AU-aware Deep Networks for Facial Expression Recognition

Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen,

Abstract— **In this paper, we propose to construct a deep architecture, AU-aware Deep Networks (AUDN), for facial expression recognition by elaborately utilizing the prior knowledge that the appearance variations caused by expression can be decomposed into a batch of local facial Action Units (AUs). The proposed AUDN is composed of three sequential modules: the first module consists of two layers, i.e., a convolution layer and a max-pooling layer, which aim to generate an over-complete representation encoding all expression-specific appearance variations over all possible locations; In the second module, an AU-aware receptive field layer is designed to search subsets of the over-complete representation, each of which aims at best simulating the combination of AUs; In the last module, multi-layer Restricted Boltzmann Machines (RBM) are exploited to learn hierarchical features, which are then concatenated for final expression recognition. Experiments on three expression databases CK+, MMI and SFEW demonstrate the effectiveness of AUDN in both lab-controlled and wild environments. All our results are better than or at least competitive to the best known results.**

I. INTRODUCTION

In the recent years, Facial expression recognition has attracted much attention due to its potential applications, such as human-computer interaction, multimedia, surveillance, and so on. However, the exploration of the specific facial expression features is still an open problem. Studies in physiology and psychology indicate that the most descriptive regions for expression are located around certain facial parts (e.g. mouth, nose, eyes, and brows) . Among them, Facial Action Coding System (FACS) [1] is a typical work, which was designed to decompose each expression into several facial Action Units (AUs) which correspond to above-mentioned facial parts. Based on the definition of FACS, facial expressions can be more precisely described compared to rigid categorization [2].

Generally, the previous works on expression recognition can be categorized into two classes: AU-based methods [3], [4] and appearance-based methods [5], [6]. In [3], they first detected a number of pre-defined AUs, and then encoded their combinations as specific expression according to FACS. Since the definitions of AUs are ambiguous in semantics, it is difficult to achieve accurate AU detection in practice. On the other hand, many appearance-based methods identify an image or a sequence as one of the basic expression categories according to appearance features (e.g. local binary patterns (LBP) [6], Gabor [7], SIFT [8], HOG [9]). Although these methods have obtained satisfactory performance in some cases, the hand-crafted descriptors applying on a whole face in their schemes are lack of semantic interpretation in expression. As different facial expressions have explicit local variations, which are corresponded to AUs, it's intuitive to make use of these distinctive spatial regions. To this end as well as tackling the difficult AUs detection, [10] and [11] proposed feature selection schemes to automatically search for descriptive feature dimensions of local patches where certain AUs locates. This strategy is shown to be effective. However, there exist two problems: Firstly, hand-crafted features are still used in their methods, i.e. LBP, and Haar, which lack in explicit semantic meaning; Secondly, selected features are directly concatenated for final expression recognition, with no further mechanism to learn more higher-level representation, which may be essential for bridging the semantic gap between AUs and certain expressions. Based on above analysis, in this paper we utilize the prior knowledge about facial AUs but propose to solve the above problems as follows: First, producing more efficient representation by convolving input image with a batch of learned descriptors instead of hand-crafted descriptor, which can explicitly encode expression-specific appearance, such as frown, grin and glare; Second, adding a multi-layer learning process after AU-aware feature selection to extract incrementally higher-level features layer-by-layer. As both solutions can be characterized as several network layers, we formulate our framework as a multi-layer Deep Networks [12] due to its hierarchical representation ability on feature learning tasks.

In the deep networks framework, for the purpose of AU-aware feature selection, we first introduce the concept of "receptive field". As many deep learning methods [14], [15], [16] have achieved high performance of object recognition using large architectures with numerous features (hidden nodes), one critical concern in such large architectures is to specify the connections of nodes (features) between adjacent layers. Traditionally, the connections are limited by restricting higher level units to receive only lower-level inputs from certain subsets (e.g. based on spatial locality [17], [18]), which is called "receptive fields". However, for various recognition scenarios, the size and shape of receptive fields may be data-dependent which make it difficult to pre-define the fields without prior knowledge. To handle this, [19] proposes to select receptive fields automatically by grouping sets of most similar features during pre-training of deep networks. Such unsupervised scheme focus most on the relationship among the features, rather than exploring the relevance between features and categories, which makes the single receptive field inefficient on description and discrim-

Mengyi Liu, Shaoxin Li, Shiguang Shan and Xilin Chen are with Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China, {mengyi.liu, shaoxin.li, shiguang.shan, xilin.chen}@vipl.ict.ac.cn.

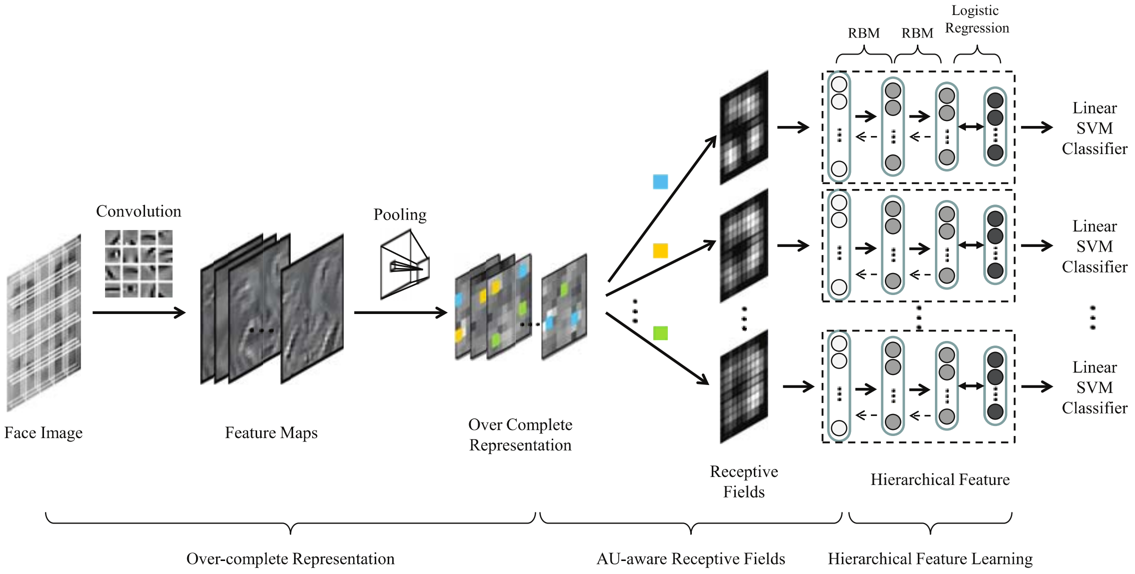

Fig. 1. The pipeline of the proposed method.

ination.

In order to construct the receptive fields automatically as well as utilize the co-occurrence information contained in disconnected facial regions, we introduce a supervised feature selection scheme to group subsets of low-level representations, which could simulate the different combinations of AUs. In this paper, we call such subsets “AU-aware Receptive fields (AURFs)” and this selection procedure is added into our whole framework for effective mid-level feature learning. An overview of the proposed method is presented in Fig.1. There are three sequential modules each of which consists of one or more network layers. As the proposed deep architecture can learn features according to the interpretation of facial AUs, we call it AU-aware Deep Networks (AUDN). Our AUDN is tested on two well known lab controlled expression database CK+ [21], MMI [22], and a wild expression database SFEW [23]. Compared to several hand-crafted appearance features, with the same linear SVM classifier [24], the learned features not only achieve state-of-the-art performance but also bears intuitive physiology and psychology appeal.

II. THE PROPOSED AUDN

This section details the three modules in the proposed AU-aware Deep Networks (AUDN). As is demonstrated in Fig.1, firstly, convolution layer and max-pooling layer are used to generate an over-complete representation. This representation can explicitly depict specific appearance presented in specific region. Then a feature selection scheme is used to find AURFs, which describe the combinations of local appearance variations. Finally, multi-layer RBM [20] is applied to each AURF respectively to learn hierarchical features for facial expression recognition.

A. *Over-complete Representation*

As accurate AUs detection is hard to achieve in static facial images, we try to design an over-complete feature representation over all possible spatial regions by convolving the dense-sampling facial patches with special filters. Previously, many patch-based learning methods tend to generate localized filters that simulate the function of simple cells in the primary visual cortex [25], [26]. These works justify the use of such filters in the standard model for object recognition. In this paper, we also attempt to learn a bank of specific filters, from all possible local patches from large number of expression images.

Among the plenty of algorithms for unsupervised learning, K-means has enjoyed wide adoption for generating codebooks of “visual words” [27], [28], which can be used to define higher-level features. This method is also applied to build layers in our framework. Suppose the patch size is u-by-u pixels, to obtain an over-complete representation, we set $K > u^2$ in K-means clustering and learn K centroids $c^{(k)}(k = 1, 2, ..., K)$ from all patches after normalizing and whitening. Then each centroid is considered as a filter to convolve with the patches in the whole facial images. For an input image with t-by-t patches, each 2D grid of t-by-t responses for a single filter is generally called a “map”. In the end we will get an t-by-t-by-K dimension representation after the convolutional layer (see Fig.2).

Depending on the first layer representation, it is hard to learn features invariant to image transformations (e.g. translation). This problem can be generally handled by incorporating “pooling” layers. In [16], it has been found that max-pooling can lead to faster convergence, superior invariant features selection, and improve generalization. Here the features go through our max-pooling layers are given by the maximum activation over adjacent, disjoint spatial blocks on each filter map.

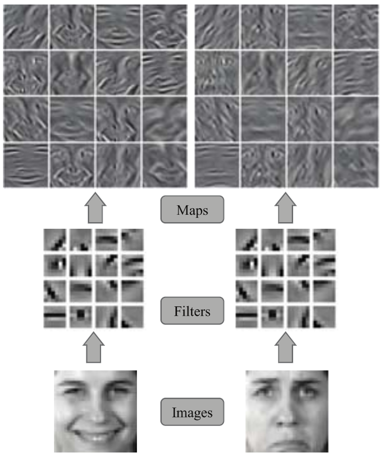

Fig. 2. Examples of filters and corresponding maps in convolutional layer.

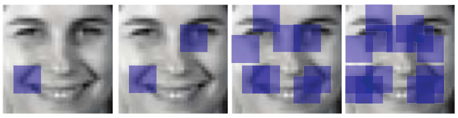

Fig. 3. Selected patches in AURF when $m = 1, 2, 8, 16$.

B. AU-aware Receptive Fields (AURFs)

In this module, we focus on selecting groups of AU-aware receptive fields from the outputs of max-pooling layer. As each feature in the outputs represents the presence of a single pattern which simulates specific AU, the selected receptive fields can depict complex combinations of appearance variations, which is expected to be consistent with the interpretation of FACS. To explore the expression-driven AU-aware features, we attempt to apply a greedy maximal relevance feature selection: iteratively selecting features with the highest relevance to the target expression category. However, in such greedy searching scheme, the combinations of individually best features do not necessarily lead to best classification performance [29]. For example, the features of adjacent regions that contained similar appearance is likely to have similar label-relevance score, but the combination of them can offer no more descriptive information to classification. On this consideration, some researchers proposed to select features with the minimal redundancy as an auxiliary condition [30], [31], [32], and one of these, based on criteria of minimal-redundancy-maximal-relevance (mRMR) [33], is a made-to-order scheme and proved to be effective in our experiments.

In our approach, relevance is characterized in terms of mutual information, which is widely used to measure dependency of variables. Given two random variables x and y, their mutual information is defined in accordance with their probabilistic density functions $p(x)$, $p(y)$, and $p(x, y)$:

$$I(x; y) = \int\int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy. \tag{1}$$

Max-relevance is to find a feature set S with m features $\{x_i\}$, which has the largest mean relevance to expression labels c. The scheme can be formulized as:

$$max\ D(S, c),\ D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c). \tag{2}$$

As the features only satisfying (2) may have unexpected redundancy, the following min-redundancy condition can be added to select relatively more diverse features:

$$min\ R(S),\ R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j). \tag{3}$$

Combining criterion presented in expression (2) and (3), an greedy search methods is used to find the local-optimal features defined by $max(D - R)$. Suppose we already have S^{k-1}, the set contained $k-1$ features. Our goal is to find the kth feature from the rest $P = X - S^{k-1}$. The incremental algorithm optimizes the following condition:

$$\underset{x_j \in P}{max}[I(x_j; c) - \frac{1}{k-1} \sum_{x_i \in S^{k-1}} I(x_j; x_i)]. \tag{4}$$

Our overall AU-aware receptive fields selection algorithm is shown in Algorithm. 1. And examples of local patches corresponding to selected features are shown in Fig.3.

Algorithm 1 : AU-aware Receptive Fields Selection

```
Input
   Over-complete representation $X = \{x_i | i = 1, 2, ..., M\}$;
   Expression labels of over-complete representation $c$;
   Selected feature dimension in each receptive field $m$;
   Number of AU-aware receptive fields $N$;
Output
   Receptive fields $RF^{(n)} = \{x_{n_i} | i = 1, 2, ..., m\}$;
Algorithm
   Initialize selected feature set $S = \phi$, the set of features
     to be selected $P = X - S$;
   for $n = 1, 2, ..., N$ do
      Set $RF^{(n)} = \phi$;
      Select feature $x_s$ from $P$ which has maximal
        mutual information with $c$:
              $\underset{x_s \in P}{argmax}\, I(x_s; c)$
      Update $RF^{(n)} = RF^{(n)} \bigcup \{x_s\}$, $S = S \bigcup \{x_s\}$,
        $P = X - S$;
      for $k = 2, ..., m$ do
         Select feature $x_s$ from $P$ based on equation (4):
              $\underset{x_s \in P}{argmax}[I(x_s; c) - \frac{1}{k-1} \sum_{x_i \in S} I(x_s; x_i)]$
         Update $RF^{(n)} = RF^{(n)} \bigcup \{x_s\}$, $S = S \bigcup \{x_s\}$,
           $P = X - S$;
      end for
   end for
```

C. Hierarchical Feature Learning

In this module, we attempt to learn even higher-level expression features from each AU-aware receptive field.

Recent neuroscience findings have provided insight into the principles governing information representation in the mammalian brain, which motivated the emergence of varies deep machine learning methods [12], [15], [18] focusing on computational models for information representation. One of the approaches in [20] applies a restricted Boltzmann machine (RBM) to model each new layer of higher-level features, which guarantees an increase on the lower-bound of the log likelihood of the data. The Single RBM is a two-layer (i.e. visible layer and hidden layer), undirected graphic model without lateral connections. The nodes in visible layer and hidden layer are represented as v_i, h_i respectively. If the visible units are real values, the configuration of v_i, h_i is characterized by an energy function as follows:

$$E(v,h) = \frac{1}{2}\sum_i v_i^2 - \sum_{i,j} v_i W_{ij} h_j - \sum_j b_j h_j - \sum_i c_i v_i, \quad (5)$$

where W_{ij} characterizes the association between visible and hidden nodes and c_i, b_j are the biases of visible layer and hidden layer respectively. Probabilistically, this is interpreted as

$$P(v,h) = \frac{exp(-E(v,h))}{Z}, \ Z = \sum_{v,h} exp(-E(v,h)). \quad (6)$$

The hidden nodes are conditionally independent given the visible layer nodes, and vice versa. The parameters of RBM can be optimized by performing stochastic gradient ascent on maximizing the log-likelihood of training data. As computing the exact gradient of log-likelihood is intractable, Contrastive Divergence (CD) approximation is used [34] which works fairly well in practice. The three-layers module in this section is formed by stacking RBMs as this way: First an RBM is trained on the receptive field layer. Then, after the training first layer RBM, the weights are frozen and hidden layer act as the input of next layer, and so forth. in the end, a supervised fine tuning step is performed to adjust the weights for improvement on particular task.

III. DEEP NETWORK DETAILS

This section details the network structure and parameters used in our experiments. As preprocessing, all the faces in images are detected automatically by Viola-Jones face detector [35], and then normalized to 32x32 based on the location of eyes.

For the first layer, we sample 6-by-6 pixel patches with a stride of 1 pixel on the 32-by-32 pixel raw images. Thus each image contains 27-by-27 small patches and K (here we set $K = 100$) filters can be learned from all these patches in training set. Then each 32-by-32 image obtains a 27-by-27-by-K representation after convolution. To achieve some extent translation invariance, we apply max-pooling over adjacent, disjoint 3-by-3 patches on each map (an image responses for a single filter). In the end this yields 9-by-9-by-100 features as an over-complete representation for each expression image.

After extracting the representation in the first module, we apply mRMR for AU-aware receptive fields selection. For the parameters referred in Algorithm.1, we set the feature dimension in each AURF as $m = 500$, and the number of AURFs N in each dataset depends on the features relevance to class labels. Our experiments show that CK+ contains less noise than the other two database (e.g. non-uniform expression, wearing accessories, lighting, and so on), so the relevance of features are much higher than that in MMI or SFEW. In practice, based on the mean relevance of each receptive field, we set $N = 9$ in CK+ and $N = 3$ in both MMI and SFEW.

In the last module, we apply two additional RBM layers to learn higher-level features in each AURF. The input layer's node number equals to the dimension of each AURF ($m = 500$). The node number of hidden layers are 400 and 300 respectively. After unsupervised pre-training, the two layers RBM can be further supervisedly fine-tuned using logistic regression. At last, for each AURF, the features in the first visible layer and each hidden layer are concatenated to construct final hierarchical feature for facial expression recognition using linear SVM classifier.

IV. EXPERIMENTS

In this section, we evaluate the learned hierarchical features for facial expression recognition. All comparisons are performed on three datasets: CK+ [21], MMI [22], and a wild expression database SFEW [23]. The experimental results of our method achieve or outperform the state-of-the-art performance.

A. *Experiments on CK+*

The CK+ database consists of 593 sequences from 123 subjects, which is an extended version of Cohn-Kanade (CK) [36] database. The validating emotion labels were only assigned to 327 sequences which were found to meet criteria for one of 7 discrete emotion (Anger (An): 45, Contempt (Co): 18, Disgust (Di): 59, Fear (Fe): 25, Happiness (Ha): 69, Sadness (Sa): 28, and Surprise (Su): 83) based on FACS. In our experiments, we make use of all these sequences from 118 subjects. For each sequence, the first image (neutral face) and three peak frames were used for prototypic expression recognition which is similar to the settings in [6], [11]. Based on the subject ID given in the dataset, we construct 10 person independent subsets by sampling in ID ascending order with step size equals 10 and adopt 10-fold cross-validation. What's more, to avoid parameter sensitivity, only linear SVM classifier is used in all of our experiments. For comparison, Table.I lists the recognition accuracies of several methods, including hand-crafted-feature-based ones and the state-of-the-art performance.

In the table, "**OR**" represents the method based on features of Over-complete Representation; "**AURF**" represents the method based on selected AU-aware Receptive Fields (AURFs); and "**AUDN**" represents the method based on the final features achieved by our deep networks. The CSPL in

TABLE I

Expression Recognition Accuracy on CK+ Database.

Methods	Accuracy	
LBP	83.87%(linear)	81.89%(RBF)
SIFT	86.39%(linear)	87.31%(RBF)
HOG	89.53%(linear)	88.61%(RBF)
Gabor	88.61%(linear)	85.09%(RBF)
CSPL [11]	89.89%(unknown)	
OR	**91.44**%(linear)	
AURF	**92.22**%(linear)	
AUDN	**92.05**%(linear)	

TABLE II

Expression Recognition Accuracy on MMI Database.

Methods	Accuracy	
LBP	52.93%(linear)	50.37%(RBF)
SIFT	57.80%(linear)	61.46%(RBF)
HOG	63.17%(linear)	65.24%(RBF)
Gabor	56.10%(linear)	57.56%(RBF)
CSPL [11]	73.53%(unknown)	
OR	**68.41**%(linear)	
AURF	**69.88**%(linear)	
AUDN	**74.76**%(linear)	

TABLE III

Expression Recognition Accuracy on SFEW Database.

Methods	Accuracy	
LBP	21.29%(linear)	23.71%(RBF)
SIFT	20.45%(linear)	21.14%(RBF)
HOG	19.52%(linear)	22.71%(RBF)
Gabor	19.29%(linear)	19.14%(RBF)
Baseline [23]	19.00%(RBF)	
OR	**24.98**%(linear)	
AURF	**23.00**%(linear)	
AUDN	**26.14**%(linear)	

TABLE IV

The confusion matrix of **AUDN** on CK+ database.

	Ne	An	Co	Di	Fe	Ha	Sa	Su
Ne	95.41	1.83	0.31	0.61	0.61	0	0.61	0.61
An	11.11	81.48	0	6.67	0	0	0.74	0
Co	7.41	1.85	77.78	0	5.56	1.85	5.56	0
Di	4.52	0	0	95.48	0	0	0	0
Fe	9.33	0	0	0	82.67	4	0	4
Ha	0	0	0	0	0.48	99.52	0	0
Sa	20.24	2.38	0	2.38	0	0	71.43	3.57
Su	1.2	0	1.2	0	0	0	0	97.59

TABLE V

The confusion matrix of **AUDN** on MMI database.

	Ne	An	Di	Fe	Ha	Sa	Su
Ne	82.93	7.32	0	1.95	0.98	5.85	0.98
An	7.53	65.59	8.6	3.23	0	15.05	0
Di	2.08	4.17	79.17	0	6.25	8.33	0
Fe	10.71	3.57	0	47.62	7.14	4.76	26.19
Ha	4.76	0	6.35	0	88.89	0	0
Sa	20.83	15.63	3.13	0	0	60.42	0
Su	15	0	0	7.5	0	0	77.5

TABLE VI

The confusion matrix of **AUDN** on SFEW database.

	An	Di	Fe	Ha	Ne	Sa	Su
An	24.11	8.93	15.18	10.71	16.07	16.96	8.04
Di	16.47	14.12	9.41	15.29	18.82	16.47	9.41
Fe	23.23	5.05	20.20	19.19	9.09	10.10	13.13
Ha	15.79	3.51	8.77	50.00	7.89	7.02	7.02
Ne	22.00	8.00	13.00	16.00	23.00	11.00	7.00
Sa	10.10	8.08	10.10	17.17	21.21	23.23	10.10
Su	9.89	9.89	9.89	13.19	23.08	12.09	21.98

[10] gets the accuracy of 89.89%, which represents state-of-the-art performance in CK+ database. It should be pointed out that CSPL only handled six expression categories from 96 subjects, while our experiment considers 8 categories (including Contempt and Neutral (denoted as "Ne")), which is much more challenging and practical compared to many existing methods.

We also compared our "learned feature" with the hand-crafted ones. The experiment settings (the same on Table.I, Table.II and Table.III) are: Image size is 32x32 pixels as same as before. LBP (944 dimensions): 16 patches with size of 8x8 pixels and 59 dimensions uniformed feature on each patch; SIFT (1152 dimensions): 9 lattice points with 128 dimensions feature on each; HOG (1568 dimensions): 49 overlapped blocks with size of 8x8 pixels, 4 cells and 8 histogram bins for each block. Gabor (24576 dimensions): convolutional images with 3 spatial scales and 8 orientations. We performed both linear SVM and RBF SVM on all the features. The grid search for parameter estimation for RBF is performed over $c = 2^k, k = 0, 1, ..., 9; g = 2^l, l = -5, -4, ..., 0$. Some results may have gaps with the existing ones [6], [7], the two main reasons are: (1) The protocols are different. we performed strict person independent test by 10-nonoverlapping-fold cross-validation. (2) The image size is 32x32 pixels, which is much smaller than other methods.

For classification details, we also show the confusion matrix of **AUDN** in Table.IV. We can see that the recognition rates of Contempt and Sadness are not as promising as that of , for example, Happiness and Surprise. This may be caused by the unevenly distribution of data, i.e. some of the categories have much fewer samples. The probability of a sample assigned to such minor categories could be overwhelmed by those of major categories.

B. *Experiments on MMI*

The MMI database [22] includes 30 subjects of both sexes and ages from 19 to 62. In the datasets, 213 sequences have been labeled with six basic expressions, in which 205 sequences are captured frontal view. We use the data from all these 205 sequences as in [11]. Similar to the settings on CK+, the neutral face and three peak frames in each sequence have been used and 10-fold cross-validation is conducted in the same way. We still perform our experiments on all seven expression categories including neutral, which is more difficult compared to six categories problem in [11]. What's more, in both CK+ and MMI experiments, the images we used are 32x32 pixels which are much smaller than 96x96 in [11], nevertheless, better results are achieved using the proposed method. Table.II demonstrate the comparisons of several methods and the confusion matrix of **AUDN** are shown in Table.V.

Compared to the results on CK+, the recognition performance degrade significantly on MMI database due to its challenging conditions: The subjects posed expressions non-uniformly, and many of them wear accessories (e.g. glasses, moustache). Thus the traditional hand-crafted features cannot offer discriminative information as such noises exist. However, our scheme apply the data-dependent dictionary rather than fixed descriptors, which is more robust dealing with the complex conditions. It shows that gradually better results have been achieved during our learning process.

C. *Experiments on SFEW*

For further validation, we apply our method to a much more challenging scenario: the expression in the wild. The Static Facial Expression in the Wild (SFEW) database which has been extracted from movies is different from the available facial expression datasets generated in highly controlled lab environments. It is the first attempt to build database depicting real-world or simulated real-world conditions for expression analysis [23]. According to Strictly Person Independent (SPI) Protocol for SFEW, the database is divided into two sets. Each set contains seven subfolders corresponding to the seven expression categories. The sets were created in strict person independent manner that there is no overlap between training and testing set. In total, there are 700 images (346 in Set1, 354 in Set2) and 95 subjects. The experiment is set to be two-fold (Fold1: train on Set1 and test on Set2; Fold2: train on Set2 and test on Set1) and we take the average accuracy of two folds for measurement.

As shown in Table.III and Table.VI, our method outperforms the baseline 19% [23] significantly. However, due to the tough imaging conditions, the faces normalized by automatically detected eyes location suffer from severe misalignment, so that none of the algorithms can work well as on CK+ or MMI.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose to construct a deep architecture especially for facial expression recognition, which is called "AUDN". Inspired by the interpretation of FACS, we applying an mRMR feature selection on an over complete representation, the obtained AURFs are shown to be able to simulate specific AUs. Additional multi-layer RBMs can extract higher-level features in each AURF and further fine tuned by logistic regression. Linear SVM classifier is applied to hierarchical feature and final result is obtained by averaging recognition results of each AURF. The proposed AUDN achieve or outperform state-of-the-art performance on three facial expression database covering both lab control and wild scenario. In future, we will try to search more complex combination of feature rather than simple non-overlapping AURF to further boost the performance of our method.

VI. ACKNOWLEDGMENTS

This paper is partially supported by Natural Science Foundation of China under contracts No. 61025010 and No. 61222211.

REFERENCES

[1] P. Ekman and W. V. Friesen. Facial action coding system. *Consulting Psychologists Press*,1,1978.

[2] C. E. Izard. The face of emotion. *New York: Appleton-Century-Crofts*,1971.

[3] Y. Tian, T. Kanade, and J. F. Cohn. Recognizing upper face action units for facial expression analysis. *CVPR*, 2000.

[4] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *TPAMI*, 29(10): 1683-1699, 2007.

[5] M. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *TPAMI*, 21(12): 1357-1362, 1999.

[6] C. Shan, S.Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27: 803-816, 2009.

[7] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 2006.

[8] U. Tariq, K. Lin, Z. Li, X. Zhou, Z. Wang, V. Le, T. S. Huang, X. Lv, and T. X. Han. Emotion recognition from an ensemble of features. *FG*, 2011.

[9] Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. Huang. Multi-view facial expression recognition. *FG*, 2008.

[10] P. Yang, Q. Liu, and D. N. Metaxas. Exploring facial expression with compositional features. *CVPR*, 2010.

[11] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. *CVPR*, 2012.

[12] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786): 504-507, 2006.

[13] A. Rao, and N. Thiagarajan. Recognizing facial expressions from videos using Deep Belief Networks. *Technical Report*, 2009.

[14] D. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. High-performance neutral networks for visual object classification. *Pre-print*, 2011. http://arxiv.org/abs/1102.0183.

[15] A. Coates and A. Y. Ng. The importance of encoding versus training with sparse coding and vector quantization. ICML, 2011.

[16] D. Scherer, A. Mller, and S. Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. *ICANN*, 2010.

[17] Y. LeCun, F. Huang, and L. Bottou. Learning methods for generic object recognition with invarience to pose and lighting. *CVPR*, 2004.

[18] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *ICML*, 2009.

[19] A. Coates and A. Y. Ng. Selecting receptive fields in deep networks. *NIPS*, 2011.

[20] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for Deep Belief Nets. *Neural Computation*, 18(7): 1527-1554, 2006.

[21] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, and Z. Ambadar. The Extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. *CVPRW*, 2010.

[22] M. F. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the MMI facial expression database. *LREC*, 2010.

[23] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. *ICCVW*, 2011.

[24] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin. LIBLINEAR: A Library for Large Linear Classification, *JMLR*, 9(2008): 1871-1874, 2008. Software available at http://www.csie.ntu.edu.tw/ cjlin/liblinear.

[25] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. *CVPR*, 2007.

[26] J. Mutch and D. G. Lowe. Object class recognition and localization using sparse features with limited receptive fields. *IJCV*, 56(6): 503-511, 2008.

[27] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *CVPR*, 2006.

[28] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. *ICCV*, 2005.

[29] A. K. Jain, R. P. W. Duin, and J. Mao. Statistcial pattern recognition: A review. *TPAMI*, 22(1): 4-37, 2000.

[30] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *PRL*, 15(11): 1119-1125, 1994.

[31] J. Jaeger, R. Sengupta, and W. L. Ruzzo. Improved gene selection for classification of microarrays. *PSB*, 2003.

[32] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *CSB*, 2003.

[33] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *TPAMI*, 27(8): 1226-1237, 2005.

[34] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14: 1771-1800, 2002.

[35] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 2001.

[36] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. *FG*, 2000.