JMLR: Workshop and Conference Proceedings 44 (2015) 60-71 NIPS 2015

The 1st International Workshop "Feature Extraction: Modern Questions and Challenges"

Learning Multi-channel Deep Feature Representations for Face Recognition

Xue-wen Chen XUEWEN.CHEN@WAYNE.EDU
Melih S. Aslan* MELIH.ASLAN@WAYNE.EDU
Kunlei Zhang* KUNLEI.ZHANG@WAYNE.EDU
Department of Computer Science, Wayne State University, Detroit, MI 48202, USA

Thomas S. Huang T-HUANG1@ILLINOIS.EDU
Department of Electrical and Computer Engineering, University of Illinois at Urbana Champaign, Urbana, IL 61801, USA

Editor: Afshin Rostamizadeh

Abstract

Deep learning provides a natural way to obtain feature representations from data without relying on hand-crafted descriptors. In this paper, we propose to learn deep feature representations using unsupervised and supervised learning in a cascaded fashion to produce generically descriptive yet class specific features. The proposed method can take full advantage of the availability of large-scale unlabeled data and learn discriminative features (supervised) from generic features (unsupervised). It is then applied to multiple essential facial regions to obtain multi-channel deep facial representations for face recognition. The efficacy of the proposed feature representations is validated on both controlled (i.e., extended Yale-B, Yale, and AR) and uncontrolled (PubFig) benchmark face databases. Experimental results show its effectiveness.

Keywords: multi-channel features, feature learning, representation learning, face recognition

1. Introduction

Feature extraction is to seek for appropriate transformations from raw data into features that can be used as the input for machine learning methods. The performance of a learning algorithm heavily depends on the features to which it is applied. Hence, extracting informative features plays a crucial role in various applications: not only does it reduce input dimensionality to alleviate the curse of dimensionality problems, but also a more meaningful representation can disentangle the underlying factors of variation (Bengio et al., 2013).

We study effective feature learning methods and explore an application to the task of automatic face recognition in this paper. Over the past two decades, there has been a substantial body of work on feature representation methods for face recognition to address challenging problems associated with intra-class or inter-class variations (e.g., arbitrary illumination, pose, expression, occlusion, etc.). Both global and local feature based approaches have been applied to face recognition problems. Eigenface (Turk and Pentland, 1991) and Fisherface (Belhumeur et al., 1997b) are two representative techniques among global feature methods. Global representations generally fail to capture higher-order statistics. Most methods for face recognition thus rely on local representations, often given by hand-crafted local descriptors such as Gabor feature (Liu and Wechsler, 2002), local binary pattern feature (Ahonen et al., 2006), SIFT (Lowe, 2004), and histograms of oriented gradients (Dalal and Triggs, 2005), or learning-based descriptor (Cao et al., 2010) such as binarized statistical image features (Kannala and Rahtu, 2012), and discriminant face descriptor (Lei et al., 2014). Recent

*. Both authors contribute equally

advances in deep learning open a new way to avoid labor-intensive feature engineering, and feature representations can be automatically learned from data through deep network architectures. They offer several advantages over those obtained through local descriptors. For example, they are capable of capturing higher-order statistics. Recently, deep learning based methods (e.g., Schroff et al., 2015; Taigman et al., 2014; Sun et al., 2014; Zhu et al., 2013; Huang et al., 2012) have shown great performance on face recognition problems.

In this paper, we propose to learn a new multi-channel deep feature representation (dubbed as 'McDFR') for face recognition. The main idea is based on unsupervised learning for generic yet descriptive features followed by supervised learning for class specific features. Specifically, face images are treated as a set of essential facial regions such as right eyes, left eyes, noses and mouths. For each facial region, we perform a deep unsupervised learning model on unlabeled data to extract generic features, which are capable of characterizing the essential representation for the region. Considering features learned in unsupervised fashion involve no supervision (i.e., label) information and thus they are not directly related to the recognition task, such generic features for each facial region are fed into a supervised learning via a deep neural network (DNN) to obtain more discriminative representations among classes. These multi-channel representations are then fused together to get the final features which are used as input to a supervised classifier for face recognition.

Different from a general deep learning method such as deep belief nets (Hinton et al., 2006b), where supervised fine-tuning is applied to the network with the original (raw) data as input, the proposed approach extracts more discriminative features through another supervised learning on the generic descriptive features. Furthermore, our approach allows for different datasets in unsupervised and supervised phases, which can fully utilize the availability of large-scale unlabeled data. In the proposed method, deep unsupervised learning is to remove redundant information from unlabeled data for distilling descriptive representations (e.g., the features to suitably represent *eyes* in general), and supervised learning is to identify descriptors for optimal discrimination among classes (e.g., the features to distinguish *John's eyes* from *David's* based on the descriptive representations for *eyes*). Figure 1 provides an overview of the proposed multi-channel deep feature representations for face recognition.

As illustrated in Figure 1, the proposed McDFR learning framework is also capable of integrating other available multimedia content, e.g., speech, by treating it as an additional channel. Compared to directly fusing heterogeneous data as a single channel, the proposed McDFR method gains more benefit from complementary information in multimedia data to achieve higher recognition accuracy (see Section 3.5).

Our main contributions in this paper are three-fold. 1) We propose to learn deep feature representations using the idea of unsupervised learning for generic features on which supervised learning is performed to yield class specific features. 2) We apply such deep feature learning approach to essential facial regions to obtain multi-channel deep facial representations for face recognition. 3) The proposed method produces the state-of-the-art recognition accuracy on several benchmark face databases.

2. Methodology

2.1 Learning Multimodal Deep Facial Representations

As shown in Figure 1, the proposed multi-channel deep facial representations consists of preprocessing, generic image feature learning using deep autoencoders, class specific feature learning using DNNs, and integration of multi-channel representations. The details of each part will be described in the following sections.

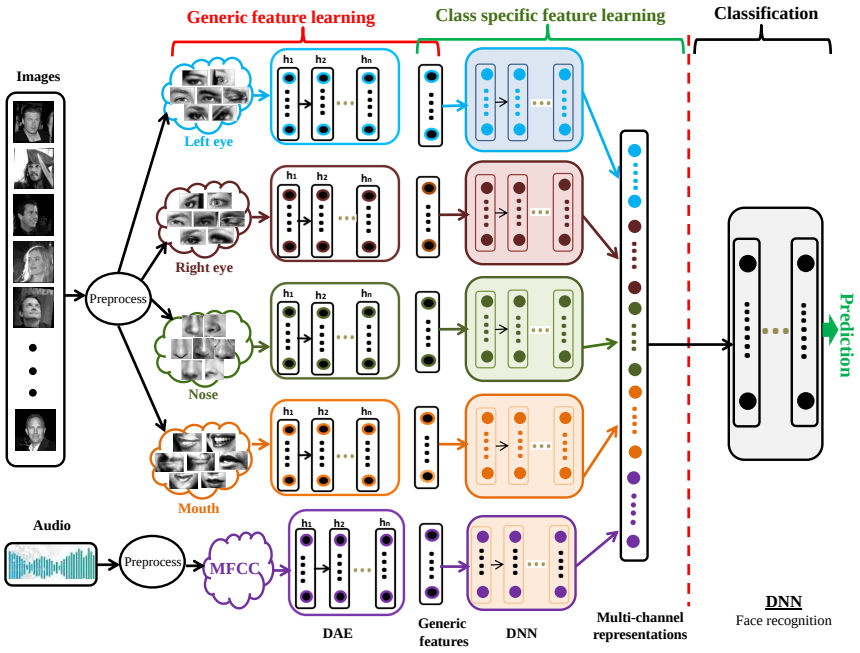

Figure 1: Outline of the proposed multi-channel deep feature representations for face recognition. The input data (images or other multimedia data) is first preprocessed. In each channel, generic features are learned on a (unlabeled) facial region or audio data through a deep autoencoder (DAE), and then class specific features are learned under supervision by feeding generic features into a DNN. The learned features from multiple channels are fused together as the final representation which is used as input to another DNN for classification.

2.1.1 Preprocessing

All the original face images are converted to grayscale images. Four essential facial regions, i.e., right eyes, left eyes, noses and mouths, are segmented from the images. The segmentation can be done automatically by using a facial region detection software[1]. Intensities of image pixels in the entire dataset are normalized to have a zero mean and unit variance. This is all the preprocessing we apply in this paper.

2.1.2 Generic Feature Learning

We carry out an unsupervised feature learning by training a deep autoencoder with stacked RBMs (Hinton et al., 2006a) to extract generic image features. The goal in this stage is to remove redundant information from unlabeled data but distill descriptive features. Other techniques such as stacked denoising autoencoder, stacked convolutional autoencoders, and their variants can also be used here.

As a building block of the deep autoencoder, an RBM is composed of a visible layer $\mathbf{v}$ and a hidden layer $\mathbf{h}$ with an array of connection weights $\mathbf{W}$ between the visible and hidden units but no connections between neurons of the same layer. The energy function of the pair of visible and hidden variables is bilinear (assume the vectors in this paper are column vectors):

$$\mathrm{E}(\mathbf{v}, \mathbf{h}) = -\mathbf{b}^T\mathbf{v} - \mathbf{c}^T\mathbf{h} - \mathbf{h}^T\mathbf{W}\mathbf{v}, \tag{1}$$

where vectors $\mathbf{b}$ and $\mathbf{c}$ are the biases of the input layer and the hidden layer, respectively. According to (Hinton et al., 2006b), the conditional distribution in RBM can be factorized due to the lack

1. OKAO Vision, http://www.omron.com/technology/index.html

of visible-visible and hidden-hidden connections. That is to say, calculation of the conditional distribution can be decomposed as $P(\mathbf{h}|\mathbf{v}) = \prod_i P(h_i|\mathbf{v})$.

In the binary case (Binary-Binary RBM) where both visible and hidden nodes take either zero or one, the probability of a hidden node taking value one happens to be a sigmoid function of the input:

$$P(h_i = 1|\mathbf{v}) = sigmoid(\mathbf{W}_i\mathbf{v} + c_i), \tag{2}$$

while the conditional distribution over the hidden nodes given the visible nodes is

$$P(v_i = 1|\mathbf{h}) = sigmoid\left(\mathbf{h}^T\mathbf{W}_i + b_i\right). \tag{3}$$

To handle continuous real-valued data, e.g., face images, we employ Gaussian-Binary Restricted Boltzmann Machine (GRBM) which is an extension of the Binary-Binary RBM. For a GRBM, the energy function is

$$\mathrm{E}(\mathbf{v}, \mathbf{h}) = \frac{1}{2}\mathbf{v}^T\Lambda\mathbf{v} - \mathbf{b}^T\Lambda\mathbf{v} - \mathbf{c}^T\mathbf{h} - \mathbf{h}^T\mathbf{W}\Lambda^{\frac{1}{2}}\mathbf{v}, \tag{4}$$

where Λ is the precision matrix of $\mathbf{v}$, which is diagonal. The conditional distributions become

$$P(\mathbf{h}|\mathbf{v}) = sigmoid(\mathbf{W}\Lambda^{\frac{1}{2}}\mathbf{v} + \mathbf{c}), \tag{5}$$

and

$$P(\mathbf{v}|\mathbf{h}) \propto \mathcal{N}(\mathbf{h}^T\mathbf{W}\Lambda^{\frac{1}{2}}, \Lambda^{-1}), \tag{6}$$

where $\mathcal{N}$ denotes a Gaussian distribution. We construct a multilayer encoder by stacking multiple Binary-Binary RBMs (except the top one which is a Binary-Gaussian RBM) on top of the GRBM.

RBMs can be learned using approximate maximum likelihood estimation with k-step contrastive divergence (Hinton, 2002). k is set to be 1 in this paper. The learning of a deep autoencoder is started by a layer-by-layer pretraining procedure which learns a stack of RBMs in the encoder network. After the pretraining phase, the RBMs are unrolled to form an autoencoder, and the decoder network is initialized with the transpose of the learned weights for the encoder network. The whole deep autoencoder network is then fine-tuned for optimal reconstruction by backpropagating the error derivatives. We use mean square difference as the error and the batch stochastic gradient descent (SGD) method with momentum is employed to update the weights.

For each essential facial region, we train one such deep autoencoder. The activation of the top layer in encoder network is considered as the learned representations which capture generic high-order nonlinear structure of the facial regions. The learned features are normalized to $[-1, 1]$: each element of a feature vector is divided by its largest absolute value, i.e.,

$$\overline{V(I_{region})} = g(I_{region}) = \frac{V(I_{region})_i}{\max(\mathrm{abs}(V(I_{region}), \varepsilon))} \tag{7}$$

where $V(I_{region})$ denotes the extracted feature vector for a facial image region I_{region} or other multimedia data, and ε is capped factor, e.g., 0.01 to avoid dividing by a too small value.

2.1.3 Class Specific Feature Learning

Since the optimization objective of unsupervised feature learning is not directly related to the recognition task, the learned features cannot preserve identity information (Zhu et al., 2013). To further improve the discriminative power among subjects, we conduct an class specific feature learning via training a supervised DNN on a multi-class face recognition task. Identity information (i.e., label) is explicitly incorporated into the objective function and directly used as the supervision of the feature learning. The stage is to identify good descriptors, based on the normalized features extracted in Section 2.1.2, for optimal discrimination among classes. The activation of the last hidden layer is treated as the learned class specific features.

For the design of the DNN model architecture, we employ several recent advanced techniques, such as the rectified linear units (ReLU) (Nair and Hinton, 2010) activation function and the dropout regularization method (Hinton et al., 2012). The ReLU, that we use for all hidden layers, produces sparse activity vectors and learns much faster than ordinary logistic units. On the top layer, we adopt a softmax function which produces a distribution over the subject labels. The probability assigned to the c-th subject is

$$p_c = \frac{\mathrm{e}^{(\mathbf{W}_c^L \mathbf{x}^{L-1} + b_c^L)}}{\sum_{j=1}^{C} \mathrm{e}^{(\mathbf{W}_j^L \mathbf{x}^{L-1} + b_j^L)}}, \tag{8}$$

where L denotes the number of layers, C denotes the number of subjects, $\mathbf{W}_c^L$ is the $c-$th row of the last layer weight matrix, and b_c^L is the c-th value in the last layer bias vector. During training, the goal is to maximize the probability of the correct subject, which can be accomplished by minimizing the cross-entropy error

$$\mathcal{L} = -\sum\nolimits_n^N \sum\nolimits_c^C y_c^n \log p_c^n, \tag{9}$$

where N denotes the number of training samples, and y denotes the true labels. Backpropagation algorithm is applied to compute the gradient of $\mathcal{L}$ with respect to the weight and bias, and batch SGD is used to update them.

To avoid the overfitting problem in DNNs, we use the dropout method as a regularization which randomly sets each unit's activation in all hidden layers to be 0 with a probability p. The value of p can be chosen based on the specific problem at hand (a typical value of p is 0.5).

2.1.4 Integration of Multi-channel Representations

Similarly, the features learned in Section 2.1.3 to represent each facial region are first normalized to $[-1, 1]$, which brings them within the same dynamic range and thus improves the stability of the classification. Then, these normalized features are fused together through concatenating one by one to get the final facial representation $V(I)$ for a given face image I:

$$V(I) = [V(I_{Leye}); V(I_{Reye}); V(I_{mouth}); V(I_{nose})] \tag{10}$$

where the feature vector for a facial region is $V(I_r) = g\left(f_\phi^{DNN}\left(g\left(f_\theta^{AE}(I_r)\right)\right)\right)$, '$r$' represents an essential facial region: Leye (left eye), Reye (right eye), mouth or nose, $f_\theta^{AE}()$ and $f_\phi^{DNN}()$ denote the deep autoencoder with parameters θ and the DNN model with parameters ϕ, respectively, and $g()$ denotes the normalized function defined in Eq. (7). When other multimedia data is available, Eq. (10) becomes

$$V(I) = [V(I_{Leye}); V(I_{Reye}); V(I_{mouth}); V(I_{nose}); V(M)], \tag{11}$$

where $V(I_M) = g\left(f_\phi^{DNN}\left(g\left(f_\theta^{AE}(I_M)\right)\right)\right)$.

The integration of multi-channel facial representations is a crucial step in the proposed method, since different essential facial regions provide complementary discriminability among subjects to a significant extent. It is found that, comparing with the feature extracted from a single region, the fused feature highly improves the recognition accuracy in the experiments (see Section 3).

2.2 Recognition Algorithm

We apply the proposed McDFR to the face identification task. This can be accomplished by training another DNN over the obtained feature vectors in Eq. (10) or Eq. (11) for a multi-class classification. The design of this DNN architecture is similar to that described in Section 2.1.3. Note that, in addition to a DNN model, other supervised classifiers, such as support vector machines, can be an alternative choice to perform face recognition here.

Once the proposed framework is trained, it can be used to identify a test face image. All the test images should be preprocessed in the same way as the training phase, then undergo the deep encoder (Section 2.1.2), the trained deep feedforward neural networks (Section 2.1.3), the same manner to integrate the multi-channel feature vectors (Section 2.1.4), and finally are classified by the trained DNN in this section to get the subject index.

3. Experiments and Results

The efficacy of the proposed framework is demonstrated on both controlled and uncontrolled benchmark face recognition databases including the extended Yale-B (Belhumeur et al., 2001), Yale (Belhumeur et al., 1997a), AR (Martınez and Benavente, 1998), PubFig83 (Pinto et al., 2011), and a multimedia data. In the experiments, we define three data subsets, i.e. DATA-A, DATA-B, and DATA-C. DATA-A does not contain label information whereas DATA-B and DATA-C contain the labels and same subjects but completely distinct images. Moreover, images and subjects in DATA-A can be joint or disjoint with those included in DATA-B. While DATA-A is employed to train the deep autoencoders for generic image feature learning, Data-B is adopted to train the DNNs for class specific feature learning, as well as train the DNN that is used for face identification. Data-C is only used for evaluation in the testing phase.

Table 1: Average recognition rates on *extended Yale-B* database with the use of features from each single facial region and the fused representation, respectively.

Region	Recognition rate (%)
Left eye	94.71 ± 0.43
Right eye	96.39 ± 0.52
Nose	94.98 ± 0.53
Mouth	95.78 ± 0.38
McDFR	**99.66 ± 0.13**

3.1 Face Recognition on extended Yale-B database

The extended Yale-B database, which contains 38 subjects, around 64 frontal images per subject and thus about 2,414 images in total, is used to asses the proposed method on severe illumination changes. Each subject was imaged under 64 different illumination conditions. We segment four essential facial regions with sizes of 26 x 34 (left eye and right eye), 25 x 47 (mouth), and 36 x 33 (nose).

In this experiment, we conduct 10 runs for training-test procedures to get the average recognition rates. For each run, a random subset with 32 images per subject is selected to get DATA-B in the training phase whereas the rest of the database is considered to be the testing set, i.e., DATA-C. We choose DATA-A to be the same as DATA-B.

Table 1 gives the average recognition rates and standard deviations across 10 runs with the use of features from each single facial region and the multi-channel representation, respectively. It is observed that the fused features provide a clear accuracy advantage over single ones. Table 2 presents the face recognition rates and the comparison with current state-of-the-art results. It is seen that the proposed method provides the highest accuracy on this database. Specifically, our method obtains an average recognition rate of 99.66% with standard deviation of 0.13%. Out of 10 randomly runs, the maximum and minimum accuracies are 99.92% and 99.42%, respectively. Even with the minimum one, to our best knowledge in the literature, we achieve the state-of-the-art

performance. This proves that the proposed multi-channel deep facial representations are robust to various illuminations on the face images.

Table 2: Comparison of face recognition rates on *extended Yale-B* database

Methods	**Recognition rate (%)**	**#Train**
Aw-SpLPP (Wang et al., 2010)	98.25	30
CRC-RLS (Zhang et al., 2011)	97.90	32
SRC (Yang et al., 2011)	99.40	32
McDFR	**99.66**	32

3.2 Face Recognition on Yale Database

The Yale database contains 15 subjects and around 11 frontal images per subject, where each image has one type of facial expressions and configurations, e.g., normal, happy, sad, sleepy, surprised, wink, and with or without glasses. Four essential facial regions are segmented as 20 x 30 (left eye and right eye), 16 x 23 (mouth), and 30 x 24 (nose).

The proposed method is applied on this dataset to assess that the extracted facial representations are tolerate to expression, and some of the other challenges such as occlusion and illumination. In this experiment, a random selection of 2, 3, 4, 5, 6, 7 or 8 images per subject is used as the training set (DATA-B), and the rest of images is used as the testing set (DATA-C). For each case, we run 10 times for the training-test procedure.

Table 3: Average recognition rates on *Yale* database with the use of features from each single facial region and the fused representation, respectively.

	Recognition rate(%)			
Region	*3 Train*	*5 Train*	*6 Train*	*8 Train*
Left eye	71.8 ± 4.1	81.9 ± 3.3	82.2 ± 4.9	86.7 ± 3.9
Right eye	73.9 ± 3.5	80.4 ± 4.1	83.2 ± 2.4	88.2 ± 3.8
Nose	62.2 ± 2.2	70.7 ± 4.6	73.2 ± 4.3	77.8 ± 6.6
Mouth	55.6 ± 4.7	65.2 ± 4.4	66.1 ± 4.1	71.9 ± 6.1
McDFR	**84.8 ± 3.1**	**94.7 ± 3.9**	**96.0 ± 3.3**	**97.8 ± 1.9**

The average recognition rates and standard deviations with the use of features from each single facial region and the fused representation, respectively, and with various numbers of images per subject for the training are presented in Table 3. Similarly, the multi-channel representation improves the recognition accuracy significantly. Table 4 compares the recognition rates obtained by the proposed method with current state-of-the-art results reported in (Gui et al., 2012). It can be seen that, for each case of training, our method outperforms all the existing methods.

3.3 Face Recognition on AR Database

The AR database (Martınez and Benavente, 1998) contains 100 subjects (50 men and 50 women), with 26 different images per subject which totals to 2,600 images taken in two sessions. In this database, there are facial expression (neural, smile, anger, scream), illumination, and occlusion (sunglass, scarf) challenges. We segment four essential facial regions with sizes of 26 x 34 (left eye and right eye), 20 x 40 (mouth), and 30 x 28 (nose).

We conduct two experiments on AR database: (1) training and testing on unoccluded images, (2) training on unoccluded images and test on occluded images. In the first experiment, we follow a

Table 4: Comparison of face recognition rates on *Yale* database, where T represents 'Train'

Methods	Recognition rate(%)						
	2 T	*3 T*	*4 T*	*5 T*	*6 T*	*7 T*	*8 T*
PCA	42.63	48.08	52.86	55.44	59.13	59.83	64.33
LPP	57.19	67.92	75.14	77.22	81.6	82.25	84.11
NDLPP	56.11	69.70	77.47	81.77	84.60	87.41	89.88
LPDP	56.74	71.75	78.90	81.78	86.73	88.17	90.67
DLPP/MMC	58.19	70.08	78.14	83.56	85.53	88.33	89.56
LDA	45.19	59.42	68.95	74.89	79.27	79.83	83.22
SNPE1	66.77	69.95	73.61	74.27	77.86	76.91	79.33
SNPE2	66.14	70.29	73.57	73.77	78.00	77.41	81.44
DSNPE1	72.33	82.33	86.85	90.61	93.60	93.41	96.00
DSNPE2	72.40	80.70	86.66	89.88	92.00	92.58	95.00
McDFR	**76.58**	**84.83**	**89.90**	**94.67**	**96.00**	**96.11**	**97.78**

scenario described in (Zang et al., 2012) which reported one of the state-of-the-art recognition rates. We select a subset of 1400 images which are composed of 14 images per subject with the facial expression and illumination changes. Various train-test image partitions are used: the training images (DATA-B) are selected randomly for each subject using the numbers $\{2, 3, 4, 5\}$, and the rest of images ($\{12, 11, 10, 9\}$) is considered as the test set (DATA-C). We conduct 10 runs for train-test procedure to get the average recognition rate for each partition. Table 5 provides the face recognition rates obtained by the proposed method and the comparison with those reported in (Zang et al., 2012). Except the case of 2 images per subject as training, our method always achieves the best performance with around 5% improvement on accuracy over the second highest values. This indicates that our method is capable of handling various facial expression and illumination challenges better than other methods.

Table 5: Comparison of face recognition rates on *AR* database when testing on unoccluded images.

Methods	Recognition rate(%)			
	2 Train	*3 Train*	*4 Train*	*5 Train*
PCA	34.94	43.44	50.71	56.13
LPP	55.07	62.60	68.12	71.58
NPE	40.45	46.02	52.95	61.12
ONPP	62.20	71.54	77.25	81.76
EPP	**72.45**	79.53	83.86	86.23
McDFR	70.92	**83.61**	**88.66**	**91.54**

In the second experiment, we follow another scenario described in (Naseem et al., 2010). For each subject, 8 images with only facial expression changes are used for training, and 4 images with occlusion (i.e., sunglass, scarf) are considered for test. The results obtained by the proposed method and the comparison with those reported in (Naseem et al., 2010) are given in Table 6. Our method obtains significantly improved accuracy over other methods in the case of testing only on images occluded by scarf. However, it performs worse than other methods when testing only on images occluded by sunglass, because in the former case three regions (left eye, right eye and nose) contribute to the recognition whereas in the latter case only two regions (nose and mouth) contribute. It is found that during the evaluation on most databases, eyes usually dominate the contribution. Overall, the proposed method yields the best results for occlusion.

3.4 Unconstrained Face Recognition on PubFig Database

In recent years, several uncontrolled databases have emerged in the literature for face recognition. Unlike the traditional face databases which are composed of images taken in controlled environments,

Table 6: Comparison of face recognition rates on *AR* database when testing on occluded images

Methods	Recognition rate(%)		
	Sunglass	*Scarf*	*Sunglass+Scarf*
PCA+NN	70.00	12.00	41.00
ICA+NN	53.50	15.00	34.25
LNMF+NN	33.50	24.00	28.75
l^2+NS	64.50	12.50	38.50
SRC	87.00	59.50	73.75
LRC	**96.00**	26.00	61.00
McDFR	63.00	**94.50**	**78.75**

face images in uncontrolled databases are generally collected from Internet sources. In particular, these images contain unrestricted varieties of expression, pose, lighting, occlusion, resolution, etc. Thus, unconstrained face recognition is a very challenging task.

Kumar et al. (2009) introduced the PubFig database, which contains 200 subjects and various numbers of images for different subjects. A modified subset of PubFig dataset, called PubFig83 (containing 83 subjects and 13,002 images), was introduced by Pinto et al. (2011) through removing duplicated photos and subjects with few photos. In this experiment, we use an aligned version of PubFig83 which is provided by Chiachia et al. (2014).

We segment four essential facial regions with sizes of 20 x 26 (left eye and right eye), 24 x 38 (mouth), and 30 x 24 (nose). Following the original protocol of (Pinto et al., 2011), we run 10 times and, for each run, take a random selection of 90 images per subject as the training set (DATA-B), the rest of images as the testing set (DATA-C). Table 7 gives the recognition accuracy and a comparison with other methods. Benefiting from the learned multi-channel deep features, our method achieves comparable accuracy, even on this challenging database, with current state-of-the-art methods.

Table 7: Comparison of face recognition rates on the *PubFig83* database.

Method	Accuracy(%)
Pinto et al., 2011	87.11±0.56
Chiachia et al., 2012	88.75±0.26
Chiachia et al., 2014	**92.28±0.28**
McDFR	**90.14±0.49**

3.5 Face Recognition on a Multimedia Database

To assess the feasibility of the proposed framework on the multimedia data, we prepare a data set containing face images and speech. We select 10 subjects along their face images from aligned version of PubFig83 database (Chiachia et al., 2014), and then download a video for each subject from YouTube to extract around 5 minute speech information. Note that the content and quality of the speech data are heterogenous. Thus, we have a multimedia data set containing 10 subjects, 1000 images, and 10 audio files.

For the images, we randomly select 50 images per subject as the training set (DATA-B), and the rest of images is used as the testing set (DATA-C). Same as in Section 3.4, four essential facial regions are used for facial representation learning. We employ the Mel frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980), and their first and second derivatives to represent the acoustic features. These features are calculated every 10 milliseconds using 25 milliseconds Hamming-window[2], and their first 12 elements are selected to form a 36-dimensional feature vector

2. We used Dan Ellis' implementation for MFCC which is available at http://www.ee.columbia.edu/ dpwe/resources/matlab/rastamat/

for each frame. In the experiment, features extracted from every 40 consecutive frames are concatenated to be a 1440-dimensional feature vector which is considered as one training/test example.

Table 8 gives the average recognition accuracy across 10 runs for direct fusion of image intensity and features of speech as a single channel, multi-channel representation with images and/or speech, respectively. Compared to the case of single channel, multi-channel representation with both images and speech provides an around 10% improvement on the recognition accuracy. This shows that the proposed framework is able to leverage multimedia data to benefit the face recognition.

Table 8: Recognition rates on a multimedia data set.

Features	**Recognition rate(%)**
McDFR (Images)	83.92 ± 0.86
Single Channel (Images + Speech)	82.24 ± 0.83
McDFR (Images + Speech)	**92.08 ± 0.41**

3.6 Compare to DBN for Face Recognition

In this experiment, we compare the proposed McDFR method to a supervised deep belief net (DBN) in the task of face recognition. We train a DBN to learn features for each essential facial region, and then perform an exactly same DNN architecture as the proposed method for recognition. Table 9 provides recognition rates on all the database for comparison. We can see that the proposed method always outperforms DBN. As stated in Section 1, we believe the reason is that, although the learning of a DBN consists of an unsupervised training over unlabeled data, the supervised fine-tuning is conducted over the original (raw) input data and network, while the proposed method extracts more discriminative features based on the learned generic features.

Table 9: Comparison between the proposed method and DBN

	Recognition rate(%)			
Methods	Yale (*8 Train*)	Yale-B (*32 Train*)	AR (*5 Train*)	PubFig83 (90 Train)
DBN	94.5 ± 1.1	99.34 ± 0.10	87.42 ± 0.28	87.53 ± 0.44
McDFR	**97.78 ± 1.9**	**99.66 ± 0.13**	**91.54 ± 0.33**	**90.14 ± 0.49**

4. Conclusions

We have presented a new approach to learning deep feature representations. We conduct unsupervised learning through deep autoencoders to extract generic features on which supervised learning through deep networks are performed to produce class specific features. Such deep feature learning is applied to each essential facial region (or other multimedia data) to learn its discriminative representation. A type of multi-channel deep facial representations is obtained by integrating the features of multiple essential facial regions (and other multimedia data). The multi-channel representations are then fed into a supervised classifier for face recognition. The proposed method achieves the state-of-the-art performance on several benchmark face databases. We believe that the proposed method can be extended to other pattern recognition problems.

Acknowledgments

The authors would like to acknowledge the support from National Science Foundation awards OIA-1028098.

References

T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28:2037–2041, 2006.

P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. isherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):711–720, 1997a.

P. Belhumeur, A. Georghiades, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23 (6):643–660, 2001.

Peter N. Belhumeur, João P Hespanha, and David Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):711–720, 1997b.

Yoshua Bengio, Aaron Courville, and Pierre Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.

Qi Cao, Zhimand Yin, Xiaoou Tang, and Jian Sun. Face recognition with learning-based descriptor. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2707–2714, 2010.

Giovani Chiachia, Nicolas Pinto, William R. Schwartz, Anderson Rocha, Alexandre X. Falcão, and David Cox. Person-Specific Subspace Analysis for Unconstrained Familiar Face Identification. In *Proc. British Machine Vision Conference*, 2012.

Giovani Chiachia, Alexandre X Falcao, Nicolas Pinto, Anderson Rocha, and David Cox. Learning person-specific representations from faces in the wild. *IEEE Trans. Information Forensics and Security*, 9(12):2089–2099, 2014.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 886–893, 2005.

Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, and S. Ji. Discriminant sparse neighborhood preserving embedding for face recognition. *Pattern Recognition*, 45(8):2884–2893, 2012.

G. E. Hinton, R. Grosse, and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006a.

G. E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation 18*, 7:1527–1554, 2006b.

G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2518–2525, 2012.

Juho Kannala and Esa Rahtu. Bsif: Binarized statistical image features. In *Proc. Int. Conf. Pattern Recognition*, pages 1363–1366. IEEE, 2012.

N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proc. IEEE Int'l Conf. Computer Vision*, 2009.

Zhen Lei, Matti Pietikainen, and Stan Z Li. Learning discriminant face descriptor. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 36(2):289–302, 2014.

C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Tran. on Image Process.*, pages 467–476, 2002.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l Jour. Comp. Vis.*, 60: 91–110, 2004.

AM Martınez and Robert Benavente. The AR face database. *Computer Vision Center, Technical Report*, 24, 1998.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proc. Int'l Conf. Machine Learning*, pages 807–814, 2010.

Imran Naseem, Roberto Togneri, and Mohammed Bennamoun. Linear regression for face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(11):2106–2112, 2010.

Nicolas Pinto, Zak Stone, Todd Zickler, and David Cox. Scaling up biologically-inspired computer vision: A case study unconstrained face recognition on facebook. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, pages 35–42, 2011.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015.

Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, 2014.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance face verification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014.

Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3 (1):71–86, 1991.

J. Wang, B. Zhang, S. Wang, M. Qi, and J. Kong. An adaptively weighted sub-pattern locality preserving projection for face recognition. *Journal of Network and Computer Applications*, 33: 323–332, 2010.

M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust sparse coding for face recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 625–632, 2011.

F. Zang, J. Zhang, and J. Pan. Face recognition using elasticfaces. *Pattern Recognition*, 45(11): 3866–3876, 2012.

L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? In *Proc. Int'l Conf. Computer Vision*, pages 471–478, 2011.

Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning identity preserving face space. In *Proc. IEEE Int'l Conf. Computer Vision*, volume 1, page 2, 2013.