# Learning Expressionlets on Spatio-Temporal Manifold for Dynamic Facial Expression Recognition

Mengyi Liu[1,2], Shiguang Shan[1], Ruiping Wang[1], Xilin Chen[1,3]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences (UCAS), Beijing, 100049, China
[3]Department of Computer Science and Engineering, University of Oulu, Finland
mengyi.liu@vipl.ict.ac.cn, {sgshan, wangruiping, xlchen}@ict.ac.cn

## Abstract

*Facial expression is temporally dynamic event which can be decomposed into a set of muscle motions occurring in different facial regions over various time intervals. For dynamic expression recognition, two key issues, temporal alignment and semantics-aware dynamic representation, must be taken into account. In this paper, we attempt to solve both problems via manifold modeling of videos based on a novel mid-level representation, i.e.* ***expressionlet****. Specifically, our method contains three key components: 1) each expression video clip is modeled as a spatio-temporal manifold (STM) formed by dense low-level features; 2) a Universal Manifold Model (UMM) is learned over all low-level features and represented as a set of local ST modes to statistically unify all the STMs. 3) the local modes on each STM can be instantiated by fitting to UMM, and the corresponding expressionlet is constructed by modeling the variations in each local ST mode. With above strategy, expression videos are naturally aligned both spatially and temporally. To enhance the discriminative power, the expressionlet-based STM representation is further processed with discriminant embedding. Our method is evaluated on four public expression databases, CK+, MMI, Oulu-CASIA, and AFEW. In all cases, our method reports results better than the known state-of-the-art.*

## 1. Introduction

Automatic facial expression recognition plays an important role in various applications, such as Human-Computer Interaction (HCI) and diagnosing mental disorders. Early research mostly focused on expression analysis from static facial images [20]. However, as facial expression can be better described as the sequential variation in a dynamic process, recognizing facial expression from video is more natural and proved to be more effective in recent research works [34, 32, 2, 33, 22].

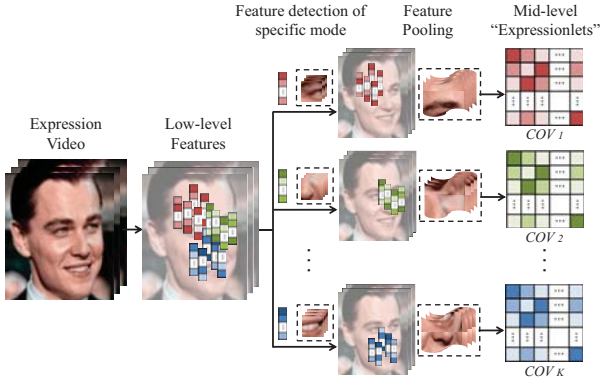



Figure 1. A schematic illustration of constructing the mid-level representation – the proposed "expressionlets" ("COV" is short for "covariance matrix"). Each strip stands for a local ST feature, and the K feature modes (similar to codewords) are pre-learned and modeled via GMM.

Among these video-based facial expression recognition methods, one of the main concerns is how to effectively encode the dynamic information in videos. Currently, the mainstream approaches of dynamic representation are based on local spatio-temporal features like LBP-TOP [34], HOG 3D [14]. These local descriptors extracted in local cuboid are then pooled over the whole video or some hand-crafted segments, to obtain a representation with certain length independent of time resolution. As the low-level features possess the property of repeatability, integrating them with pooling leads to robustness to intra-class variations and deformations of different expression styles. However, this kind of technique lacks of consideration of two important issues: 1) **Temporal alignment**. Expressions are inherently dynamic events consisting of onset, apex, and offset phases. Intuitively, the recognition should conduct matching among corresponding phases, which thus requires globally temporal alignment among different sequences. The rigid pooling has inevitably dropped those sequential relations and temporal correspondences. 2) **Semantics-aware dynamic**

1063-6919/14 $31.00 © 2014 IEEE
DOI 10.1109/CVPR.2014.226

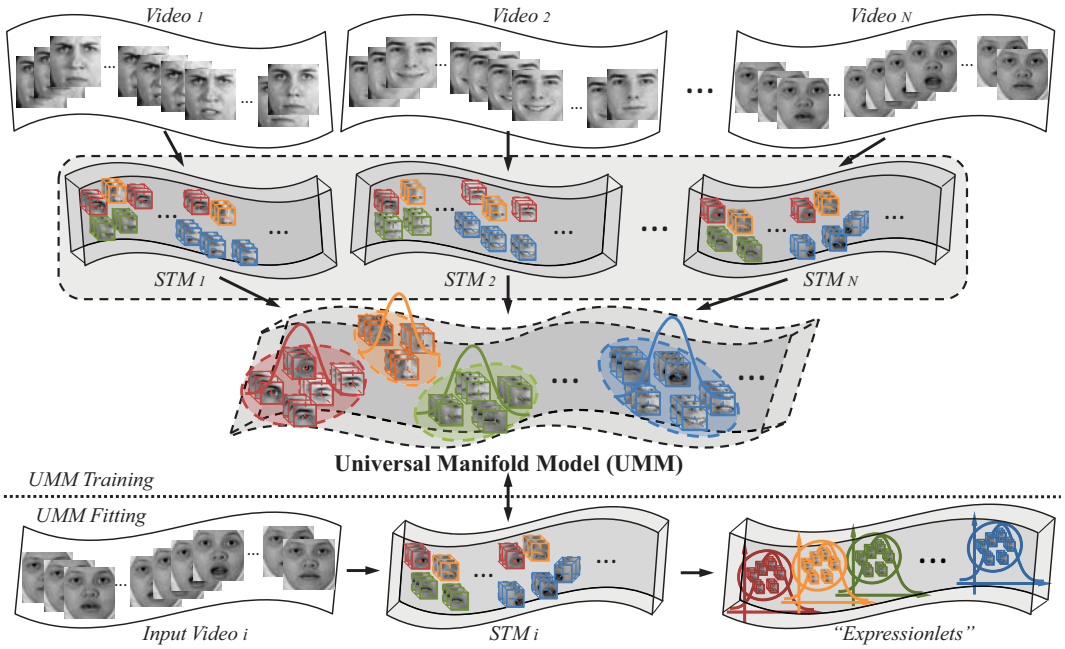



Figure 2. The schema of the proposed method. Given an individual video clip, we intend to model it as a Spatio-Temporal Manifold (STM) spanned by local spatio-temporal features, which however leads to difficulty of aligning different STMs. To statistically unify and thus facilitate the alignment of STMs, we propose a Universal Manifold Model (UMM), represented as a number of universal local ST modes, which can be learned by EM-like methods. With UMM constructed, the local modes on each STM can be instantiated by fitting to UMM and thus aligned mutually, then the corresponding expressionlet is built to model the variations in each local ST mode. Thus we obtain an expressionlet-based representation of STM. Please note that, for UMM training, we exploit both appearance and spatio-temporal location information of the local features in order to enforce some degree of locality both spatially and temporally.

**representation**. Each expression can be decomposed into a group of semantic action units, which exhibit in different facial regions with varying sizes and last for different lengths of time. Since the manually designed cuboids can only capture low-level information short of representative and discriminative ability, they are incapable of modeling the expression dynamic in higher semantic level.

In this paper, we attempt to take both issues into account via spatio-temporal manifold modeling based on a set of mid-level representations, i.e. **expressionlets**. The proposed mid-level expressionlet is a kind of modeling that aims to characterize the variations among a group of low-level features as shown in Figure 1. The notation "-let" means that it serves as a local (both spatially and temporally) dynamic component within a whole expression process, which shares similar spirit with "motionlet" [27] in action recognition community. Thus expressionlet bridges the gap between low-level features and high-level semantics desirably. Specifically, given an individual video clip, we first model it as a Spatio-Temporal Manifold (STM) spanned by its low-level features. To conduct spatio-temporal alignment among STMs, we build a Universal Manifold Model (UMM), represented as a number of universal local ST modes, which can be learned by EM-like methods on all possible low-level features. By fitting to UMM, the local modes on each STM can be instantiated respectively and all of the different STMs are inherently well-aligned to UMM via these corresponding modes. Finally, our expressionlet is constructed by modeling each local mode on STM. To characterize the correlations and variations among low-level features within each mode, the expressionlet is represented as the covariance matrix of the feature set in a statistic manner, which also makes it robust to local misalignment [25, 11, 29].

To further enhance the discriminative ability of expressionlet, we perform a discriminant learning with these mid-level representations on all of the STMs. Inspired by [28], while only considering the "margin" among corresponding expressionlets, we exploit a graph-embedding [31] method by constructing partially connected graphs to keep the links between expressionlets with the same semantics. In the end, the embedded features are correspondingly concatenated into a long vector as the final manifold (video) representation for classification. Hence, the proposed expressionlet has the following characteristics: 1) **Flexible spatio-temporal range**. i.e. varying sizes of spatial regions and temporal durations. 2) **Variation modeling**. It encodes the local variations caused by expression using a covariance matrix. 3) **Discriminative ability**. It is descriptive and contains category information for recognition.

The proposed method is evaluated on four public expression databases: CK+ [19], MMI [26], Oulu-CASIA VIS [33], AFEW [6], including both lab-controlled and wild scenarios. The discussion of parameters shows the benefits of considering the two crucial issues mentioned above, i.e. temporal alignment and semantics-aware dynamic representation. All of our results achieve the state-of-the-art.

## 2. Expressionlet Construction

In this section, we present how to generate the semantic mid-level representation, i.e. expressionlet, for video representation. The whole process is illustrated in Figure 2 and summarized in Algorithm 1.

### 2.1. Spatio-Temporal Manifold (STM)

For clarification, we first present the spatio-temporal manifold (STM) for modeling each video clip. The STM is spanned by 3D (i.e. spatio-temporal) blocks densely sampled from the video volume, which cover a variety of local variations in both spatial and temporal space. Instead of hand-crafted descriptors e.g. LBP-TOP, HOG 3D, we employ a bank of learned filters to extract low-level features from the blocks. Such kind of features are directly learned from data and proved to be more adaptive and generalizable to the data domain [15].

We adopt $K$-means on the smaller-scale cuboids densely sampled from all the training videos. The cluster centroids are used to form a bank of filters to be applied to the spatio-temporal blocks. To obtain a translation invariant represention, we conduct a 3D convolution followed by a max-pooling operation over adjacent spatial regions. The extracted feature is denoted as $a_{xyt}$, where $x, y, t$ are spatio-temporal index of the block on the STM.

### 2.2. Universal Manifold Model (UMM) – a statistic model of STMs

In this work, Universal Manifold Model (UMM) is defined to statistically model the STMs from different people with different expressions. As a person-independent and expression-independent model, UMM facilitates the robust parameterized modeling of the STMs. Inspired by [10, 16], we employ a Gaussian Mixture Model (GMM) to learn the UMM by estimating the appearance and location distribution of all the 3D block features.

To consider the manifold structure information, for all the blocks we augment the appearance features with their spatio-temporal coordinates, i.e. $f = \{a_{xyt}, x/w, y/h, t/l\}$, where $a_{xyt}$ is the appearance feature of the block located at $\{x, y, t\}$, and $w, h, l$ are the numbers of blocks on width, height and time length direction on the STM. Simply we train a GMM with spherical Gaussian components, i.e.,

$$P(f|\Theta) = \sum_{k=1}^{K} w_k G(f|\mu_k, \sigma_k^2 I), \quad (1)$$

where $\Theta = (w_1, \mu_1, \sigma_1, ..., w_K, \mu_K, \sigma_K)$; $K$ is the number of Gaussian mixture components; $I$ is identity matrix; $w_k, \mu_k, \sigma_k$ are the mixture weight, mean, and diagonal covariance of the $k$-th Gaussian component $G(f|\mu_k, \sigma_k)$.

We use typical Expectation Maximization (EM) algorithm to estimate the paremeters of GMM by maximizing the likelihood of the training feature set. After training the UMM, each Gaussian component builds correspondence of a group of block features from different STMs, which constitute a local ST mode universally.

**Algorithm 1 : Expressionlet Construction**

**Input:**
Low-level feature sets of each STM: $F^1, ..., F^N$

**Output:**
Experssionlet sets of each STM: $E^1, ..., E^N$

1: Initialize GMM parameter: $\Theta = \{(\omega_k, \mu_k, \sigma_k)\}$
2: Use EM algorithm to learn optimal GMM parameters: $\Theta^* = argmax_\Theta \sum_{i,b,k} \omega_k G(f_b^i|\mu_k, \sigma_k^2 I)$
3: **for** i:=1 to N **do**
4: **for** k:=1 to K **do**
5: Find top $T$ block features $F_k^i = \{f_{k_t}^i\}_{t=1}^T$ with the largest $T$ probabilities on $G_k$: $G(f_{k_t}^i|\mu_k^*, (\sigma_k^*)^2 I) > G(f_{k_{t+1}}^i|\mu_k^*, (\sigma_k^*)^2 I)$
6: Calculate $k$-th expressionlet of $F^i$ as: $C_k^i = \frac{1}{T-1} \sum_{t=1}^T (f_{k_t}^i - \overline{f_k^i})(f_{k_t}^i - \overline{f_k^i})^T$
7: **end for**
8: $E^i = \{C_1^i, C_2^i, ..., C_K^i\}$
9: **end for**
10: **return** $\Theta^*, E^1, ..., E^N$

### 2.3. Expressionlet Modeling

The UMM learned above can be regarded as a container with K-components GMM. Then, given any STM, we aim to formulate it as a parameterized instance of the UMM. For this purpose, our basic idea is assigning some of the local ST features of the STM into the K Gaussian "buckets" and further modeling the distribution of the local features in each Gaussian bucket with their covariance matrix.

Formally, an expression manifold $M^i$ can be presented as block feature set $F^i = \{f_1^i, ..., f_{B_i}^i\}$, where $B_i$ is the number of features on $M^i$. For the $k$-th Gaussian component, we can calculate the probabilities of each $f_b^i$ in $F^i$ as

$$P_k^i = \{p_k(f_b^i) \mid p_k(f_b^i) = w_k G(f_b^i|\mu_k, \sigma_k^2 I)\}_{b=1}^{B_i}. \quad (2)$$

We sort the block features $f_b^i$ in descending order of $P_k^i$, and the features with the largest $T$ probabilities are selected for the $k$-th local model construction, which can be represented as $F_k^i = \{f_{k_1}^i, ..., f_{k_T}^i\}$.

Considering the correlations and variations among the features in a local model, we calculate the covariance matrix of set $F_k^i$ as the representation of an expressionlet:

$$C_k^i = \frac{1}{T-1}\sum_{t=1}^{T}(f_{k_t}^i - \overline{f_k^i})(f_{k_t}^i - \overline{f_k^i})^T, \tag{3}$$

where $\overline{f_k^i}$ is the mean of the block features in set $F_k^i$. The diagonal entries of $C_k^i$ represent the variance of each individual feature, and the non-diagonal entries are their respective correlations. As the expressionlets are globally aligned via UMM, the covariance modeling can provide a desirable locally tolerance to spatial-temporal misalignment.

In the end, the manifold $M^i$ is represented as a set of expressionlets, i.e. $E^i = \{C_1^i, C_2^i, ..., C_K^i\}$. An overall procedure is summarized in Algorithm 1.

### 2.4. Discussion

In this section, we compare our "Expressionlets" with two well-known works: Action Units (AU) [7] and Bag-of-Visual-Words (BoVW) [4].

**Expressionlets vs. Action Units (AU)** Action Units (AU) are fundamental actions of individual or groups of facial muscles for encoding facial expression based on Facial Action Coding System (FACS). Similarly, our expressionlets are designed to model expression variations over local spatio-temporal regions in the same spirit as AUs. However, there are two differences between expressionlets and AUs. 1) AUs are manually defined concepts that are independent of person and category, while expressionlets are some mid-level representations extracted from data using learning scheme. 2) According to FACS, each expression is encoded by a certain number of AUs. Instead of the binary coding manner, in our method, an expression can be represented by various expressionlet patterns which provide more flexible and rich information.

**Expressionlets vs. Bag-of-Visual-Words (BoVW)** In our method, we extract dense local ST features and construct a codebook (by GMM), in which each codeword can be considered as a representative of several similar local features. Both of the two operations are typical steps in BoVW framework. In pooling stage, BoVW estimates histogram(s) of occurrences of each codeword over the whole image or separated regions. However, in our method, different from simply counting the number, we make use of the second-order statistics by estimating the covariance of all the local features (augmented with location information) falling into each bucket (codeword). In this way, the local features are pooled to keep more variations, under the constraint of expressionlet. In addition, in our method, by limiting the number ($T$ in Algorithm 1) of local features falling into each bucket, not all local features are necessarily taken into account by the second-order pooling, which is also different from traditional BoVW methods. We believe such a strategy can alleviate the influence of unexpected noise or signal distortions (e.g. caused by occlusion).

## 3. Discriminant Learning with Expressionlets

In this section, we attempt to enhance the discriminative power of expressionlets. As the expressionlet possess the property of spatio-temporal locality, an effective way is to consider the "margin" among corresponding expressionlets instead of globally. By utilizing the graph-embedding [31] framework, we formulate our learning scheme as follows.

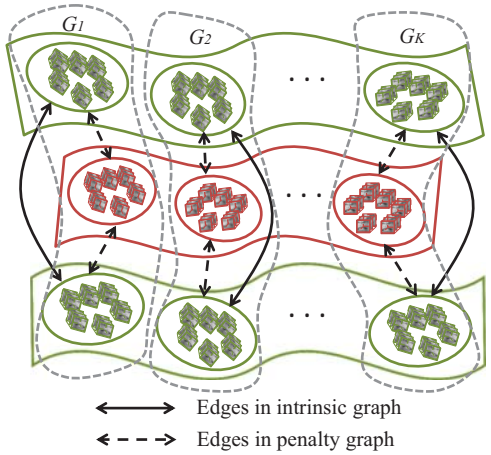



Figure 3. The adjacency relationships of the intrinsic and penalty graphs for the discriminative learning with expressionlets ($G_i$ represents the $i$-th component of GMM).

In the overall expressionlet set $E$, given the $m$-th expressionlet, which corresponds to the $p$-th mode on $M^i$, denoted as $C_p^i$ and the $n$-th expressionlet, which corresponds to the $q$-th mode on $M^j$, denoted as $C_q^j$ (if all STMs are ordered, we can denote $m = (i-1) * K + p$ and similarly $n = (j-1) * K + q$. The indices $m$ and $n$ are used for better illustration), with the class label $l_i, l_j$ for $M_i, M_j$ respectively, the intrinsic graph $W_w$ and penalty graph $W_b$ can be defined as (See Figure 3):

$$W_w(m,n) = \begin{cases} 1, \ if \ l_i = l_j, and \ p = q \\ 0, \ otherwise \end{cases} \tag{4}$$

$$W_b(m,n) = \begin{cases} 1, \ if \ l_i \neq l_j, and \ p = q \\ 0, \ otherwise \end{cases} \tag{5}$$

We aim to learn an embedding function $\phi$ to maximize the discriminative power while simultaneously preserve the

correspondence of expressionlets from the same Gaussian component. According to $W_w$ and $W_b$, the within-class scatter $S_w$ and between-class scatter $S_b$ can be defined as:

$$S_w = \sum_{m,n} Dis(\phi(C_p^i), \phi(C_q^j)) W_w(m,n), \quad (6)$$

$$S_b = \sum_{m,n} Dis(\phi(C_p^i), \phi(C_q^j)) W_b(m,n), \quad (7)$$

where $Dis(\phi(C_p^i), \phi(C_q^j))$ denotes the distance between two embedded expressionlets $\phi(C_p^i)$ and $\phi(C_q^j)$. Here the expressionlets $C_p^i, C_q^j$ are Symmetric Positive Definite (SPD) matrices (i.e. nonsingular covariance matrices), lie on a Riemannian manifold. We exploit a Log-Euclidean Distance (LED) [1] to project these points to Euclidean vector space, where standard vector learning methods are ripely studied, as advocated in [29].

Given two covariance matrices $C_p^i, C_q^j$, the LED metric between them is defined as:

$$d_{LED}(C_p^i, C_q^j) = ||log(C_p^i) - log(C_q^j)||_F. \quad (8)$$

Let $C = U\Sigma U^T$ be the eigen-decomposition of SPD matrix $C$, its $log$ can be computed by

$$log(C) = Ulog(\Sigma)U^T. \quad (9)$$

Thus we can obtain a vector representation $x_m$ of the $m$-th expressionlet, i.e. $C_p^i$, where $x_m$ is a vector spanned by $log(C_p^i)$. Now let us return to the embedding function, simply consider a linear projection $v$, we can reformulate the embedded features and the distance between them in classical Euclidean space as

$$\phi(C_p^i) = v^T x_m, \phi(C_q^j) = v^T x_n, \quad (10)$$

$$Dis(\phi(C_p^i), \phi(C_q^j)) = ||v^T x_m - v^T x_n||^2. \quad (11)$$

Accordingly, we only need to learn the projection $v$ instead of $\phi$, by maximizing the between-class scatter $S_b$ while minimizing the within-class scatter $S_w$:

$$v_{opt} = \arg\max \frac{v^T X(D_b - W_b)X^T v}{v^T X(D_w - W_w)X^T v}, \quad (12)$$

where $D_w$ and $D_b$ are diagonal matrices with diagonal elements $D_w(m,m) = \sum_n W_w(m,n)$ and $D_b(m,m) = \sum_n W_b(m,n)$. Let $L_w$ and $L_b$ be the Laplacian matrices of two graph $W_w$ and $W_b$. The columns of an optimal $v$ are the generalized eigenvectors corresponding to the $l$ ($l = 100$ in this paper) largest eigenvalues in

$$XL_bX^T v = \lambda XL_wX^T v. \quad (13)$$

With the learned embedding function $\phi$, the $K$ expressionlets from $M_i$ can be represented as $\{\phi(C_1^i), ..., \phi(C_K^i)\}$. These $K$ features are concatenated as a long vector for the final expression manifold (video) representation. In the end, we use multi-class linear SVM implemented by Liblinear [8] for classification.

# 4. Experiments

Our method is evaluated on four databases. The experimental details are shown and analyzed in this section.

## 4.1. Databases and protocols

**CK+ Database**.The CK+ database [19] consists of 593 sequences from 123 subjects, which is an extended version of Cohn-Kanade (CK) [13] database. The image sequence vary in duration from 10 to 60 frames and incorporate the onset (neutral face) to peak formation of the facial expression. The validated expression labels are only assigned to 327 sequences which are found to meet the criteria for 1 of 7 discrete emotions (Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise) based on Facial Action Coding System (FACS). Based on the subject ID in the database, we constructed 10 person-independent subsets by sampling in ID ascending order with a step size of 10 to adopt 10-fold cross-validation as in several previous work [24, 9, 35, 18].
**MMI Database**. The MMI database [26] includes 30 subjects of both sexes and ages from 19 to 62. In the database, 213 sequences have been labeled with six basic expressions, in which 205 sequences were captured frontal view. Each of the sequence reflects the whole temporal activation patterns (onset $\rightarrow$ apex $\rightarrow$ offset) of a single facial expression type. In our experiments, all of these data were used as in [22, 35] and also a person-independent 10-fold cross-validation was conducted in the same way with CK+. Compared to CK+, MMI is thought to be more challenging for the subjects pose expressions non-uniformly and usually wear some accessories (e.g. glasses, moustache).
**Oulu-CASIA VIS Database**. The Oulu-CASIA VIS database [33] is a subset of the Oulu-CASIA NIR-VIS database, in which all the videos were taken under the visible (VIS) light condition. We evaluated our method only on the normal illumination condition (i.e. strong and good lighting). It includes 80 subjects between 23 and 58 years old, with six basic expressions (i.e. anger, disgust, fear, happiness, sadness, and surprise) of each person. Each video starts at a neutral face and ends at the apex of expression as the same settings in CK+. Similar to [33] and [9], we adopted person-independent 10-fold cross-validation scheme on the total 480 sequences.
**AFEW Database**. For further validation, we apply our method to a much more difficult wild scenario. The Acted Facial Expression in Wild (AFEW) [6] database has been collected from movies showing close-to-real-world conditions, which depicts or simulates the spontaneous expressions in uncontrolled environment. Some exemplar images are shown in Figure 4. According to the protocols defined in Emotion Recognition in the Wild Challenge (EmotiW 2013) [5], the database is divided into three sets: training, validation and test. The task is to classify each video clip into one of the seven expression categories (six basic class-

Figure 4. Exemplar images from AFEW database, which contains large variations in pose, facial expression, illumination, etc.

es plus a neutral class). As the ground truth of test set still remains unreleased, here we only report our results on validation set for comparison.

### 4.2. Discussion of Parameters

For preprocessing, all the faces images are normalized to 48x48 pixels based on the locations of two eyes. In the STM construction step, we learned 100 spatio-temporal 3D filters with the size of 9x9x3 pixels (See Figure 5). The low-level 3D blocks are 16x16x3 pixels samples with a stride of 4 pixel in spatial dimension and 1 frame in temporal dimension. Adopted convolution and 4x4 spatial pooling operations, each 3D block yields a 2x2x100-dimension vector. As the estimation of covariance may be inaccurate when the dimension is much larger than the number of samples $T$, we first reduce the dimension of block features from 403 to 100 via PCA. And the dimension of covariance (i.e. expressionlet) is reduced to $l = 100$ after discriminant learning.

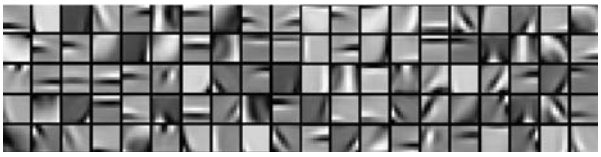
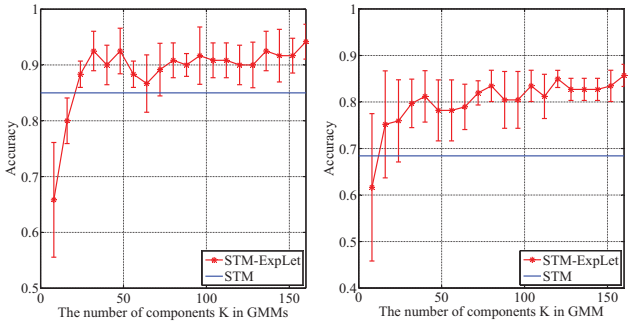


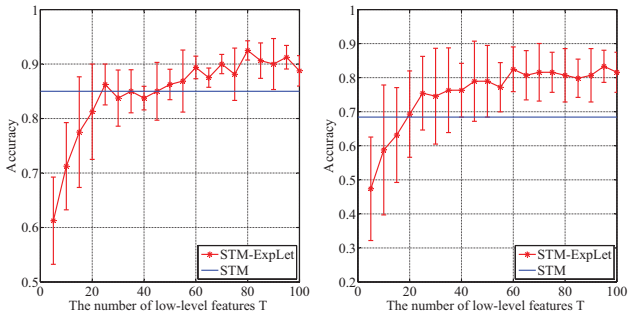
Figure 5. Examples of the spatio-temporal filters. For ease of visualization. we ignore the temporal dimension and only show the middle frame.

To evaluate the effectiveness of the proposed expressionlet, we compared our complete framework which has expressionlet learning scheme (denoted as "**STM-ExpLet**") with our baseline method which used the single global covariance matrix over the whole STM as feature (denoted as "**STM**"). Two key parameters in the expressionlet construction step were also studied: 1) The number of components $K$ in GMM, which is also the number of expressionlets used to represent a video clip. 2) The number of low-level features $T$ used to calculate covariance matrix for each expressionlet. The relations between recognition performance and each of the two parameters are presented in Figure 6 and Figure 7 respectively. As for validation purpose only, we




(a) (b)

Figure 6. Recognition accuracy with different number of components $K$ in GMM. (a) CK+ database. (b) MMI database.




(a) (b)

Figure 7. Recognition accuracy with different number of low-level features $T$ to construct an expressionlet. (a) CK+ database. (b) MMI database.

conduct such experiments on one fold of CK+ and MMI.

In Figure 6, the red curves are the mean accuracy and standard deviation with respect to the number of components $K$ in GMM, where for a certain choice of K, the mean and standard deviation are calculated by varying $T$. As shown, as the number of GMM component $K$ increases, the recognition accuracy gradually improves and the standard deviation also decreases accordingly. Similarly, the mean accuracy and its standard deviation with respect $T$ is shown in Figure 7 (the mean and standard deviation are calculated by varying $K$ under each certain T). We can also observe a rising trend of the curve as $T$ increases. This demonstrates the benefits brought from both the global temporal alignment by GMM and local variation modeling by expressionlet. We tuned $K$ and $T$ according to the validation results in all the experiments of four databases. The typical settings of these parameters are $K = 176$, $T = 30$ in CK+, MMI, Oulu-CASIA VIS, and $K = 88$, $T = 50$ in AFEW.

We also conduct comparisons with four popular local spatio-temporal descriptors: LBP-TOP [34], 3D SIFT [23], HOG 3D [14], HOE [27]. All the experiments were conducted using their released code and the parameters are tuned to better-adapt to our task. For LBP-TOP, we divided the spatial image into non-overlapping 6x6 blocks (i.e.

each block is 8x8 pixels), the local features from each block (i.e. volume in spatio-temporal space) were concatenated to create one feature vector. For 3D SIFT, the video was divided into five equal volumes according to temporal length, in each volume, four interested points were sampled to calculate the feature. All the features from 4x5 points were concatenated to one vector. For HOG 3D and HOE, we first divided the video into 6x6x4 volumes, then each volume was further divided into 2x2x2 grids. The descriptors were applied on each grids by computing histogram of oriented gradient, where the orientation are quantized into 8 bins. Similar to the others above, the overall feature vector can be obtained by concatenating all these local features. In classification stage, we used a multi-class linear SVM by feeding all these comparative features as well as our proposed features ("**STM**" and "**STM-ExpLet**") to the classifier.

### 4.3. Results Comparisons

We demonstrate the proposed method for expression recognition on four databases: CK+, MMI, Oulu-CASIA VIS, and AFEW. As shown in Table 1,2,3,4, we separate the results into three categories with horizontal lines. The first block shows the experimental results of four local spatio-temporal descriptors, which are obtained by using exactly the same data and protocols with ours. The second block shows the state-of-the-art results directly cited from their publications. The third block lists two methods proposed in this paper: "**STM**" and "**STM-ExpLet**".

| Method | Accuracy(%) |
|---|---|
| 3D SIFT [23] | 81.35 |
| HOE [27] | 82.26 |
| LBP-TOP [34] | 88.99 |
| HOG 3D [14] | 91.44 |
| ITBN [30] (15-fold) | 86.3 |
| CERT [17] | 87.21 |
| MCF [3] (LOSO) | 89.4 |
| MSR [21] | 91.4 |
| TMS [12] (4-fold) | 91.89 |
| Cov3D [22] (5-fold) | 92.3 |
| **Ours STM** | **91.13** |
| **Ours STM-ExpLet** | **94.19** |

Table 1. Experimental Results on CK+ Database.

For CK+ (Table.1), many state-of-the-art methods adopted different protocols on the same data. ITBN [30] performed a 15-fold cross subject validation; CERT [17] only took a subset of the database for evaluation; MCF [3] adopted a Leave-One-Subject-Out (LOSO) cross-validation; TMS [12] adopted a 4-fold cross-validation; and Cov3D [22] adopted a 5-fold cross-validation. For MMI (Table.2), ADL [24], CPL [35], and CSPL [35] conducted the experiments

| Methods | Accuracy(%) |
|---|---|
| HOE [27] | 54.15 |
| LBP-TOP [34] | 59.51 |
| HOG 3D [14] | 60.89 |
| 3D SIFT [23] | 64.39 |
| ADL [24] | 47.78 |
| HMM [30] | 51.5 |
| ITBN [30] | 59.7 |
| CPL [35] | 49.36 |
| CSPL [35] | 73.53 |
| **Ours STM** | **65.37** |
| **Ours STM-ExpLet** | **75.12** |

Table 2. Experimental Results on MMI Database.

| Methods | Accuracy(%) |
|---|---|
| 3D SIFT [23] | 55.83 |
| HOE [27] | 61.25 |
| LBP-TOP [34] | 68.13 |
| HOG 3D [14] | 70.63 |
| AdaLBP [33] | 73.54 |
| Atlases [9] | 75.52 |
| **Ours STM** | **68.96** |
| **Ours STM-ExpLet** | **74.59** |

Table 3. Experimental Results on Oulu-CASIA VIS Database.

| Methods | Accuracy(%) |
|---|---|
| HOE [27] | 19.54 |
| 3D SIFT [23] | 24.87 |
| LBP-TOP [34] | 25.13 |
| HOG 3D [14] | 26.90 |
| EmotiW [5] | 27.27 |
| **Ours STM** | **29.19** |
| **Ours STM-ExpLet** | **31.73** |

Table 4. Experimental Results on AFEW Database.

using the same data (i.e. 205 sequences) and same protocols (i.e. person-independent 10-fold cross-validation) with ours. However they only considered several manually selected apex frames for recognition, which need more ground truth information compared to our settings. HMM [30] and ITBN [30] both adopted 20-fold cross subject validation for classification based on whole sequence. For the last two database, i.e. Oulu and AFEW, the same evaluate protocol was used for all comparison methods and the results are shown in Table 3 and Table 4 respectively. We can see that on all the databases "**STM**" outperforms most of the hand-crafted local descriptors. As the mid-level representation, i.e. expressionlet, provides more accurate feature correspondence and more effective local appearance encoding, the proposed "**STM ExpLet**" achieves even more convincing results on all the databases.

## 5. Conclusion

In this paper, we propose a new method for dynamic facial expression recognition. By considering two critical issues of the problem, i.e. temporal alignment and semantics-aware dynamic representation, a kind of variation modeling is conducted among well-aligned spatio-temporal regions to obtain a group of expresssionlets, which serve as the mid-level representations to bridge the gap between low-level features and high-level semantics. As evaluated on four state-of-the-art facial expression benchmarks, the proposed expressionlet has shown its superiority over traditional methods for video based facial expression recognition. As the framework is quite general and not limited to the task of expression recognition, an interesting direction in the future is to exploit its applications in other video related vision tasks, such as action recognition and object tracking.

## Acknowledgement

The work is partially supported by Natural Science Foundation of China under contracts nos.61025010, 61222211, 61390511, 61379083, and the FiDiPro program of Tekes.

## References

[1] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007.

[2] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, and S. Sridharan. Person-independent facial expression detection using constrained local models. In *FG*, 2011.

[3] S. W. Chew, S. Lucey, P. Lucey, S. Sridharan, and J. F. Conn. Improved facial expression recognition via uni-hyperplane classification. In *CVPR*, 2012.

[4] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCVW*, 2004.

[5] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *ICMI*, 2013.

[6] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):0034, 2012.

[7] P. Ekman and W. V. Friesen. Facial action coding system: a technique for the measurement of facial movement. *Consulting Psychologists Press*, 1978.

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[9] Y. Guo, G. Zhao, and M. Pietikäinen. Dynamic facial expression recognition using longitudinal facial expression atlases. In *ECCV*, 2012.

[10] T. Hasan and J. H. Hansen. A study on universal background model training in speaker verification. *IEEE T ASLP*, 19(7):1890–1899, 2011.

[11] X. Hong, H. Chang, S. Shan, X. Chen, and W. Gao. Sigma set: A small second order statistical region descriptor. In *CVPR*, 2009.

[12] S. Jain, C. Hu, and J. K. Aggarwal. Facial expression recognition with temporal modeling of shapes. In *ICCVW*, 2011.

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

[14] A. Klaser and M. Marszalek. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[15] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, 2011.

[16] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *CVPR*, 2013.

[17] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *FG*, 2011.

[18] M. Liu, S. Li, S. Shan, and X. Chen. Au-aware deep networks for facial expression recognition. In *FG*, 2013.

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

[20] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE T PAMI*, 22(12):1424–1445, 2000.

[21] R. Ptucha, G. Tsagkatakis, and A. Savakis. Manifold based sparse representation for robust expression recognition without neutral subtraction. In *ICCVW*, 2011.

[22] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell. Spatio-temporal covariance descriptors for action and gesture recognition. In *WACV*, 2013.

[23] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACM MM*, 2007.

[24] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *IVC*, 27(6):803–816, 2009.

[25] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE T PAMI*, 30(10):1713–1727, 2008.

[26] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *LRECW*, 2010.

[27] L. Wang, Y. Qiao, and X. Tang. Motionlets: Mid-level 3d parts for human motion recognition. In *CVPR*, 2013.

[28] R. Wang and X. Chen. Manifold discriminant analysis. In *CVPR*, 2009.

[29] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, 2012.

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

[31] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE T PAMI*, 29(1):40–51, 2007.

[32] P. Yang, Q. Liu, X. Cui, and D. N. Metaxas. Facial expression recognition using encoded dynamic features. In *CVPR*, 2008.

[33] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen. Facial expression recognition from near-infrared videos. *IVC*, 29(9):607–619, 2011.

[34] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE T PAMI*, 29(6):915–928, 2007.

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