features. The locally applied smaller convolution filters seem to align the way that human process emotions with the deformation of local muscles. Note that we did not specifically search for the architectures that obtain the absolute best accuracies on some datasets. Our goal is to confirm our generalized metric learning loss function and the unified two-branch FC layer joint learning framework perform well.

Combing the (N+M)-tuplet clusters loss and softmax loss is an intuitive improvement to reach a better performance. However, conducting them directly on the last FC layer is sub-optimal. The basic idea of building a two-branch FC layers after the deep convolution groups is combining two losses in different level of tasks. We learn the detailed features shared between the same expression class with the expression classification (EC) branch, while exploiting semantic representations via the metric learning (ML) branch to handle the significant appearance changes from different subjects. The connecting layer embeds the information learned from the expression label-based detail task to the identity label-based semantical task, and balances the scale of weights in two task streams. This type of combination can effectively alleviate the interference of identity-specific attributes. The inputs of connecting layer are the output vectors of the former FC layers- FC_2 and FC_3, which have the same dimension denoted as D_{input}. The output of the connecting layer, denoted as FC_4 with dimension D_{ouput}, is the feature vector fed into the second layer of the ML branch. The connecting layer concatenates two input feature vectors into a larger vector and maps it into a D_{output} dimension space:

$$FC_4 = \mathbf{P}^{\mathrm{T}}[FC_2\,; FC_3] = \mathbf{P}_{\mathbf{1}}^{\mathrm{T}} FC_2 + \mathbf{P}_{\mathbf{2}}^{\mathrm{T}} FC_3 \qquad (13)$$

where $\mathbf{P}$ is a $2(D_{\text{input}} \times D_{\text{output}})$ matrix, $\mathbf{P_1}$ and $\mathbf{P_2}$ are $D_{\text{input}} \times D_{\text{output}}$ matrices.

Regarding the sampling strategy, every training image is used as a query example in an epoch. In practice, the softmax loss will only be calculated for the query example. The importance of two loss functions is balanced by a weight α. During the testing stage, this framework takes one facial image as input, and generates the classification result through the EC branch with the softmax loss function.

5. Experiments and analysis

5.1. Preprocessing

For a raw image in the database, face registration is a crucial step for good performance. The bidirectional warping of Active Appearance Model (AAM) [30] and a Supervised Descent Method (SDM) called IntraFace model [45] are used to locate the 49 facial landmarks. Then, face alignment is done to reduce in-plane rotation and crop the region of interest based on the coordinates of these landmarks to a size of 60×60. The limited images of FER datasets is a bottleneck of deep model implementation. Thus, an augmentation procedure is employed to increase the volume of training data and alleviate the chance of over-fitting. We randomly crop the 48×48 size patches, flip them horizontally and transfer them to grayscale images. All the images are processed with the standard histogram equalization and linear plane fitting to remove unbalanced illumination. Finally, we normalize them to a zero mean and unit variance vector. In the testing phase, a single center crop with the size of 48×48 is used as input data.

5.2. Implementation Details

Following the experimental protocol in [28,48], we pre-train our convolutional groups and EC branch FC layers on the FER2013 database [9] for 300 epochs, optimizing the softmax loss using stochastic gradient decent with a momentum of 0.9. The initial network learning rate, batch size, and weight decay parameter are set to 0.1, 128, 0.0001, respectively. If the training loss increased more than 25% or the validation accuracy does not improve for ten epochs, the learning rate is halved and the previous network with the best loss is reloaded. Then the ML branch is added and the whole network is trained by 204,156 frontal viewpoints (-45° to 45°) face images selected from the CMU Multi-pie [10] dataset. There contains 337 people displaying disgust, happy, surprise and neutron. The size of both the positive and negative set are fixed to 3 images. The weights of two loss functions are set equally. We select the highest accuracy training epoch as our pre-trained model.

In the fine-tuning stage, the positive and negative set size are fixed to 6 images (for CK+ and SFEW) or 5 images (for MMI). For a query example, the random searching is employed to select the other 6 (or 5) same expression images to form the positive set. Identity labels are required for negative mining in our method. CK+ and MMI have the subject IDs while the SFEW need manually label. In practice, an off-the-shelf face recognition method can be used to produce this information. When the query example lacks some expression images from the same subject, the corresponding expression images sharing the same ID with the any other positive examples are used. The tuplet-size is set to 12, which means 12×(6+6) =144 (or 12×(5+5) =120) images are fed in each training iteration. We use Adam [19] for stochastic optimization and other hyper-parameters such as learning rate are tuned accordingly via cross-validation. All the CNN architectures are implemented with the widely used deep learning tool “Caffe [14].”

5.3. Experimental Evaluations

To evaluate the effectiveness of the proposed method, extensive experiments have been conducted on three well-known publicly available facial expression databases: CK+, MMI and SFEW. For the fair comparison, we follow the

Table 1. Average confusion matrix obtained from proposed method on the CK+ database [26].

		Predict						
		AN	CO	DI	FE	HA	SA	SU
Actual	AN	**91.1%**	0%	0%	1.1%	0%	7.8%	0%
	CO	5.6%	**90.3%**	0%	2.7%	0%	5.6%	0%
	DI	0%	0%	**100%**	0%	0%	0%	0%
	FE	0%	4%	0%	**98%**	2%	0%	8%
	HA	0%	0%	0%	0%	**100%**	0%	0%
	SA	3.6	0%	0%	1.8%	0%	**94.6%**	0%
	SU	0%	1.2%	0%	0%	0%	0%	**98.8%**

Table 2. Average confusion matrix obtained from proposed method on the MMI database [32].

		Predict					
		AN	DI	FE	HA	SA	SU
Actual	AN	**81.8%**	3%	3%	1.5%	10.6%	0%
	DI	10.9%	**71.9%**	3.1%	4.7%	9.4%	6%
	FE	5.4%	8.9%	**41.4%**	7.1%	7.1%	30.4%
	HA	1.1%	3.6%	0%	**92.9%**	2.4%	0%
	SA	17.2%	7.8%	0%	1.6%	**73.4%**	0%
	SU	7.3%	0%	14.6%	0%	0%	**79.6%**

protocol used by previous works [28,48]. Three baseline methods are employed to demonstrate the superiority of the novel metric learning loss and two-branch FC layer network respectively, i.e., adding the (*N*+*M*)-tuplet clusters loss or (*N*+1)-tuplet loss with softmax loss after the EC branch, denoted as 1B(*N*+1)Softmax or 1B(*N*+*M*)Softmax, and combining the (*N*+1)-tuplet loss with softmax loss via the two-branch FC layer structure, as 2B(*N*+1)Softmax. We do not compare with the triplet loss here, because the number of triplets grows cubically with the number of images, which makes it impractical and inefficient. With randomly selected triplets, the loss failed to converge during the training phase.

The extended Cohn-Kanade database (CK+) [26] includes 327 sequences collected from 118 subjects, ranging from 7 different expressions (i.e., anger, contempt, disgust, fear, happiness, sadness, and surprise). The label is only provided for the last frame (peak frame) of each sequence. We select and label the last three images, and obtain 921 images (without neutral). The final sequence-level predictions are made by selecting the class with the highest possibility of the three images. We split the CK+ database to 8 subsets in a strict subject independent manner, and an 8-fold cross-validation is employed. Data from 6 subsets is used for training and the others are used for validation and testing. The confusions matrix of the proposed method evaluated on the CK+ dataset is reported in Table 1. It can be observed that the disgust and happy expressions are perfectly recognized while the contempt expression is relatively harder for the network because of the limited training examples and subtle muscular movements. As shown in Table 3, the proposed 2B(*N*+*M*)Softmax outperforms the human-crafted feature-based methods, sparse coding-based methods and the other deep learning methods in comparison. Among them, the 3DCNN-DAP, STM-Explet and DTAGN utilized temporal information extracted from sequences. Not surprisingly, it also beats the baseline methods obviously benefit from the combination of novel deep metric learning loss and two-branch architecture.

The MMI database [32] includes 31 subjects with frontal-view faces among 213 image sequences which contain a full temporal pattern of expressions, i.e., from neutral to one of six basic expressions as time goes on, and then released. It is especially favored by the video-based methods to exploit temporal information. We collect three frames in the middle of each image sequence and associate them with the labels, which results in 624 images in our experiments. We divide MMI dataset into 10 subsets for person-independent ten-fold cross validation. The sequence-level predictions are obtained by choosing the class with the highest average score of the three images. The confusion matrix of the proposed method on the MMI database is reported in Table 2. As shown in Table 3, the performance improvements in this small database without causing overfitting are impressive. The proposed method outperforms other works that also use static image-based features and can achieve comparable and even better results than those video-based approaches.

Table 3. Recognition accuracy comparison on the CK+ database [26] in terms of seven expressions, and MMI database [32] in terms of six expressions.

Methods	CK+	MMI
MSR [33]	91.4%	N/A
ITBN [44]	91.44%	59.7%
BNBN [25]	96.7%	N/A
IB-CNN [11]	95.1%	N/A
3DCNN-DAP [23]	92.4%	63.4%
STM-Explet [24]	94.19%	75.12%
DTAGN [16]	97.25%	70.2%
Inception [28]	93.2%	77.6%
1B(*N*+1)Softmax	93.21%	77.72%
2B(*N*+1)Softmax	94.3%	78.04%
1B(*N*+M)Softmax	96.55%	77.88%
2B(*N*+*M*)Softmax	**97.1%**	**78.53%**

The static facial expressions in the wild (SFEW) database [6] is created by extracting frames from the film clips in the AFEW data corpus. There are 1766 well-labeled images (i.e., 958 for training, 436 for validation and 372 for testing) being assigned to be one of the 7 expressions. Different from the previous two datasets, it targets for unconstrained facial expressions, which has large variations reflecting real-world conditions. The confusion matrix of our method on the SFEW validation set is reported in Table 4. The

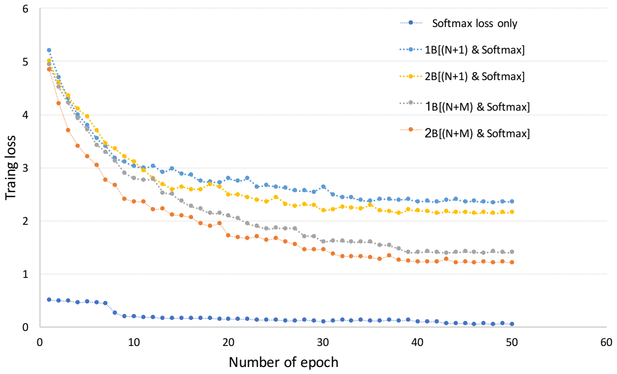

Figure 5. The training loss of different methods on SFEW validation set.

recognition accuracy of disgust and fear are much lower than the others, which is also observed in other works. As illustrated in Table 5, the CNN-based methods dominate the ranking list. With the augmentation of deep metric learning and two-branch FC layer network, the proposed method works well in the real world environment setting. Note that Kim et al. [18] employed 216 AlexNet-like CNNs with different architectures to boost the final performance. Our network performs about 25M operations, almost four times fewer than a single AlexNet. With the smaller size, the evaluation time in testing phase takes only 5ms using a Titan X GPU, which makes it applicable for real-time applications.

Overall, we can see that joint optimizing the metric learning loss and softmax loss can successfully capture more discriminative expression-related features and translate them into the significant improvement of FER accuracy. The (*N*+*M*)-tuplet clusters loss not only inherits merits of conventional deep metric learning methods, but also learns features in a more efficient and stable way. The two-branch FC layer can further give a boost in performance. Some nice properties of the proposed method are verified by Fig. 5, where the training loss of 2B(*N*+*M*)Softmax converges after about 40 epochs with a more steady decline and reaches a lower value than those baseline methods as we expect. As Fig. 6 illustrates, the proposed method and the baseline methods achieve better performance in terms of the validation accuracy on the training phase.

Table 4. Average confusion matrix obtained from proposed method on the SFEW validation set [6].

		Predict						
		AN	DI	FE	HA	NE	SA	SU
Actual	AN	**66.24%**	1.3%	0%	6.94%	9.09%	5.19%	10.69%
	DI	21.74%	**4.35%**	4.35%	30.34%	13.04%	4.35%	21.74%
	FE	27.66%	0%	**6.38%**	8.51%	10.64%	19.15%	27.66%
	HA	0%	0%	0%	**87.67%**	6.85%	1.37%	4.11%
	NE	5.48%	0%	2.74%	1.37%	**57.53%**	5.48%	27.4%
	SA	22.81%	0%	1.75%	7.02%	8.77%	**40.35%**	19.3%
	SU	1.16%	0%	2.33%	5.81%	17.44%	0%	**73.26%**

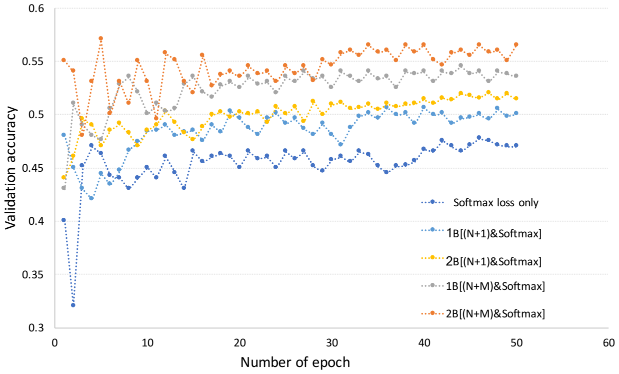

Figure 6. The validation accuracies of different methods on SFEW validation set.

Table 5. Recognition accuracy comparison on the SFEW database [6] in terms of seven expressions.

Methods	Validation
Kim et al. [18]	53.9%
Yu et al. [48]	55.96%
Ng et al. [31]	48.5%
Yao et al. [46]	43.58%
Sun et al. [38]	51.02%
Zong et al. [53]	N/A
Kaya et al.[17]	53.06%
Mao et al.[27]	44.7%
Mollahosseini [28]	47.7%
1B(*N*+1)Softmax	49.77%
2B(*N*+1)Softmax	50.75%
1B(*N*+*M*)Softmax	53.36%
2B(*N*+*M*)Softmax	**54.19%**

6. Conclusion

We derive the (*N*+*M*)-tuplet clusters loss and combine it with softmax loss in a unified two-branch FC layer joint metric learning CNN architecture to alleviate the attribute variations introduced by different identities on FER. The efficient identity-aware negative-mining and online positive-mining scheme are employed. After evaluating performance on the posed and spontaneous FER dataset, we show that the proposed method outperforms the previous softmax loss-based deep learning approaches in its ability to extract expression-related features. More appealing, the (*N*+*M*)-tuplet clusters loss function has clear intuition and geometric interpretation for generic applications. In future work, we intend to explore the use of it to the person or vehicle re-identifications.

Acknowledgements

The funding support from Youth Innovation Promotion Association, CAS (2017264), Innovative Foundation of CIOMP, CAS (Y586320150) and Hong Kong Government General Research Fund GRF (Ref. No.152202/14E) is greatly appreciated.

References

[1] S. Arora, A. Bhaskara, R. Ge and T. Ma. Provable bounds for learning some deep representations. *arXiv preprint arXiv:1310.6343*, 2013.

[2] E. Barsoum, C. Zhang, C. C. Ferrer and Z. Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ICMI*, pages 279-283, 2016.

[3] T. Baltrusaitis, M. Mahmoud and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Automatic Face and Gesture Recognition (FG)*, pages 6:1-6, 2015.

[4] S. Chopra, R. Hadsell and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

[5] J. F. Cohn and P. Ekman. Measuring facial action. *The new handbook of methods in nonverbal behavior research*, pages 9–64, 2005.

[6] A. Dhall, Abhinav, O. V. R Murthy, R. Goecke, J. Joshi and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *ICMI*, pages 423–426, 2015.

[7] S. Ding, L. lin, G Wang and H. Chao. Deep Feature Learning with Relative Distance Comparison for Person Re identification. *Pattern Recognition*, 48: 2993-3003, 2015.

[8] Y. Dong, B Du, L. Zhang, L. Zhang and D. Tao. LAM3L: Locally adaptive maximum margin metric learning for visual data classification. *Neurocomputing,* 235:1–9, 2017.

[9] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D Lee, Y Zhou, C Ramaiah, F Feng, R Li and X Wang. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, page 117–124, 2013.

[10] R. Gross, I. Matthews, J. Cohn, T. Kanade and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[11] S. Han, Z. Meng, A. S. Khan and Y. Tong, Incremental Boosting Convolutional Neural Network for Facial Action Unit Recognition, In *NIPS*, pages 109–117, 2016.

[12] S. Horiguchi, D. Ikami and K. Aizawa. Significance of softmax-based features over metric learning-based features. In *ICLR*, 2017.

[13] S. Jain, C. Hu and J.K. Aggarwal. Facial expression recognition with temporal modeling of shapes. In *CVPRW*, pages 1642–1649, 2011.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[15] B. Jiang, M. Valstar and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face and Gesture Recognition (FG)*, pages 314-321, 2011.

[16] H. Jung, S. Lee, J. Yim, S. Park and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *ICCV*, pages 2983–2991, 2015.

[17] H. Kaya and A. Salah. Combining modality-specific extreme learning machines for emotion recognition in the wild. *Journal on Multimodal User Interfaces* 10(2):139-149, 2016.

[18] B.K. Kim, J. Roh, S.Y. Lee and J. Roh. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, 10(2): 173-189, 2016.

[19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[20] I. Kotsia, S. Zafeiriou and S. Fotopoulos. Affective gaming: A comprehensive survey. In *CVPRW*, pages 663–670, 2013.

[21] Z. Li, S Chang, F Liang, T.S. Huang, L Cao and J R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, pages 3610-3617. 2013.

[22] H. Liu, Y. Tian, Y. Yang, L. Pang and T. Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *CVPR*, pages 2167-2175, 2016.

[23] M. Liu, S. Li, S. Shan, R. Wang and X. Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *ACCV*, pages 143–157, 2014.

[24] M. Liu, S. Shan, R. Wang and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *CVPR*, pages 1749–1756, 2014.

[25] P. Liu, S. Han, Z. Meng and Y. Tong. Facial expression recognition via a boosted deep belief network. In *CVPR*, pages 1805–1812, 2014.

[26] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews. The extended cohn-kanade dataset (ck+): A complete expression dataset for action unit and emotion-specified expression. In *CVPRW*, pages. 94-101, 2010.

[27] Q. Mao, Q. Rao, Y. Yu and M. Dong. Hierarchical Bayesian Theme Models for Multi-pose Facial Expression Recognition. *IEEE Transactions on Multimedia*, 19(4): 861 – 873, 2017.

[28] A. Mollahosseini, D. Chan and M.H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *WACV,* pages 1–10, 2016.

[29] A. Mollahosseini, B. Hassani, M.J. Salvador, H. Abdollahi, D. Chan and M.H. Mahoor. Facial expression recognition in the world wild web. In *CVPR,* pages 58-65, 2016.

[30] A. Mollahosseini and M. H. Mahoor. Bidirectional warping of active appearance model. In *CVPRW*, pages 875–880, 2013.

[31] H.W. Ng, V.D. Nguyen, V. Vonikakis and S Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *ICMI*, pages 443–449, 2015.

[32] M. Pantic, M Valstar, R Rademaker, and L Maat. Web-based database for facial expression analysis. In *ICME,* pages5, *2005.*

[33] S. Rifai, Y Bengio, A. Courville, P. Vincent and M Mirza. Disentangling factors of variation for facial expression recognition. In *ECCV*, pages 808–822, 2012.

[34] E. Sariyanidi, H. Gunes and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 37:1113–1133, 2015.

[35] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCVW*, pages 1003-1011, 2015.

[36] H. Shi, Y. Yang, X. Zhu, L. Zhen, W. Zheng and S.Z. Li. Embedding deep metric for person re-identification: a study against large variations. In *ECCV*, pages 732-748, 2016.

[37] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, pages 1849-1857, 2016.

[36] H. Shi, Y. Yang, X. Zhu, L. Zhen, W. Zheng and S.Z. Li. Embedding deep metric for person re-identification: a study against large variations. In *ECCV*, pages 732-748, 2016.

[37] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, pages 1849-1857, 2016.

[38] B. Sun, L. Li, G. Zhou, X. Wu, J. He, L. Yu, D. Li and Q. Wei. Combining multimodal features within a fusion network for emotion recognition in the wild. In *ICMI*, pages 497–502, 2015.

[39] Y. Sun, Y. Chen, X. Wang and X. Tang, Deep learning face representation by joint identification-verification. In *NIPS*, pages 1988–1996, 2014.

[40] Y. Tian, T. Kanade and J. F. Cohn. Facial expression analysis. In *Handbook of face recognition*, pages 247–275, 2005.

[41] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *ICCV*, pages 593-600, 2013.

[42] M.F. Valstar, M. Mehu, B. Jiang, M. Pantic and K. Scherer. Meta-analysis of the first facial expression recognition challenge. *IEEE Trans. Syst., Man Cybern. B, Cybern,* 42(4): 966–979, 2012.

[43] Q. Wang, W. Zuo, L. Zhang and P. Li. Shrinkage expansion adaptive metric learning. In *ECCV*, pages 456–471, 2014.

[44] Z. Wang, S Wang and Q Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *CVPR*, pages 3422–3429, 2013.

[45] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013.

[46] A. Yao, J. Shao, N. Ma and Y. Chen. Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *ICMI*, pages 451–458, 2015.

[47] D. Yi, Z. Lei, S. Liao and S.Z. Li, Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[48] Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning. In *ICMI*, pages 435–442, 2015.

[49] A. Yüce, H. Gao and J. P. Thiran. Discriminant multi-label manifold embedding for facial action unit detection. In *Automatic Face and Gesture Recognition (FG),* pages 9:1-6, 2015.

[50] S. Zafeiriou, A. Papaioannou, I. Kotsia, M. Nicolaou and G. Zhao. Facial Affect "In-the-Wild": A Survey and a New Database. In *CVPRW*, pages 1487-1498, 2016.

[51] L. Zhang, D. Tjondronegoro and V. Chandran. Random Gabor based templates for facial expression recognition in images with facial occlusion. *Neurocomputing,* 145:451-464, 2014.

[52] X. Zhang, F. Zhou, Y. Lin and S. Zhang, Embedding label structures for fine-grained feature representation. In *CVPR*, pages 1114–1123, 2016.

[53] Y. Zong, W. Zheng, X. Huang, J. Yan, and T. Zhang. Transductive transfer lda with riesz-based volume lbp for emotion recognition in the wild. In *ICMI*, pages 491–496, 2015.