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.
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) 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 and scene categorization. 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., 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. proposed a spatial LDA to include the spatiotemporal structures among visual words. Hanna considered word order information by incorporating n-gram statistics. Hospedales et al. 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 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) on facial image sequences to track the movement of facial features. Figure(a) shows the shape model consisting of 58 facial points which is identical with the one given in [4]. Figure displays the facial feature localization results of one subject’s six