Computer Vision and Image Understanding 147 (2016) 95–101

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Learning prototypes and similes on Grassmann manifold for spontaneous expression recognition

Mengyi Liu, Ruiping Wang, Shiguang Shan*, Xilin Chen

Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China

ARTICLE INFO

Article history:
Received 28 March 2015
Accepted 17 August 2015

Keywords:
Expression prototype
Simile representation
Grassmann manifold
Spontaneous expression recognition

ABSTRACT

Video-based spontaneous expression recognition is a challenging task due to the large inter-personal variations of both the expressing manners and the executing rates for the same expression category. One of the key is to explore robust representation method which can effectively capture the facial variations as well as alleviate the influence of personalities. In this paper, we propose to learn a kind of typical patterns that can be commonly shared by different subjects when performing expressions, namely “prototypes”. Specifically, we first apply a statistical model (i.e. linear subspace) on facial regions to generate the specific expression patterns for each video. Then a clustering algorithm is employed on all these expression patterns and the cluster means are regarded as the “prototypes”. Accordingly, we further design “simile” features to measure the similarities of personal specific patterns to our learned “prototypes”. Both techniques are conducted on Grassmann manifold, which can enrich the feature encoding manners and better reveal the data structure by introducing intrinsic geodesics. Extensive experiments are conducted on both posed and spontaneous expression databases. All results show that our method outperforms the state-of-the-art and also possesses good transferable ability under cross-database scenario.

1. Introduction

In recent years, facial expression recognition has become a popular research field due to its wide applications in many areas such as biometrics, psychological analysis, human-computer interaction, and so on. In the early stage, many works have been done to classify human posed expressions in static images [1]. However, as facial expression can be viewed as a sequentially dynamic process, it is natural and proved to be more effective to be recognized from video clips [2–5]. For spontaneous expression recognition in video, one of the main challenges is the large inter-personal variations of expressing manners and executing rates for the same expression category. The key issue to cope with the challenge is to develop a more robust representation for facial expression, which can better capture the subtle facial variations as well as alleviate the influence of personalities in performing expression.

According to the theory from physiology and psychology, facial expressions are the outcome of facial muscle motions over various time intervals. When captured by cameras, an observed expression can be decomposed into a set of local appearance variations produced by the motions occurring in different facial regions. In spite of the large inter-personal variations, there still exist some typical motion patterns, that can be commonly shared by different subjects in performing expressions. The similar idea is also reflected in a pioneering work Facial Action Coding System (FACS) [6], where a number of Action Units (AU) are manually defined to describe some emotion-related facial actions aroused by muscle motions. Then each expression is represented by the existence of these AUs in a binary coding manner.

In light of such theory, we propose to explore a batch of commonly shared typical patterns, i.e. “prototypes”, using data-driven approach, and then design a prototype-based encoding manner to generate the feature representation for each sample. An schema of our basic idea is illustrated in Fig. 1. Specifically, we first apply a statistical model (i.e. linear subspace) on facial regions to model the local variations of local patterns, which can generate the specific expression patterns for each video sample. Then a clustering algorithm is employed on all these expression patterns, and each cluster mean can be regarded as a “prototype”, which integrates the common properties of the samples assigned to this cluster. Note that, all of the original patterns and the learned “prototypes” are represented as linear (orthogonal) subspaces lying on Grassmann manifold, thus intrinsic geodesic distance [7] and Karcher means [8] are employed in this procedure for accurate estimation. To obtain the unified prototype-based representation, we further design “simile” features to measure the similarities of personal specific patterns to our learned “prototypes” on Grassmann manifold. The idea is derived from [9] for face verification, which

* Corresponding author. fax: +86 10 6260 0548.
E-mail address: sgshan@ict.ac.cn (S. Shan).

http://dx.doi.org/10.1016/j.cviu.2015.08.006
1077-3142/© 2015 Elsevier Inc. All rights reserved.

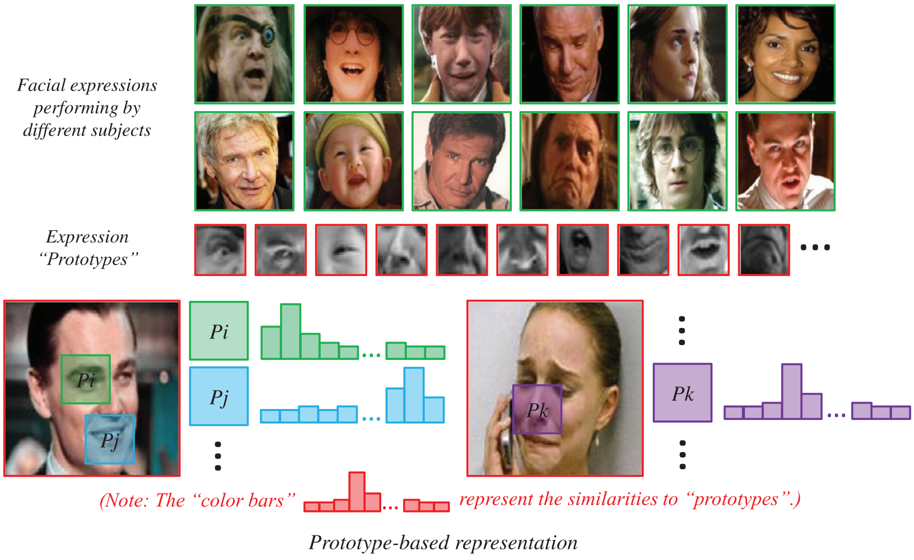

Facial expressions performing by different subjects
Expression "Prototypes"
Pi
Pj
Pk
(Note: The "color bars" represent the similarities to "prototypes".)
Prototype-based representation

Fig. 1. An schema of our basic idea (best viewed in color).

assumed that an unseen face can be described as having a mouth that *looks like* *A*'s and a nose that *looks like* *B*'s, where *A* and *B* are individuals in the reference set. In our method, the static facial attributes are replaced by some dynamic variation manners of facial regions when performing expressions. However, different from [9] which introduced an auxiliary reference set, we measure the similarities referring to the "prototypes" directly explored from the data, thus brings favorable robustness against the bias in the construction of reference set.

The main contributions of this paper are summarized as follows: (1) We propose a novel approach for modeling expression patterns and learning "prototypes" using statistical model on Grassmann manifold; (2) "Similes" are designed to explore the relations among common prototypes and specific patterns, which provides a new viewpoint to analyze the generality and specificity in the manner of human performing spontaneous expressions. (3) Comprehensive experiments are conducted with different parameter settings. The transferable ability of prototypes and similes are further discussed in cross-database test.

The rest of this paper is structured as follows. Section 2 reviews several most related work for video-based facial expression recognition. Section 3 introduces the essential components of our proposed method, including facial expression patterns, prototypes learning, and simile representation. In Section 4, we provide comprehensive evaluations on the whole framework as well as discussing the important parameters. Finally, we conclude the work and discuss possible future efforts in Section 5.

2. Related works

For video-based facial expression recognition, there is always strong interest in modeling the temporal dynamics among video frames. The mainstream approaches of dynamic representation are based on local spatial-temporal descriptors. For example, Yang et al. [3] designed Dynamic Binary Patterns (DBP) mapping based on Haar-like features. Zhao et al. [2] proposed LBP-TOP to extract the spatial-temporal information from three orthogonal planes (i.e. X–Y, X–T, Y–T) in image volumes. Hayat et al. [10] conducted a comprehensive evaluation based on various descriptors, e.g. HOG3D [11], HOG/HOF [12], 3D SIFT [13], using bag of features framework for facial expression recognition. All these hand-crafted methods possess favorable computational efficiency and generalization ability due to the independency of data.

Another line of methods attempt to explore the specific characteristics in expression evolution using dynamic graphic models. For instance, Shang et al. [14] employed a non-parametric discriminant Hidden Markov Model (HMM) on tracked facial features to for dynamic expressions modeling. Jain et al. [15] proposed to model the temporal variations within facial shapes using Latent-Dynamic Conditional Random Fields (LDCRFs), which can obtain the entire video prediction and continuously frame labels simultaneously. Wang et al. [4] proposed Interval Temporal Bayesian Networks (ITBN) to represent the spatial dependencies among primary facial actions as well as the variety of time-constrained relations, which characterize the complex activities both spatially and temporally. Although these schemes can better reveal the intrinsic principles of facial expressions, the optimization requires lots of domain knowledge and large computational cost.

More recently, statistical models were employed to encode the appearance variations occurring in dynamic facial expressions, which proved to be more effective when dealing with real-world data [16,17]. In [16], linear subspace were applied on the feature set of successive image frames to model the facial feature variations during the temporal evolution of expression. For more robust modeling, [17] integrated three different types of statistics into the framework, i.e. linear subspace, covariance matrix, gaussian distribution, to model the feature variations from different perspectives. As these statistical models all reside on Riemannian manifold, intrinsic geodesics or extrinsic kernel methods were exploited to perform representation and classification on Riemannian manifold instead of traditional Euclidean space. A common property of these two methods is to model the facial variation globally, which could be easily affected by misalignment or partial occlusion. Since localization has been proved to be more natural and effective when processing face-related application, in this paper, we focus on modeling local variation based on statistical models and extracting generic local dynamic patterns to unify the description of facial expressions performed by different subjects. We believe that this scheme can make balance between the generalization ability by considering repeatable local features and the modality specificity by learning directly from data.

3. The proposed method

In this section, we introduce the proposed method in three stages: facial expression patterns generation, prototypes learning, and simile representation.

3.1. *Facial expression pattern*

To consider locality, we focus on modeling facial patches rather than the whole region. Specifically, suppose the patch size is s-by-s pixels, which results in s^2-dimensional gray feature vector, we collect patches from the same location of all image frames in the video to construct a feature set $F = \{f_1, f_2, \ldots, f_T\}$, where $f_t \in R^{s^2}$ denotes the patch features of the tth frame, and T is the number of frames in the video. The set can be statistically represented by a linear subspace $X \in R^{s^2 \times r}$ via SVD on its feature covariance matrix as:

$$\sum_{t=1}^{T} (f_t - \overline{f})(f_t - \overline{f})^T = X \Lambda X^T, \tag{1}$$

where $X = [x_1, x_2, \ldots, x_r]$, x_j is the jth leading eigenvector, and r is the dimension of the linear subspace. For all the feature sets of patches over different facial regions from different videos, we can obtain a collection of X as $\{X_i | i = 1, 2, \ldots, N\}$. Since the linear subspace encodes the feature correlations among successive frames, which describes the temporal variations occurring on facial regions, it can serve as a kind of expression features, the so-called *facial expression pattern* in this paper. Based on these basic patterns, we design the following algorithm to obtain more compact and descriptive representation for expression recognition.

3.2. *Prototypes learning*

In this paper, “prototypes” are defined to be some kind of typical patterns describing common facial variations which can be shared by different subjects when performing expressions. To discover prototypes automatically from data, we utilize a clustering algorithm on all the facial expression patterns to obtain a bank of centroids, each of which can be regarded as an “exemplar” of the patterns in its cluster. However, the linear subspaces $\{X_i | i = 1, 2, \ldots, N\}$ are a collection of points residing on Grassmann manifold [7], traditional clustering algorithm in Euclidean space cannot be directly applied. Following [18], we employ the spectral clustering based on the affinity matrix calculated by Grassmann kernels. Specifically, the similarity between two linear subspaces X_i and X_j can be measured by projection kernel [7] as:

$$A_{ij} = ||X_i^T X_j||_F^2, \tag{2}$$

where A is the affinity (Grassmann kernel) matrix for spectral clustering on Grassmann manifold. As in [19], the affinity matrix is first normalized by:

$$L = D^{-1/2} A D^{-1/2}, \quad \text{where } D_{ii} = \sum_j A_{ij}. \tag{3}$$

Then the top l eigenvectors of L are computed to construct a new matrix U. By treating each row of U as a point in R^l, we apply the general K-means in the new embedded space and obtain the cluster assignment, which can also be easily made correspondence to the original facial expression patterns X_i on Grassmann manifold. The whole clustering procedure is summarized in Algorithm 1.

According to the cluster assignment, we can calculate the centroid of each cluster to generate the “prototype”. To obtain a more accurate estimation of the cluster mean, we employ the Karcher mean [8,20] which considers the intrinsic metric on Grassmann manifold specifically for linear subspaces. Formally, given the set of facial expression patterns in the kth cluster C_k: $\{X_{k_n} | X_{k_n} \in C_k\}$, the Karcher mean is defined to be a point residing on the manifold which minimizes the sum of squared geodesic distances [8]:

$$\widehat{X} = \arg\min_{X \in \mathcal{M}} \sum_{k_n} d_g^2(X_{k_n}, X), \tag{4}$$

where $\mathcal{M}$ denotes the Grassmann manifold, and $d_g : \mathcal{M} \times \mathcal{M} \rightarrow R^+$ is the geodesic distance defined on the manifold. Specifically, d_g can be measured by exponential map $exp_X(\cdot)$ and logarithm map $log_X(\cdot)$, to switch between the manifold and its tangent space at the point X. Thus $\widehat{X}$ is the solution to $\sum_{k_n} log_X(X_{k_n}) = 0$, which can be solved iteratively as in Algorithm 2. After this step, we obtain K cluster means $\widehat{X}_1, \widehat{X}_2, \ldots, \widehat{X}_K$ to serve as the “prototypes”. An illustration of prototypes learning is shown in Fig. 2.

Algorithm 1 Spectral clustering on Grassmann manifold.

```
Input:
    Facial expression patterns extracted from all videos: $\{X_i | i = 1, 2, \ldots, N\}$
Output:
    $K$ clusters: $\{C_k | k = 1, 2, \ldots, K\}$, where $C_k = \{X_{k_n} | X_{k_n} \in C_k\}$

Compute the affinity (kernel) matrix $A \in R^{N \times N}$ where $A_{ij} = ||X_i^T X_j||_F^2$.
Define $D$ as the diagonal matrix where $D_{ii} = \sum_j A_{ij}$, and compute the normalized matrix $L = D^{-1/2} A D^{-1/2}$.
Find the top $l$ eigenvectors of $L$: $u_1, u_2, \ldots, u_l$.
Form a matrix $U$ containing $u_1, u_2, \ldots, u_l$ as columns and renormalize to unit norm by $U_{ij} = U_{ij} / (\sum_j U_{ij}^2)^{1/2}$.
Take each of the $N$ row of $U$ as a point in $R^l$ and apply K-means to these $N$ points, where the $i$th point corresponds to $X_i$.
Assign the original $X_i$ to $C_k$ if the $i$th row of $U$ is assigned to cluster $k$.
```

Algorithm 2 Karcher mean in cluster.

```
Input:
    Facial expression patterns in $C_k$: $\{X_{k_n} | X_{k_n} \in C_k\}$
Output:
    Karcher mean of the $k$th cluster: $\widehat{X}_k$

Randomly select a sample from $C_k$ as the initial Karcher mean $\widehat{X}_k^{(0)}$
Set iteration index $p = 0$
while $p < max\_iter$ do
    For each $X_{k_n}$, compute the tangent vector:
    $V_{k_n} = log_{\widehat{X}^{(p)}}(X_{k_n})$
    Compute the mean vector $\overline{V_k} = \sum_n V_{k_n} / \#n$ in tangent space
    if $||\overline{V_k}||^2 < epsilon(a\ small\ value)$,
        break;
    else
        Move $\overline{V_k}$ back onto the manifold:
        $\widehat{X}_k^{(p+1)} = exp_{\widehat{X}_k^{(p)}}(\overline{V_k})$
    end if
    $p = p + 1$
end while
```

3.3. *Simile representation*

In order to obtain a unified representation based on the prototypes, “simile” features are designed to explore the relationships among the common prototypes and specific patterns, which explicitly considers the generality and specificity in the manner of human performing expressions. Specifically, the relationship between two samples is presented by calculating their similarity on Grassmann manifold. Given a sample video containing M specific facial expression patterns, for each pattern, we calculate the its similarities to all the K prototypes via the projection kernel mentioned above, thus results in $M*K$ similarities to construct a “simile” representation. Different from the traditional features extracted using descriptors, each dimension of the simile possesses a kind of mid-level semantics due to the characterizing of “relationship”. Such simile feature is believed

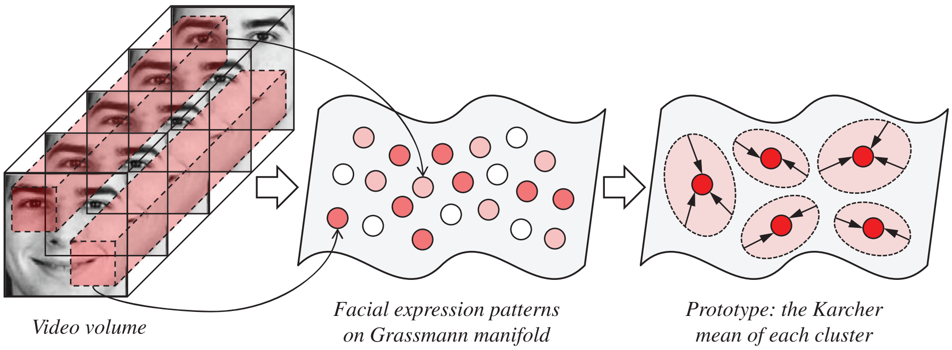

Video volume
Facial expression patterns on Grassmann manifold
Prototype: the Karcher mean of each cluster

Fig. 2. An illustration of the expression prototypes learning. For each video volume, we first generate a number of specific facial expression patterns via linear subspace modeling. Then a clustering algorithm is employed on all patterns extracted from different videos, to obtain a bank of "prototype" calculated by Karcher mean of each cluster. The prototypes encode the typical properties of the basic motion patterns, which can be commonly shared by different subjects when performing expressions.

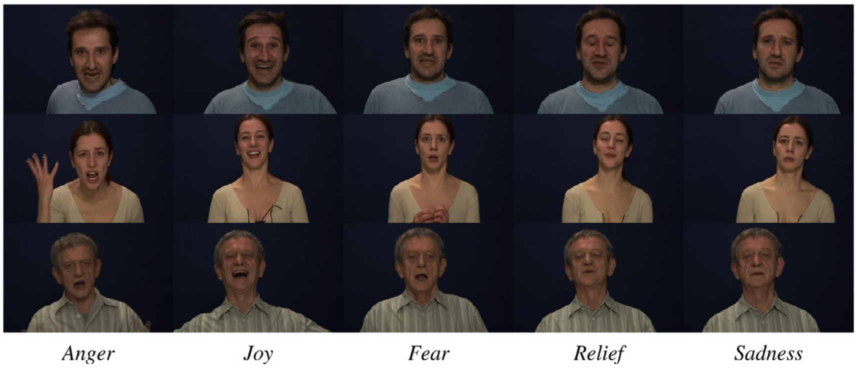


Fig. 3. The sample facial expression images extracted from FERA database.

to be more compact and descriptive, which is also proved in our experiments.

3.4. Discussions

The most related work of our method are [16] and [17], which can be regarded as a special case of the proposed framework. Specifically, the final features in [16] and [17] are kernel matrix calculated by taking the training samples as the "prototypes"; and for each sample, the specific patterns degenerate to a single global pattern. It is obvious that the local modeling manner can capture more detailed information and expect to be less sensitive to misalignment and occlusion. Moreover, with the property of repeatability among different samples, the local patterns are more robust to intra-class variation and much easier to be generalized to other data sources.

Another similar work is *Bag of Words (BoW)*, which also generate a bank of common prototypes (i.e. codebook) using clustering algorithm. The difference is that BoW is always based on low-level features (i.e. words) and the final feature only estimates the occurrence of each codeword, while our method provides more descriptive mid-level "words" and more elaborate representation "simile". However, from a broader view, our method can be regarded as a fine-assignment version of BoW conducting on Grassmann manifold. Thus the possible future effort may be to extend methods in the family of BoW, e.g. fisher vector, to Grassmann manifold or some other non-Euclidean space.

4. Experiments

4.1. Data and protocols

CK+ database [21] contains 593 videos of 123 different subjects, which is an extended version of CK database. All of the image sequences vary in duration from 10 to 60 frames and start from the neutral face to the peak expression. Among these videos, 327 sequences from 118 subjects are annotated with the seven basic emotions (i.e. Anger (An), Contempt (Co), Disgust (Di), Fear (Fe), Happy (Ha), Sadness (Sa), and Surprise (Su)) according to FACS [6].

MMI database [22] includes 30 subjects of both sexes and ages from 19 to 62. In the database, 213 image sequences have been labeled with six basic expressions, in which 205 are with frontal face. Different from CK+, the sequences in MMI cover the complete expression process from the onset apex, and to offset. In general, MMI is considered to be more challenging for the subjects usually wear some accessories (e.g. glasses, mustaches), and there are also large inter-personal variations when performing the same expression.

FERA database is a fraction of the GEMEP corpus [23] that has been put together to meet the criteria for a challenge on facial AUs and emotion recognition. As the labels on test set are unreleased, we only use the training set for evaluation. The training set includes seven subjects, and 155 sequences have been labeled with five expression categories: Anger (An), Fear (Fe), Joy (Jo), Sadness (Sa), and Relief (Re). FERA is more challenging than CK+ and MMI because the expressions are **spontaneous** in natural environment. Fig. 3 shows some examples from FERA database.

AFEW database is collected from movies showing close-to-real-world conditions [24]. Here we use a subset of AFEW which is provided for EmotiW2013 [25]. According to the protocol, the videos are divided into three sets: training, validation, and testing. The task is to classify a sample video into one of the seven expression categories: Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Neutral (Ne), Sadness (Sa), and Surprise (Su). As far as we know that AFEW is considered to be the most challenging emotion database due to the extremely large varitions in the wild. Fig. 4 shows some examples from FERA database.

We also provide the number of samples for each expression category of the four databases in Table 1. For evaluation, we adopt the strictly person-independent protocols on all these databases.

Fig. 4. The sample facial expression images extracted from AFEW database.

Table 1
The number of samples for each expression of the four database.

	An	Co	Di	Fe	Ha	Ne	Re	Sa	Su	Total
CK+	45	18	59	25	69	–	–	28	83	327
MMI	31	–	32	28	42	–	–	32	40	205
FERA	32	–	–	31	30	–	31	31	–	155
AFEW	117	–	90	104	127	118	–	116	104	776

Table 2
Mean accuracy with different parameters on four databases.

(a) CK+						
Image scale	SVM			PLS		
	K = 64	K = 128	K = 256	K = 64	K = 128	K = 256
32x32	83.5	84.4	89.4	81.2	87.1	90.2
48x48	81.4	85.2	86.1	83.4	88.4	88.8
64x64	82.9	85.2	84.2	87.9	89.0	**90.9**
(b) MMI						
Image scale	SVM			PLS		
	K = 64	K = 128	K = 256	K = 64	K = 128	K = 256
32x32	55.6	52.8	59.4	54.7	54.3	59.4
48x48	57.5	57.5	58.9	58.9	59.3	60.1
64x64	56.5	58.9	56.5	58.6	**61.5**	60.0
(c) FERA						
Image scale	SVM			PLS		
	K = 64	K = 128	K = 256	K = 64	K = 128	K = 256
32x32	58.6	58.1	66.5	54.1	59.4	62.0
48x48	64.6	65.2	62.7	65.1	64.4	63.9
64x64	66.0	**67.9**	64.2	64.7	65.4	64.1
(d) AFEW						
Image scale	SVM			PLS		
	K = 64	K = 128	K = 256	K = 64	K = 128	K = 256
32x32	23.9	25.9	25.5	24.2	24.2	25.1
48x48	24.6	25.7	25.6	25.8	25.3	26.1
64x64	25.8	26.4	26.1	25.0	26.2	**27.5**

In detail, experiments are performed based on leave-one-subject-out cross validation on CK+ and 10-fold cross validation in MMI. As these labels on test set are unknown for both FERA and AFEW, we conduct leave-one-subject-out cross validation on FERA's training set and two-fold cross-validation on the training-validation set on AFEW.

4.2. Evaluation of the parameters

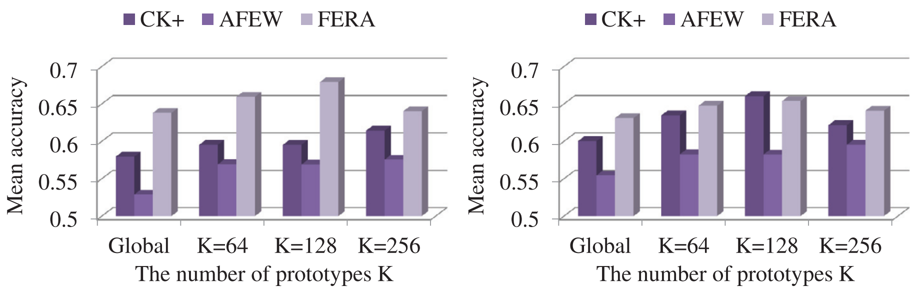
For preprocessing, we first detect the face in each video frames and normalize them based on the locations of two eyes. In the facial expression patterns generation step, the patch size is fixed as 16-by-16 (i.e. $s = 16$) pixels with the sampling step of eight pixels, while the global facial image scale is a tunable parameter varying in 32, 48, 64. To calculate the linear subspace, each patch are represented as a 256-dimension gray feature vector, and the dimension of subspace r is fixed as 15 in all cases. Another important parameter in our framework is the number of prototypes, i.e. the number of clusters K in Algorithm 1. In our experiment, we discuss three different values of K as 64, 128, 256, to answer the two following questions: (1) How many prototypes are enough for representing the typical patterns generally or is it the more the better? (2) Are there any differences between lab-controlled data and spontaneous data regarding to the selection of K?

We evaluation the effects of these parameters on all the four databases. And for the final recognition, we employ two kinds of classifiers: linear SVM and one-vs-rest PLS [16]. The experimental results are listed in Table 2, from where we can obtain several important observations as follows: (1) The larger image scale, i.e. 64x64, consistently performs better when employing PLS classifier. For SVM, the same trend can also be observed when setting $K = 64, 128$, however, there are always exceptions when $K = 256$ on most of the databases. Generally, the image with a larger scale can capture more subtle appearance variations on faces, which benefits the prototypes learning even with a smaller K. (2) The larger K usually performs well on most of the databases except for FERA with the image scale of 48x48 and 64x64. We conjecture the reason may be that a larger number of prototypes leads to a more elaborate description of the observable data, which is easy to cause overfitting due to the lack of training data (i.e. only about 130 videos in each fold) on FERA. (3) For the two kinds of classifiers, PLS usually performs better than SVM, perhaps the one-vs-rest manner can especially deal with several difficult and confusion categories, thus contributes more in calculating the mean accuracy over all classes.

4.3. Comparison with related work

In this section, we compare our method with the most related work, i.e. the global modeling strategy proposed in [16], which generate a single linear subspace for each video sample. For further boosting performance, we conduct a decision level fusion of the results obtained by employing multiple image scales. As shown in Table 3, our method significantly outperforms the global model in all cases.

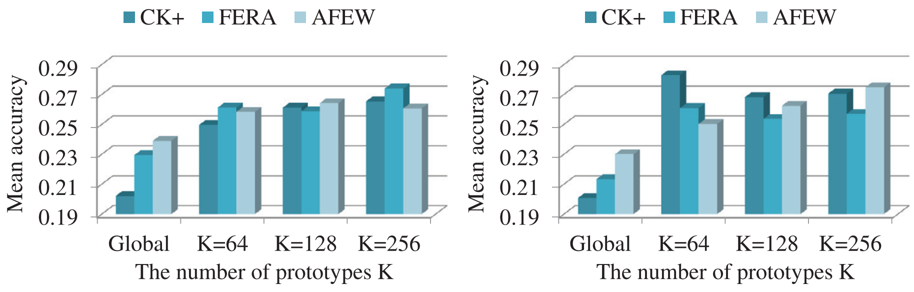
In addition, we also conduct comparison between the global model and proposed prototypes regarding to the transferable ability in cross-database test. Specifically, for each **target** video, we first extract its specific facial expression patterns, then generate the simile features according to the prototypes calculated from other **source** data. In global scheme, each prototype is degenerated to a linear subspace modeling of the whole video sample. For evaluation, we



(a) SVM

(b) PLS

Fig. 5. Mean accuracy on FERA with different classifiers. (a) SVM (b) PLS.



(a) SVM

(b) PLS

Fig. 6. Mean accuracy on AFEW with different classifiers. (a) SVM (b) PLS.

Table 3
Mean accuracy (mAcc) and overall accuracy (Acc) on four databases.

(a) CK+				
Method	SVM		PLS	
	mAcc	Acc	mAcc	Acc
Global model	83.6	89.9	85.7	91.1
Prototype & Simile	88.2	92.1	**94.6**	**96.3**
(b) MMI				
Method	SVM		PLS	
	mAcc	Acc	mAcc	Acc
Global model	56.6	59.5	56.4	60.0
Prototype & Simile	**64.5**	**67.3**	63.7	66.3
(c) FERA				
Method	SVM		PLS	
	mAcc	Acc	mAcc	Acc
Global model	63.8	63.9	63.1	63.2
Prototype & Simile	69.1	69.0	**70.4**	**70.3**
(d) AFEW				
Method	SVM		PLS	
	mAcc	Acc	mAcc	Acc
Global model	23.9	23.9	23.0	23.9
Prototype & Simile	28.3	29.5	**29.3**	**30.9**

take two spontaneous databases, i.e. FERA and AFEW, as the **target** data. The corresponding **source** data are CK+/AFEW for FERA, and CK+/FERA for AFEW. By fixing the image scale as 64x64 and varying K in 64, 128, 256, we can obtain all the results as listed in Tables 4 and 5. For easy of comparison, we also provide the results obtained by using the same source and target in the last line of each table.

According to Figs. 5 and 6, we can observe that for global model, the performance degrade significantly when employing different source and target data. While for our method, prototypes learned from other source data show favorable transferable ability on both FERA and AFEW. Even on AFEW, the results based on CK+ outperform that obtained by using its own data for training. The reason may be that AFEW is in real-world conditions with undesirable noises coming from both image quality and manual annotations, thus makes it challenging to explore the purified expression-related information, however which can be provided by the prototypes learned from CK+. This may also inspire us that spontaneous expressions also share some general variation patterns with the posed expressions. Since so far the spontaneous data are difficult to collect and annotate, we can make full use of the prototype knowledge derived from the large amount of well-organized lab-controlled data.

4.4. Comparison with state-of-the-art

Finally we conduct comparison with state-of-the-art methods as shown in Table 6, where “PSG” is short for our method “Prototype and Simile on Grassmann manifold”. For all methods, we directly cite the results reported in their publications, except for the “Liu13∗[16]” on AFEW, which is implemented by using gray features as other databases rather than the convolutional features in the original method. Compared to the results on other databases,the recognition

Table 4
Mean accuracy on FERA with prototypes transferring from different sources.

Sources	SVM				PLS			
	Global	K = 64	K = 128	K = 256	Global	K = 64	K = 128	K = 256
CK+	57.9	59.5	59.5	61.4	60.0	63.5	66.1	62.2
AFEW	52.9	56.9	56.9	57.6	55.4	58.3	58.2	59.5
FERA	63.9	66.0	**67.9**	64.2	63.1	64.7	65.4	64.1

Table 5
Mean accuracy on AFEW with prototypes transferring from different sources.

Sources	SVM				PLS			
	Global	K = 64	K = 128	K = 256	Global	K = 64	K = 128	K = 256
CK+	20.2	25.0	26.1	26.5	20.1	**28.3**	26.8	27.1
FERA	22.9	26.1	25.9	27.4	21.3	26.1	25.4	25.7
AFEW	23.9	25.8	26.4	26.1	23.0	25.0	26.2	27.5

Table 6
Comparison with state-of-the-art on four databases.

(a) CK+					
	Chew11[26]	Lucey10[21]	Chew12[27]	**PSG**$_{SVM}$	**PSG**$_{PLS}$
mAcc	74.4	83.3	89.4	88.2	**94.6**
Acc	82.3	88.3	–	92.1	**96.3**
(b) MMI					
	Wang13[4]	Wang13[4]	Liu14[28]	**PSG**$_{SVM}$	**PSG**$_{PLS}$
mAcc	51.5	59.7	62.2	**64.5**	63.7
Acc	–	60.5	63.4	**67.3**	66.3
(c) FERA					
	Ptucha[29]	Chew12[27]	Liu14[28]	**PSG**$_{SVM}$	**PSG**$_{PLS}$
mAcc	56.6	65.6	56.3	69.1	**70.4**
Acc	–	–	56.1	69.0	**70.3**
(d) FERA					
	Dhall13[25]	Liu13∗[16]	**PSG**$_{SVM}$	**PSG**$_{PLS}$	
mAcc	26.7	23.9	28.3	**29.3**	
Acc	27.3	23.9	29.5	**30.9**	

performance degrades significantly on AFEW due to its challenging image conditions, e.g. large variations of pose and illumination.

5. Conclusions

In this paper, we present a novel framework for spontaneous facial expression recognition in videos. To handle the large inter-personal variations, we propose to learn a bank of typical patterns named “prototypes”, which can be commonly shared by different subject when performing expressions. Accordingly, we further design simile features to model the relationships between the individually specific patterns and generally common prototypes. Our experiments on both posed and spontaneous data demonstrate the effectiveness of the method. The evaluation of transferable ability also provides inspiration on how to deal with spontaneous expression recognition in real-world by utilizing large available lab-controlled data. In the future, we will continually focus on this issue as well as explore more descriptive features for a robust representation.

Acknowledgment

This work is partially supported by 973 Program under contract No. 2015CB351802, Natural Science Foundation of China under contract nos. 61390511, 61222211, and 61379083, and Youth Innovation Promotion Association CAS No. 2015085.

References

[1] M. Pantic, L.J.M. Rothkrantz, Automatic analysis of facial expressions: The state of the art, IEEE Trans. Patt. Anal. Mach. Intell. 22 (12) (2000) 1424–1445.
[2] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Patt. Anal. Mach. Intell. 29 (6) (2007) 915–928.
[3] P. Yang, Q. Liu, X. Cui, D.N. Metaxas, Facial expression recognition using encoded dynamic features, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. CVPR, 2008.
[4] Z. Wang, S. Wang, Q. Ji, Capturing complex spatio-temporal relations among facial muscles for facial expression recognition, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2013.
[5] M. Liu, S. Shan, R. Wang, X. Chen, Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2014.
[6] P. Ekman, W.V. Friesen, Facial action coding system: a technique for the measurement of facial movement, Consulting Psychologists Press (1978).
[7] J. Hamm, D.D. Lee, Grassmann discriminant analysis: a unifying view on subspace-based learning, in: Proceedings of International Conference on Machine learning, 2008.
[8] H. Karcher, Riemannian center of mass and mollifier smoothing, Commun. Pure Appl. Math. 30 (5) (1977) 509–541.
[9] N. Kumar, A.C. Berg, P.N. Belhumeur, S.K. Nayar, Attribute and simile classifiers for face verification, in: Proceedings of IEEE International Conference on Computer Vision, 2009.
[10] M. Hayat, M. Bennamoun, A. El-Sallam, Evaluation of spatiotemporal detectors and descriptors for facial expression recognition, in: Proceedings of International Conference on Human System Interactions (HSI), 2012.
[11] A. Klaser, M. Marszalek, A spatio-temporal descriptor based on 3d-gradients, in: Proceedings of International Conference on British Machine Vision Association, 2008.
[12] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[13] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of international conference on Multimedia, ACM, 2007.
[14] L. Shang, K.-P. Chan, Nonparametric discriminant hmm and application to facial expression recognition, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[15] S. Jain, C. Hu, J.K. Aggarwal, Facial expression recognition with temporal modeling of shapes, in: Proceedings of IEEE International Conference Computer Vision Workshops (ICCV Workshops), 2011.
[16] M. Liu, R. Wang, Z. Huang, S. Shan, X. Chen, Partial least squares regression on grassmannian manifold for emotion recognition, in: Proceedings of ACM on International Conference on Multimodal Interaction, ACM, 2013.
[17] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, X. Chen, Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild, in: Proceedings of ACM on International Conference on Multimodal Interaction, ACM, 2014.
[18] S. Shirazi, M.T. Harandi, C. Sanderson, A. Alavi, B.C. Lovell, Clustering on grassmann manifolds via kernel embedding with application to action analysis, in: Proceedings of IEEE International Conference on Image Processing (ICIP), 2012.
[19] A.Y. Ng, M.I. Jordan, Y. Weiss, et al., On spectral clustering: analysis and an algorithm, in: Proceedings of IEEE Neural Information Processing Systems (NIPS), 2002.
[20] P. Turaga, A. Veeraraghavan, A. Srivastava, R. Chellappa, Statistical computations on Grassmann and stiefel manifolds for image and video-based recognition, IEEE Trans. Patt. Anal. Mach. Intell. 33 (11) (2011) 2273–2286.
[21] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, 2010.
[22] M. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition to the mmi facial expression database, in: Proceedings of IEEE Leading Real Estate Companies of the World (LRECW), 2010.
[23] T. Bänziger, K.R. Scherer, Introducing the geneva multimodal emotion portrayal (gemep) corpus, Blueprint for Affective Computing: A Sourcebook, 2010, pp. 271–294.
[24] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Collecting large, richly annotated facial-expression databases from movies, IEEE MultiMed. 19 (3) (2012) 34–41.
[25] A. Dhall, R. Goecke, J. Joshi, M. Wagner, T. Gedeon, Emotion recognition in the wild challenge 2013, in: Proceedings of International Conference on Multimodal Interaction, ACM, 2013.
[26] S.W. Chew, P. Lucey, S. Lucey, J. Saragih, J.F. Cohn, S. Sridharan, Person-independent facial expression detection using constrained local models, in: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG 2011), 2011.
[27] S.W. Chew, S. Lucey, P. Lucey, S. Sridharan, J.F. Conn, Improved facial expression recognition via uni-hyperplane classification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[28] M. Liu, S. Li, S. Shan, R. Wang, X. Chen, Deeply learning deformable facial action parts model for dynamic expression analysis, in: Proceedings of IEEE Conference on Asian Conference on Computer Vision, 2014.
[29] R. Ptucha, G. Tsagkatakis, A. Savakis, Manifold based sparse representation for robust expression recognition without neutral subtraction, in: Proceedings of IEEE International Conference Computer Vision Workshops (ICCV Workshops), 2011.