# A Temporal Latent Topic Model for Facial Expression Recognition

Lifeng Shang and Kwok-Ping Chan

The University of Hong Kong, Pokfulam, Hong Kong
{lfshang,kpchan}@cs.hku.hk

**Abstract.** In this paper we extend the latent Dirichlet allocation (LDA) topic model to model facial expression dynamics. Our topic model integrates the temporal information of image sequences through redefining topic generation probability without involving new latent variables or increasing inference difficulties. A collapsed Gibbs sampler is derived for batch learning with labeled training dataset and an efficient learning method for testing data is also discussed. We describe the resulting temporal latent topic model (TLTM) in detail and show how it can be applied to facial expression recognition. Experiments on CMU expression database illustrate that the proposed TLTM is very efficient in facial expression recognition.

## 1 Introduction

Facial expression recognition has become an active research topic in recent years due to its potential applications in human computer interfaces, data-driven animation, etc. Most facial expression recognition methods attempt to recognize six prototypic expressions (namely joy, surprise, anger, disgust, sadness and fear) proposed by Ekman [6]. Over the past decade, many techniques (e.g. Neural networks [22]) have been applied to still facial images recognition. Psychological studies show that facial image sequences often produce more accurate and robust recognition compared to mug shots [1]. Therefore, recent attention has been moving to model the facial expression dynamics through integrating temporal information [12] [18] [19].

The approaches to modeling temporal behaviors of facial expressions are generally classified as designing dynamic features (e.g. Dynamic Texture [27]) or constructing sequential data modeling tools (e.g. Dynamic Graphical Model [26]). Yang et al. [24] designed a dynamic Haar-like feature to represent facial image sequences. Zhao et al. [27] extended the well-known local binary feature (LBP) to the temporal domain and applied it to facial expression recognition. Yeasin et al. [25] captured the dynamics of facial image sequences by Hidden Markov Models (HMMs). To better model the relative change of emotional magnitude, Zhang et al. [26] presented a probabilistic framework by integrating the Dynamic Bayesian networks (DBNs) with the facial action units (AUs) [6]. Their methods can reflect the evolution of a spontaneous expression. DBNs are natural for modeling facial expression variations, and can be easily extended by combining them

R. Kimmel, R. Klette, and A. Sugimoto (Eds.): ACCV 2010, Part IV, LNCS 6495, pp. 51–63, 2011.

with other models (e.g. Neural Networks) or incorporating semantic relationships between AUs. Nevertheless, modeling the temporal order of facial expression explicitly is risky, because noise in the facial features can easily propagate through the model. Moreover, these models often suffer from too many latent variables or too complex model structures, which makes learning and inference difficult.

Recently, in the statistical text community latent topic models (e.g. LDA [2]) have achieved significant success in semantic clustering. Besides modeling text generation, LDA has also been widely used to solve computer vision problems, e.g. object discovery [23] and scene categorization [15]. However, directly applying a language model to computer vision problems has some difficulties, since in LDA the "bag-of-words" representation relies on the assumption that the order of words or documents can both be ignored. As pointed out by Wang et al. [23], the spatial and temporal structure of documents or words are meaningless in a language model, but important for many computer vision problems. Therefore, studies on extending the LDA to model the spatiotemporal structures of words, topics, documents or corpora have gained more and more attention. Wang et al. [23] proposed a spatial LDA to include the spatiotemporal structure among visual words. Hanna [8] considered word order information by incorporating $n$-gram statistics. Hospedales et al. [9] combined HMM with LDA to model behavior dynamics. In this paper, we propose a new latent topic model (TLTM) which considers the temporal structure of facial image sequences. In TLTM, facial expression dynamics is included by redefining topic generation probability to ensure that successive images are most likely to have the similar topic distributions. Compared to existing extensions, our TLTM does not use new latent variables nor increase inference difficulties, which makes it as efficient as LDA. Experiments on CMU facial expression dataset [11] show that our generative TLTM model outperforms the generally used HMM models and achieves comparable performance as some discriminant models.

The rest of this paper is organized as follows. In Section 2, we describe the feature extraction method. In Section 3, we introduce the proposed TLTM and apply it to facial expression recognition. In Section 4, the performance of proposed method is evaluated by the CMU dataset. Section 5 summarizes this paper.

## 2 Feature Extraction and Indication

In facial expression recognition, there are two types of facial features: permanent and transient features. The permanent facial features are the shapes and locations of facial components (e.g. eyebrows, eye lids, nose, lips and chin). The transient features are the wrinkles and bulges appeared with expressions. In this work, we do not consider transient features and use the movement of facial features away from neutral positions to measure facial expression variation.

We applied the well-known Active Appearance Model (AAM) [5] on facial image sequences to track the movement of facial features. Figure 1(a) shows the shape model consisting of 58 facial points which is identical with the one given in [4]. Figure 2 displays the facial feature localization results of one subject's six

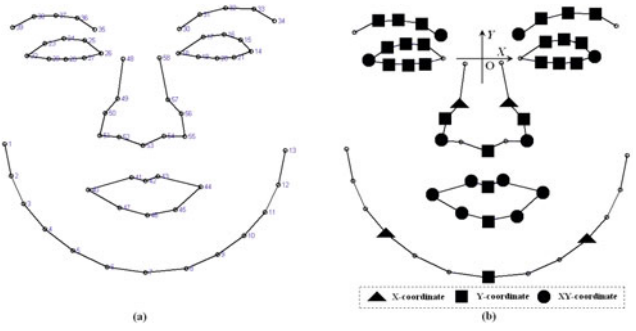


**Fig. 1.** (a) The facial landmarks(58 facial points) and (b) selected feature points

basic expressions. In [4], the $(x, y)$ coordinates of the 58 localized facial points forming a 116-dimensional vector are used to represent an image. Based on the analysis of facial action coding system (FACS) [6], we found that the movements of some facial points (e.g. facial points 1 and 13) are not essential to measuring facial deformation, so a subset is selected from the 58 facial points as feature points which are depicted in Fig. 1(b), in which the solid triangles and rectangles represent that only the X or Y-coordinates are used as feature and the solid circles represent that both the X and Y-coordinates are used. The midpoint of the inner corners of the two eyes (facial points 18 and 26) is defined as the origin. A facial image is thus represented by a 52-dimensional feature vector.


**Fig. 2.** The tracking results of one subject's six basic expressions

To further reduce the inter-personal variations with regard to the amplitudes of facial actions, feature points are quantized into a fixed number of words according to movements away from neutral positions. The movement in the X-axis direction is quantized into a word of the vocabulary $\text{VocabularyX} = \{\text{Left}_i, \text{Right}_i, \text{MotionlessX}_i | i = 1, 2, \cdots, 58\}$, where the word $\text{Left}_i$ ($\text{Right}_i$) represents that the $i$-th facial point moves at lest two pixels left (right) to its neutral position, otherwise it will be quantized to the word $\text{MotionlessX}_i$. Similarly, the vocabulary describing the movement types in the Y-axis direction is defined as $\text{VocabularyY} = \{\text{Up}_i, \text{Down}_i, \text{MotionlessY}_i | i = 1, 2, \cdots, 58\}$. With the two vocabularies at hand, for a given facial image $d_i$, its 52-dimensional feature vector is changed to a bag-of-words representation $\{w_{i,1}, \cdots, w_{i,52}\}$. Our image collection (corpus) is constructed by concatenating these bag-of-words representations one after the other.

## 3 TLTM for Facial Expression Recognition

Facial expressions can be described by the FACS, in which each expression is characterized by the co-occurrence of atomic facial AUs which are represented by some low-level features. LDA is a hierarchical generative topic model, which is very suitable for discovering the co-occurrence of low-level visual words (or higher-level topics). We can find there is a good correspondence between the FACS and LDA model. When LDA is applied to modeling facial expression variations the low-level visual words (i.e. the movements of feature points away from neutral positions) are clustered into higher level topics which correspond to atomic facial action units. In this section we will first briefly review LDA and establish notations, then particularly study how to extend LDA to model facial expression dynamics.

### 3.1 Latent Dirichlet Allocation

LDA is a generative model for topic discovery which has attracted a lot of interest from the field of machine learning, language processing and computer vision community. Figure 3 shows the graphical model of LDA. In this model, documents are represented as random mixtures over latent topics, which are characterized by discrete distributions over words.

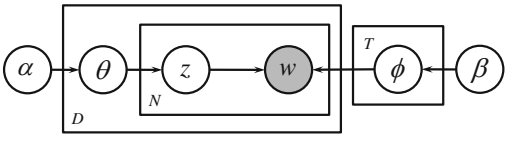


**Fig. 3.** Plate notation for LDA

Each individual word token $w_n$ in a corpus $\boldsymbol{w} = \{w_1, w_2, \cdots, w_N\}$ is assumed to have been generated by a latent topic $z_n$, which is drawn from a document-specific distribution over $T$ topics. The probability of generating a word $w$ from a topic $t$ is defined by $\phi_{w|t} = P(w_n = w|z_n = t)$. These probabilities are recorded by a $T \times W$ matrix $\boldsymbol{\Phi}$, where $W$ is the size of vocabulary and $T$ is the number of topics. Similarly, the topic generation is characterized by another conditional probability $\theta_{t|d} = P(z_n = t|d_n = d)$. These probabilities are recorded by a $D \times T$ matrix $\boldsymbol{\Theta}$, where $D$ is the number of documents in the corpus. Thus the joint probability of the corpus $\boldsymbol{w}$ and a set of corresponding latent topics $\boldsymbol{z} = \{z_1, \cdots, z_N\}$ is

$$P(\boldsymbol{w}, \boldsymbol{z}|\Phi, \Theta) = \prod_{n=1}^{N} \phi_{w_n|z_n} \theta_{z_n|d_n} \tag{1}$$

where $w_n$ is the $n$-th word of the corpus $\boldsymbol{w}$, $z_n$ is the topic assignment for the $n$-th word and $d_n$ is the document number of the $n$-th word.

To make the model fully Bayesian, symmetric Dirichlet priors with hyper parameters $\alpha$ and $\beta$ are placed over $\Theta$ and $\Phi$

$$P(\Theta|\alpha) = \prod_d \text{Dirichlet}(\theta_d|\alpha) \text{ and } P(\Phi|\beta) = \prod_t \text{Dirichlet}(\phi_t|\beta) \tag{2}$$

where $\theta_d$ is the $d$-th row of the matrix $\Theta$, $\phi_t$ is the $t$-th row of the matrix $\Phi$. Combining the two priors with equation (1) and integrating over $\Theta$ and $\Phi$ gives the joint probability of corpus and latent topics given hyperparameters: $P(\boldsymbol{w}, \boldsymbol{z}|\alpha, \beta)$. Consequently the posterior probability for latent topics $\boldsymbol{z}$ is calculated

$$P(\boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) = \frac{P(\boldsymbol{w}, \boldsymbol{z}|\alpha, \beta)}{\sum_{\boldsymbol{z}} P(\boldsymbol{w}, \boldsymbol{z}|\alpha, \beta)}. \tag{3}$$

Unfortunately, exact inference is intractable for LDA, since computing the equation (3) involves evaluating a probability distribution on a large discrete state space. However, there have been three approximating methods to learn LDA, EM with variation inference [2], EM with expectation propagation [16], and Gibbs sampling [7]. In this work, we adopted Gibbs sampling, since this method is better tolerant to local optima and its performance is comparable with the other two methods.

To sample from the posterior distribution (3) using the Gibbs sampling method, we need the full conditional distribution

$$P(z_n = t|\boldsymbol{z}_{-n}, \boldsymbol{w}, \alpha, \beta) \propto \frac{N_{-n,t}^{(w_n)} + \beta}{N_{-n,t}^{(\cdot)} + W\beta} \frac{N_{-n,t}^{(d_n)} + \alpha}{N_{-n}^{(d_n)} + T\alpha} \tag{4}$$

where $\boldsymbol{z}_{-n}$ denotes all the $z_j$ with $j \neq n$, $N_{-n,t}^{(w_n)}$ is the number of times the word $w_n$ assigned to topic $t$ and $N_{-n,t}^{(\cdot)}$ is the number of words assigned to topic $t$. $N_{-n,t}^{(d_n)}$ is the number of times topic $t$ occurring in document $d_n$ and $N_{-n}^{(d_n)}$ is the number of words in document $d_n$. All the four numbers do not include the current assignment of $z_n$. With a set of samples the parameters $\Theta$ and $\Phi$ can be estimated from $\boldsymbol{w}$ and $\boldsymbol{z}$ by

$$\hat{\theta}_{t|d} = \frac{N_t^{(d)} + \alpha}{N^{(d)} + T\alpha}, \text{ and } \hat{\phi}_{w|t} = \frac{N_t^{(w)} + \beta}{N_t^{(\cdot)} + W\beta}. \tag{5}$$

In the context of facial expression recognition, low-level visual words are clustered into higher level topics by LDA which correspond to atomic facial action units. In the next section, our TLTM model will be built based on LDA by including temporal information of image sequences.

### 3.2 Temporal Latent Topic Model

Before using LDA to model facial expressions dynamics, we need to first define the meaning of "document" for facial expression recognition. If we treat each facial image sequence as a document, the document order information will be

changed to word order information. However, in the standard LDA words are exchangeable, so document structure information will be ignored. To include word order information, Hanna [8] incorporated $n$-gram statistics. If we define each image as a document, LDA still misses document order information, since LDA is developed for unstructured documents. In [9], Hospedales et al. introduced a Markov chain to model the temporal structure of image sequences. In TLTM, we adopt the latter way that the bag-of-words representation of one facial image is defined as a document. In order to include the temporal information of facial image sequences, we modify the topic generation probability $\theta_{t|d}$ to be $\theta_{t|d,\mathrm{pre}(d)}$, where $\mathrm{pre}(d)$ is the index of the previous image of the $d$-th image. Since our image collection is constructed by stacking image sequences one after the other and preserving the inner sequence structure, the value of $\mathrm{pre}(d)$ will be $(d-1)$ or null if image $d$ is the first slice of a sequence. In the case of null, the topic generation probability will be reduced to $\theta_{t|d}$.

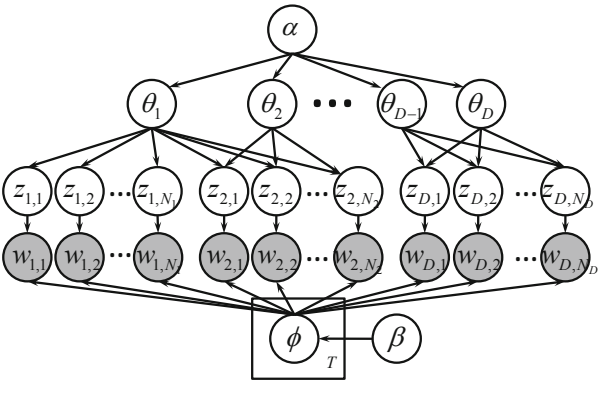


**Fig. 4.** The graphical model of TLTM

Figure 4 shows the graphical model of TLTM, in which the repeated topic and word generation within the corpus is explicitly drawn. In this figure, $N_d$ is the number of words in the $d$-th document (image) and has the value of 52 in this work. $w_{i,j}$ is the $j$-th word of the $i$-th document and its topic assignment is $z_{i,j}$. The index of the word $w_{i,j}$ in the whole corpus is calculated as $n = ((i-1) \times 52 + j)$. Compared to the standard LDA as shown in Fig. 3, it can be observed that the topic generation probability $\theta_{t|d}$ does not only depend on $\theta_d$ but also depends on $\theta_{(d-1)}$. Use the topic assignment of the word $w_{2,1}$ as an example, the generation probability for $z_{2,1}$ depends both on $\theta_1$ and $\theta_2$. The generation probability of $z_{2,1}$ is changed from $P(z_{2,1}|\theta_2)$ to $P(z_{2,1}|\theta_1, \theta_2)$. The joint probability $P(\boldsymbol{w}, \boldsymbol{z}|\Phi, \Theta)$ becomes to

$$P(\boldsymbol{w}, \boldsymbol{z}|\Phi, \Theta) = \prod_{n=1}^{N} \phi_{w_n|z_n} \theta_{z_n|d_n,(d_n-1)}. \tag{6}$$

According to the Bayes'rule, the distribution $\theta_{t|d,(d-1)}$ can be calculated from the distributions $\theta_{t|d}$ and $\theta_{t|(d-1)}$ as follows

$$\begin{aligned} P(t|\theta_d, \theta_{(d-1)}) &= \frac{P(t)P(\theta_d, \theta_{(d-1)}|t)}{P(\theta_d, \theta_{(d-1)})} = \frac{P(t)P(\theta_d|t)P(\theta_{(d-1)}|t)}{P(\theta_d, \theta_{(d-1)})} \\ &= \frac{P(\theta_d)P(\theta_{(d-1)})}{P(\theta_d, \theta_{(d-1)})} \frac{P(t|\theta_d)P(t|\theta_{(d-1)})}{P(t)} \\ &\propto \frac{P(t|\theta_d)P(t|\theta_{(d-1)})}{P(t)}, \end{aligned} \tag{7}$$

where $P(t)$ is the prior probability of topic $t$ and the prior knowledge here is defined as the set of words, which have the same sequence number as $w_n$ does, and their corresponding topic assignments. So the prior probability $P(t)$ can be regarded as a sequence level topic generation probability with respect to document level topic generation probability (i.e. $\theta_{t|d}$), and it is characterized by a conditional probability $\psi_{t|s} = P(z_n = t|s_n = s)$. These probabilities are recorded by a $S \times T$ matrix $\Psi$, where $S$ is the number of sequences in the image collection.

As in LDA, we place symmetric Dirichlet priors with hyper parameters $\alpha$, $\beta$ and $\gamma$ over $\Theta$, $\Phi$ and $\Psi$, respectively. $P(\Theta|\alpha)$ and $P(\Phi|\beta)$ are given as in equation (2). $P(\Psi|\gamma)$ is given as follows

$$P(\Psi|\gamma) = \prod_s \text{Dirichlet}(\psi_s|\gamma) \tag{8}$$

where $\psi_s$ is the $s$-th row of the matrix $\Psi$. Combining the three priors with equation (6) and integrating over $\Theta$, $\Phi$ and $\Psi$ gives the joint probability of corpus and latent topics given hyperparameters:

$$P(\boldsymbol{w}, \boldsymbol{z}|\alpha, \beta, \gamma) = \prod_{t=1}^{T} \frac{\mathrm{B}(C_t^T + \boldsymbol{\beta})}{\mathrm{B}(\boldsymbol{\beta})} \prod_{d=1}^{D} \frac{\mathrm{B}(C_d^D + C_{d+1}^D + \boldsymbol{\alpha})}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{s=1}^{S} \frac{\mathrm{B}(\boldsymbol{\gamma})}{\mathrm{B}(C_s^S + \boldsymbol{\gamma})}, \tag{9}$$

where $\mathrm{B}(\cdot)$ is the multinomial beta function, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors with const elements $\alpha$, $\beta$ and $\gamma$, respectively. $C^T$, $C^D$ and $C^S$ are three count matrixes. $C_t^T$, $C_d^D$ and $C_s^S$ are the $t$-, $d$- and $s$-th row of the matrixes $C^T$, $C^D$ and $C^S$, respectively. The $(t, w)$-th element of $C^T$ is the number of times that topic $t$ is assigned to word $w$. The $(d, t)$-th element of $C^D$ is the number of times that topic $t$ is assigned to words in document $d$. The $(s, t)$-th element of $C^S$ is the number of times that topic $t$ is assigned to words in sequence $s$. Finally, the Gibbs sampling update for the topic $z_n$ is obtained as follows

$$\begin{aligned} P(z_n = t|\boldsymbol{z}_{-n}, \boldsymbol{w}, \alpha, \beta, \gamma) &= \frac{P(z_n = t, \boldsymbol{w}, \boldsymbol{z}_{-n}|\alpha, \beta, \gamma)}{P(\boldsymbol{w}, \boldsymbol{z}_{-n}|\alpha, \beta, \gamma)} \\ \propto \frac{N_{-n,t}^{(w_n)} + \beta}{N_{-n,t}^{(\cdot)} + W\beta} & \frac{N_t^{(d_n-1)} + N_{-n,t}^{(d_n)} + \alpha}{N^{(d_n-1)} + N_{-n}^{(d_n)} + T\alpha} \frac{N_{-n,t}^{d_n} + N_t^{(d_n+1)} + \alpha}{N_{-n}^{d_n} + N^{(d_n+1)} + T\alpha} \frac{N_{-n}^{(s_n)} + T\gamma}{N_{-n,t}^{(s_n)} + \gamma} \end{aligned} \tag{10}$$

where $s_n$ is the sequence number of the word $w_n$, $N_{-n,t}^{(s_n)}$ is the number of times topic $t$ occurring in the the facial image sequence $s_n$ and $N_{-n}^{(s_n)}$ is the number of words in the sequence $s_n$(both excluding $z_n$). With a set of samples from the posterior distribution $P(\boldsymbol{z}|\boldsymbol{w})$, we can estimate $\Theta$, $\Phi$, and $\Psi$ from $\boldsymbol{w}$ and $\boldsymbol{z}$ by equations

$$\hat{\theta}_{t|d} = \frac{N_t^{(d)} + N_t^{(d+1)} + \alpha}{N^{(d)} + N^{(d+1)} + T\alpha}, \hat{\phi}_{w|t} = \frac{N_t^{(w)} + \beta}{N_t^{(\cdot)} + W\beta}, \text{ and } \hat{\psi}_{t|s} = \frac{N_t^{(s)} + \gamma}{N^{(s)} + T\gamma}. \quad (11)$$

### 3.3 Applying TLTM to Facial Expression Recognition

In facial expression recognition, TLTMs are learned for facial expression training dataset. The learned TLTM for the $i$-th facial expression is denoted by a compact notation $\text{TLTM}^{[\text{Tr}^i]} = (\boldsymbol{w}^{[\text{Tr}^i]}, \boldsymbol{z}^{[\text{Tr}^i]}, \Theta^{[\text{Tr}^i]}, \Phi^{[\text{Tr}^i]}, \Psi^{[\text{Tr}^i]})$, here $\boldsymbol{w}^{[\text{Tr}^i]}$ is the image corpus of the $i$-th facial expression and $\boldsymbol{z}^{[\text{Tr}^i]}$ is the learned latent topic assignments. For a new facial image not contained in training dataset, we need to quickly assess the topic assignments, while the standard inference method described above is offline. Recently, some online [3] or efficient inference methods have been proposed [20], we adopt the efficient Monte Carlo algorithm as described in [20]. The basic idea of this method is to run only on the word tokens in the new image.

Given a testing facial image sequence $\{d_1^{[\text{Te}]}, d_2^{[\text{Te}]}, \cdots, d_j^{[\text{Te}]}, \cdots\}$, where the $j$-th image $d_j^{[\text{Te}]}$ has the bag-of-words representation $\{w_{j,1}^{[\text{Te}]}, w_{j,2}^{[\text{Te}]}, \cdots, w_{j,52}^{[\text{Te}]}\}$. The trained TLTMs are used to classify the current image slice into one of the six basic expressions. Let $l_j^{[\text{Te}]}$ denote the label of the $j$-th testing image. Once the $j$-th image is obtained, we will sample new assignments of words to topics by applying equation (10) only to the word tokens in the $j$-th image. After several sampling iterations (20 iterations in our simulation), we can get the topic assignment $z_{j,k}^{[\text{Te}]}$ for each word in $d_j^{[\text{Te}]}$. The topic generation probabilities for image $d_j^{[\text{Te}]}$ in both document and sequence levels can be estimated by equation (11), and the probability $\theta_{t|d_j^{[\text{Te}]},d_j^{[\text{Te}]}-1}^{[\text{Te}]}$ can thus be calculated by equation (7). Finally, the observation probability of $d_j^{[\text{Te}]}$ conditioned on the $i$-th expression is calculated by

$$P(d_j^{[\text{Te}]}|l_j^{[\text{Te}]} = i) = \prod_{k=1}^{52} \sum_{t=1}^{T} \phi_{w_{j,k}^{[\text{Te}]}|t}^{[\text{Tr}^i]} \theta_{t|d_j^{[\text{Te}]},d_j^{[\text{Te}]}-1}^{[\text{Te}]}. \quad (12)$$

According to the Bayes's rule, the probability of sequence $\{d_1^{[\text{Te}]}, d_2^{[\text{Te}]}, \cdots, d_j^{[\text{Te}]}\}$ classified to the $i$-th expression is calculated as

$$P(l_j^{[\text{Te}]} = i|\{d_1^{[\text{Te}]}, d_2^{[\text{Te}]}, \cdots, d_j^{[\text{Te}]}\}) \propto \frac{1}{N(j)} P(d_j^{[\text{Te}]}|l_j^{[\text{Te}]} = i)$$

$$\sum_{k=1}^{6} P(l_j^{[\text{Te}]} = i|l_{(j-1)}^{[\text{Te}]} = k) P(l_{(j-1)}^{[\text{Te}]} = k|\{d_1^{[\text{Te}]}, d_2^{[\text{Te}]}, \cdots, d_{(j-1)}^{[\text{Te}]}\}), \quad (13)$$

here $N(j)$ is a scale factor to ensure $\sum_{i=1}^{6} P(l_j^{[\mathrm{Te}]} = i|\{d_1^{[\mathrm{Te}]}, d_2^{[\mathrm{Te}]}, \cdots, d_j^{[\mathrm{Te}]}\}) = 1$ and $P(l_j^{[\mathrm{Te}]} = i|l_{(j-1)}^{[\mathrm{Te}]} = k)$ is the transition probability from expression $k$ to $i$. To facilitate the computation of transition probabilities, a $6 \times 6$ matrix $R$ is constructed. The $(k, i)$-th entry of $R$ records the number of times transmitting from the expression $k$ to $i$ in two consecutive time slices. $R$ is initialized to a matrix with all ones. The transition probability $P(l_j^{[\mathrm{Te}]} = i|l_{(j-1)}^{[\mathrm{Te}]} = k)$ is simply calculated as $R_i^k / \sum_i R_i^k$, where $R_i^k$ is the $(k, i)$-th entry of the matrix $S$. All the probabilities involved in (13) are obtained, a testing facial image sequence is classified to expression $i^*$

$$i^* = \arg \max_{i=1,\ldots,6} P(l_j^{[\mathrm{Te}]} = i|\{d_1^{[\mathrm{Te}]}, d_2^{[\mathrm{Te}]}, \cdots, d_j^{[\mathrm{Te}]}\}), \tag{14}$$

if $P(l_j^{[\mathrm{Te}]} = i^*|\{d_1^{[\mathrm{Te}]}, d_2^{[\mathrm{Te}]}, \cdots, d_j^{[\mathrm{Te}]}\}) > 0.40$, otherwise Neutral expression is assigned.

## 4 Experiments

### 4.1 Dataset and Parameter Settings

We use the Cohn-Kanade Database to evaluate the performance of TLTM. This database consists of 100 university students ranging in age from 18 to 30 years. Sixty-five percent were female, fifteen percent were African-American and three percent Asian or Latino. For our experiments, we selected 72 whole image sequences (totally, 1085 images) from the database. Each expression contains 12 sequences. The original frames are normalized to $170 \times 210$ pixels facial images based on the positions of two eyes. Before using TLTM and LDA, we need to first set the hyperparameters $\alpha$, $\beta$, $\gamma$ and the number of latent topics $T$. For all runs of our algorithm, we set $\alpha$, $\beta$ and $\gamma$ to constant values $\alpha = 50/T$, $\beta = 0.1$ and $\gamma = 60/T$. $T$ is a very influential parameter for any latent topic models, and some Dirichlet Processes based methods have been proposed to estimate the value of $T$ automatically [21]. In our simulation, we used the generally acceptable empirical methods to determine the optimal value for $T$. We run our model for different $T$ values and found five latent topics provides the best recognition rate.

**Table 1.** The recognition rates (%) of LDA and TLTM

| | JOY | SUR | ANG | DIS | SAD | FEA | Overall Rate |
|---|---|---|---|---|---|---|---|
| LDA | 66.67 | 100.00 | 83.33 | 94.44 | 100.00 | 88.89 | 88.89 |
| TLTM | 75.00 | 100.00 | 91.67 | 100.00 | 100.00 | 95.83 | 93.75 |

### 4.2 Experimental Results

We used a three-fold cross validation in our experiments to verify the benefits of using TLTM to model facial expression dynamics. Table 1 presents the recognition results of LDA and TLTM. It can be observed that the TLTM method

outperforms the LDA method for the recognition of joy, anger, disgust and fear expressions, which confirms the benefit of using temporal information of image sequences. Furthermore, we can see that both methods perform relatively worse for the joy expression, since the joy expression mainly includes two AUs: AU6(Cheek raiser) + AU12(Lip corner puller) and the AU6 depends on some transient features such as nasolabial furrows presence and eye wrinkles increased, however AAM is not particularly suitable to track these features. Other tools (e.g. Canny edge operator) will be used to quantify the intensity of furrows and wrinkles in future work to obtain better performance.

**Table 2.** Comparisons with other methods

| Methods | ParzenHMM | KnnHMM | DynamicLBP | SVMLBP | LDA | TLTM |
|---|---|---|---|---|---|---|
| Overall Rate | 86.11 | 91.67 | 96.26 | 92.10 | 88.89 | 93.75 |

Table 2 summarizes a comparison to some other representative approaches. Here "ParzenWHMM" denotes a modified HMM in which the generation probability is estimated by a nonparametric density estimation method-Parzen Windows [10]. "KnnHMM" denotes a discriminate HMM proposed by Lefevre [14], in which the discrimination ability at hidden state level is improved by a k-nearest neighbors (k-NN) estimation method. "SVMLBP" denotes the method proposed by Shan [17], they used LBP to represent facial features and SVM as classifier. The "DynamicLBP" method used dynamic LBP to represent facial features and used SVM as classifier [27]. It can be observed that the LDA method performs better than the HMM based method and slightly worse than its discriminant version KnnHMM. Our method achieves the similar performance as the SVMLBP method, although TLTM is a generative model and does not use the information of other classes in the training stage. The main difference of the methods DynamicLBP and SVMLBP is LBP is replaced by dynamic LBP, which confirms the benefit of considering the temporal information for sequential data classification. The DynamicLBP method performs better than our generative model, since SVM is a discriminant model which uses the information of other classes in the training stage. Recently, some works on increasing the discriminant ability of LDA have been proposed such as DisLDA [13] and MedLDA [28]. We will use some discriminant rules to train our TLTM in future work to get higher recognition rate.

### 4.3 Some Recognition Examples

In this section, we will use two examples to illustrate the efficiency of the proposed method in an intuitive way. In the first example, we created a short image sequence as shown in Figure 5(a) in which the subject performed smiling with blinking her eyes in the frames 4 and 5. We can observe that from the second frame lip corners begin to be pulled obliquely and cheeks are raised. From Fig. 5(b), we can see the probabilities of the six expressions are close in the first

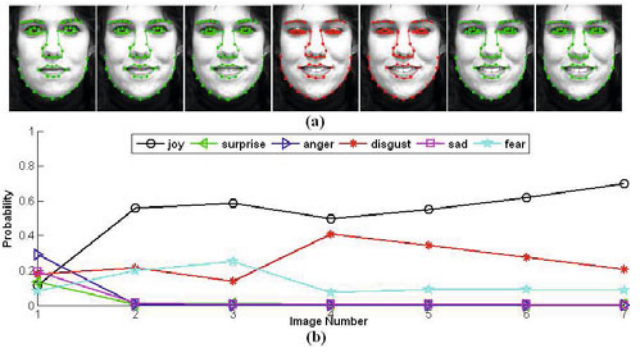


**Fig. 5.** Example 1: (a) An image sequence shows a subject performing smiling with blinking eyes in the frames 4 and 5, (b) the probability distributions of facial expressions

frame. As the expression progresses with time the probability of joy increases gradually and decreases in the frames 4 and 5 resulted by the eyes blinking action. In the 7-th frame, the probability of joy rises to nearly 0.7 and implies that this frame has the apex joy expression. This experiment illustrates that our method can well model the evolution of facial expression.

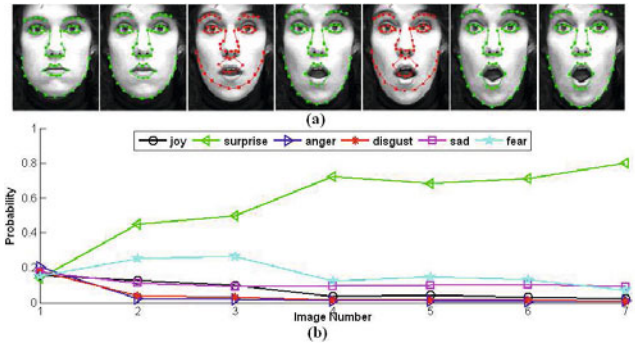


**Fig. 6.** Example 2: (a) An image sequence shows a subject performing surprise with tracking error in the frames 3 and 5, (b) the probability distributions of facial expressions

Figure 6(a) shows another image sequence in which the subject performed surprise with some frames mis-tracked. In frames 3 and 5, we can see that the locations of mouth and chin are tracked in error. Fig. 6(b) gives the result of our method, from which we can observe that although the probability of surprise visibly decreases in the 5-th frame, the facial expression can still be correctly recognized. This example illustrates that our method is robust to tracking error.

## 5 Conclusions and Future Work

This paper proposed a new latent topic model TLTM for facial expression analysis by integrating the temporal information of image sequences. We redefined the topic generation probability without involving new latent variables or increasing inference difficulties. Experiments on CMU expression database confirmed the efficiency of the TLTM in facial expression recognition. In future work, we will pay more attention to feature extraction and use some discriminant training rules to increase the discriminant ability of TLTM to get better performance.

## References

1. Bassili, J.: Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. Personality and Social Psychology, 2049–2059 (1979)
2. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. JMLR 3(2-3), 993–1022 (2003)
3. Canini, K.R., Shi, L., Griffiths, T.L.: Online inference of topics with latent Dirichlet allocation. In: AISTATS (2009)
4. Chang, Y., Hu, C., Turk, M.: Probabilistic expression analysis on manifolds. In: CVPR, pp. 520–527 (2004)
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. on PAMI 23(6), 681–685 (2001)
6. Ekman, P., Friesen, W.V.: Facial Action Coding System (FACS): Manual. Consulting Psychologists Press, Palo Alto (1978)
7. Griffiths, T.L., Steyvers, M.: Finding scientific topics. PNAS 101, 5228–5235 (2004)
8. Hanna, M.W.: Topic modeling: beyond bag-of-words. In: ICML (2006)
9. Hospedales, T., Gong, S., Xiang, T.: A Markov clustering topic model for mining behaviour in video. In: ICCV (2009)
10. Jin, N., Mokhtarian, F.: A non-parametric HMM learning method for shape dynamics with application to human motion recognition. In: ICPR, pp. 29–32 (2006)
11. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: FG, pp. 46–53 (2000)
12. Kumano, S., Otsuka, K., Yamato, J., Maeda, E., Sato, Y.: Pose-invariant facial expression recognition using variable-intensity templates. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part I. LNCS, vol. 4843, pp. 324–334. Springer, Heidelberg (2007)
13. Lacoste-Julien, S., Sha, F., Jordan, M.I.: DiscLDA: Discriminative learning for dimensionality reduction and classification. In: NIPS, pp. 897–904 (2008)
14. Lefevre, F.: Nonparametric probability estimation for HMM-based automatic speech recognition. Computer Speech and Language 17(2-3), 113–136 (2003)
15. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR, pp. 524–531 (2005)
16. Minka, T., Lafferty, J.: Expectation propagation for the generative aspect model. In: UAI, pp. 352–359 (2002)
17. Shan, C., Gong, S., McOwan, P.W.: Robust facial expression recognition using local binary patterns. In: ICIP, pp. 370–373 (2005)
18. Shang, L., Chan, K.P.: Temporal Exemplar-Based Bayesian Networks for Facial Expression Recognition. In: ICMLA, pp. 16–22 (2008)

19. Shang, L., Chan, K.P.: Nonparametric Discriminant HMM and Application to Facial Expression Recognition. In: CVPR, pp. 2090–2096 (2009)
20. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic Author-Topic Models for Information Discovery. In: KDD, pp. 306–315 (2004)
21. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.: Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. In: NIPS (2004)
22. Tian, Y., Kanade, T., Cohn, J.: Recognizing action units for facial expression analysis. IEEE Trans. on PAMI 23(2), 97–115 (2001)
23. Wang, X., Grimson, E.: Spatial latent dirichlet allocation. In: NIPS (2007)
24. Yang, P., Liu, Q., Metaxas, D.N.: Boosting coded dynamic features for facial action units and facial expression recognition. In: CVPR, pp. 1–6 (2007)
25. Yeasin, M., Bullot, B., Sharma, R.: From facial expression to level of interest: a spatio-temporal approach. In: CVPR, pp. 922–927 (2004)
26. Zhang, Y., Ji, Q.: Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Trans. on PAMI 27(5), 699–714 (2005)
27. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. on PAMI 29(6), 915–928 (2007)
28. Zhu, J., Ahmed, A., Xing, E.P.: MedLDA: Maximum margin supervised topic models for regression and classification. In: ICML (2009)