Statistical Motion Characterization for Video Content Classification

Chiou-Ting Hsu and Ching-Wei Lee
Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
{cthsu,br871528}@cs.nthu.edu.tw

Abstract

This paper proposed using a unified model to characterize the motion variations along both the spatial and temporal domains. To this end, we estimate the motion quantities from the pixelwise normal flow and represent the motion distribution using two Gibbs models: temporal and spatial Gibbs models. We measure the potential values of the two Gibbs models by maximum likelihood criterion. To demonstrate the effectiveness of the proposed model, we have applied the motion model for the application of video content classification. Experimental results show that using the proposed model indeed improves the classification performance.

1. Introduction

Motion is one of the most important features to enable efficient video browsing and classification for voluminous video databases. A number of researches [1-5] have investigated various representations to characterize the motion behaviors. In [6], a nonparametric motion model is proposed to represent the video motion as the “temporal texture” and employ a causal Gibbs model to characterize the motion variation along the temporal domain. In comparison with [1-5], this nonparametric model [6] can better characterize complex motion behaviors, such as motion of river, flames, crowds and etc, which are very difficult to be represented by trajectories or parametric motion models.

Nevertheless, the motion model proposed in [6] is insufficient to differentiate videos with similar motion variations along the temporal domain but with very dissimilar behaviors along the spatial domain. For example, two video clips, one contains a sports game with many small moving players and the other contains only a walking person, may have similar motion distributions along the time domain [7]. However, spatial variations of their motion distributions are very different and completely neglected by the temporal Gibbs model [6].

Therefore, in this paper, we proposed modeling the spatial and temporal variations of motion behaviors in a unified model. The rest of this paper is organized as follows: In Section 2, we review the previous work in [6]. Section 3 introduces our proposed framework and the potential estimation. In Section 4, we present the issues of using the proposed model to the application of video classification. Experimental results are presented in Sec. 5 and the conclusion is given in Sec. 6.

2. Review of Motion Characterization by Temporal Gibbs Model

Here, we review the nonparametric motion model proposed in [6], in which the motion content within a video clip is modeled using a causal Gibbs distribution. Let \( x_k \) denote the quantized motion-related quantities in a video clip containing \( K + 1 \) frames and \( M_k \) denote the underlying Gibbs model. Thus, the motion content along the video clips are assumed to be a realization of a first-order Markov chain

\[
P_{M_k}(x) = P_{M_k}(x_0) \prod_{k=0}^{K} P_{M_k}(x_k | x_{k-1}).
\]

In [6], the authors assume that the random variables in the \( k \)th frame are conditionally independent when given the variables in the \((k-1)\)th frame and that the variable for each pixel \( p \) in the \( k \)th frame depends only on the temporal neighborhoods of the pixel \( p \) in the \((k-1)\)th frame. Hence, the conditional probability is factorized as

\[
P_{M_k}(x_k | x_{k-1}) = \prod_{p \in R} P_{M_k}(x_k(p) | x_{k-1}(p))
\]

where \( R \) is the image grid and \( x_{k-1}(p) \) is the temporal neighborhoods of \( p \) and is illustrated in Fig. 1(a).
Using the Markov-Gibbs equivalence properties [9], the conditional probability function can be written as

\[ P_{M_k}(x, \{x_p\} | \{x_{\eta_k}\}) = \frac{\exp\left[\sum_{k=1}^K \Psi_{M_k}^r (x, \{x_p\}, \{x_{\eta_k}\})\right]}{Z_{M_k}(x, \{x_p\})} \tag{3} \]

where \( \Psi_{M_k}^r = \left\{ \Psi_{M_k}^r (\nu, \nu') \right\}_{\nu, \nu' \in \Theta} \) is the potential for the temporal clique \( a \in A, \Lambda \) is the range of the quantized motion quantities. In (3), the pixel \( p_a \) is the temporal neighborhood of the pixel \( p \) for the clique \( a \) in \( \Phi \), and \( Z_{M_k}(k, p, \{x_{\eta_k}\}) \) is the normalization constant.

Next, the temporal co-occurrence matrix is used to represent the set of possible motion quantities along the temporal domain. Let \( \Gamma^T_x(x) = \left[ \Gamma^T_x(x, x') \right]_{x, x' \in \Theta} \) be the co-occurrence matrix for the clique \( a \). The inner product between the potential \( \Psi_{M_k}^r \) and the co-occurrence matrix \( \Gamma^t_x(x) \) is defined as

\[ \Psi_{M_k}^r(\nu, \nu') \cdot \Gamma^t_x(x) = \sum_{(\nu, \nu') \in \Theta} \Psi_{M_k}^r(\nu, \nu') \Gamma^t_x(\nu, \nu') \tag{4} \]

where \( \Gamma^t_x(\nu, \nu') \) is defined by

\[ \Gamma^t_x(\nu, \nu') = \sum_{t=1}^T \delta(\nu - x_t(p)) \delta(\nu' - x_{t-1}(p)) \tag{5} \]

Therefore, the joint distribution \( P_{M_k}(x) \) is rewritten as

\[ P_{M_k}(x) = \frac{\exp\left[\sum_{k=1}^K \Psi_{M_k}^r(\nu, \nu') \cdot \Gamma^t_x(x)\right]}{Z_{M_k}(x)} \tag{6} \]

3. Proposed Motion Model

3.1. Motion distribution

In addition to the temporal variation, we aim to include the spatial variation of motion quantities into a unified motion distribution model. We assume that the motion variation along the temporal domain is irrelevant to that of the spatial domain and propose using the Gibbs model to represent each distribution individually. Thus, the proposed motion distribution for the sequence of motion quantities \( x = \{x_k\}_{k=0, \ldots, K} \) is represented by

\[ P_{M}(x) = P_{M_t}(x)P_{M_s}(x) \tag{7} \]

In (7), the temporal domain distribution \( P_{M_t}(x) \) is similarly modeled by (1) and (6). For the spatial domain distribution \( P_{M_s}(x) \), our goal is to discriminate between two motion distributions having similar temporal variations but with dissimilar spatial characteristics. For example, we wish to characterize the difference between one large object with uniformly moving pixels from a fragmental set of individually moving pixels. To this end, we consider the spatial 4-neighborhood system (as shown in Fig. 1(b)) and let the Gibbs model in the spatial domain be isotropic. Thus, the potential values of the spatial Gibbs model for different clique types will become the same [9].

We assume that the spatial domain distribution of motion quantities is the joint probabilities of all frames and the probability for each pixel \( p \) in the \( k \)th frame depends only on its spatial neighborhoods \( \eta_p \) in the same frame. Thus, we factorize the model as

\[ P_{M_t}(x) = \prod_{k=0}^K P_{M_t}(x_k), \quad \text{and} \quad P_{M_s}(x) = \prod_{p \in \Theta} P_{M_s}(x_p | x_t(\eta_p)) \tag{8} \]

To further simplify the model into a more tractable form, we also assume that the spatial neighborhoods of pixel \( p \) are independent. Therefore we have

\[ P_{M_s}(x_p | x_t(\eta_p)) = \prod_{p \in \Theta} P_{M_s}(x_p | x_t(\eta_p)) \tag{9} \]

Similarly, we employ the Markov-Gibbs equivalence properties to derive the following equation:

\[ P_{M_s}(x_p | x_t(\eta_p)) = \frac{\exp[\Psi_{M}(x_p(x_t(\eta_p)))]}{Z_{M_s}(k, p, x_t(\eta_p))} \tag{10} \]

where \( \Psi_{M_s} = \left\{ \Psi_{M_s}(\nu, \nu') \right\}_{\nu, \nu' \in \Theta} \) is the potential for the spatial model \( M_s \), which is defined by \( \theta^s \) potential values. The normalization term \( Z_{M_s}(k, p, x_t(\eta_p)) \) is given by

\[ Z_{M_s}(k, p, x_t(\eta_p)) = \exp[\Psi_{M_s}(\nu, \nu')] \tag{12} \]

Therefore, the joint probability for the spatial domain becomes

\[ P_{M_s}(x) = \prod_{k=0}^K \prod_{p \in \Theta} \frac{\exp[\Psi_{M_s}(x_p(x_t(\eta_p)))]}{Z_{M_s}(k, p, x_t(\eta_p))} \tag{13} \]

Next, we represent the spatial variation of the motion quantities using the spatial co-occurrence matrix. The model \( P_{M_s}(x) \) is rewritten as

\[ P_{M_s}(x) = P_{M_s}(x) \frac{\exp[\Psi_{M_s}(x_p(x_t(\eta_p)))]}{Z_{M_s}(x)} \tag{14} \]

where \( \Gamma_s(x) = \left[ \Gamma_s(x, x') \right]_{x, x' \in \Theta} \) is the spatial co-occurrence matrix and \( \Gamma_s(x, x') \) is defined by

\[ \Gamma_s(x, x') = \sum_{k=1}^K \sum_{p \in \Theta} \delta(x_t(p) - x_k(z_k)) \delta(x_{t-1}(p) - x_{k-1}(z_{k-1})) K(z, p - l, p') \tag{15} \]

where \( K(z) \) is the kernel function defined by.
\[
K() = \exp(-\|\|). \tag{16}
\]

3.2. Estimation of potentials

Our proposed motion model in (7) consists of temporal and spatial Gibbs models, which are defined by their corresponding potentials. We follow the derivation in [6] and use the maximum likelihood (ML) criterion to estimate these potential values. We seek the model that maximizes the conditional log-likelihood of the observed motion quantities:

\[
\hat{M} = \arg \max_M \ln(P_M(x)). \tag{17}
\]

Since our motion model \( \hat{M} \) is composed of two independent models \( M_1 \) and \( M_2 \), the ML estimation can be rewritten as

\[
\hat{M}_1 = \arg \max_{M_1} \ln(P_{M_1}(x)) \quad \text{and} \quad \hat{M}_2 = \arg \max_{M_2} \ln(P_{M_2}(x)). \tag{18}
\]

We assume the a priori models \( P_{M_1}(x) \) and \( P_{M_2}(x) \) are uniform and rewrite (18) into

\[
\hat{M}_1 = \arg \max_{M_1} \sum_{x \in \Lambda} \Psi_{M_1}(x, \nu | x) - \ln Z_{M_1}(x) \quad \text{and} \quad \hat{M}_2 = \arg \max_{M_2} \sum_{x \in \Lambda} \Psi_{M_2}(x, \nu | x) - \ln Z_{M_2}(x). \tag{19}
\]

Setting the first-order derivatives of these log-likelihood functions to zero with respect to their corresponding potentials \( \Psi_{M_1}(x, \nu | x) \) and \( \Psi_{M_2}(x, \nu | x) \), we get the following results.

a) The temporal potentials

\[
\forall (a, \nu, \nu') \in A \times \Lambda,
\sum_{k \in S} P_{M_a}(x_k(p) = \nu | x_{(k-1)}(p)) = \Gamma_a(\nu, \nu' | x), \tag{20}
\]

where \( S = \{k, p \in \{0, \ldots, K\} \times R \mid x_{(k)}(p) = \nu'\} \).

When dealing with the simplest temporal clique model \( A = [a_0] \), i.e., the pixel \( p \) in the \( k \)th frame depends only on the same position in the \( (k-1) \)th frame, the potential values is derived as [6]

\[
\Psi_{M_a}(\nu, \nu') = \ln \left( \Gamma_a(\nu, \nu' | x) \right). \tag{21}
\]

b) The spatial potentials

\[
\forall (\nu, \nu') \in \Lambda^2,
\sum_{k, p \in S'} P_{M_a}(x_k(p) = \nu | x_k(p')) = \Gamma_a(\nu, \nu' | x), \tag{22}
\]

where \( S' = \{k, p, p' \in \{0, \ldots, K\} \times R \mid x_{(k)}(p') = \nu'\} \).

Then, following the similar derivation in the temporal case, we obtain the following equation:

\[
\Psi_{M_a}(\nu, \nu') = \ln \left( \frac{\sum_{x \in \Lambda} \Gamma_a(\nu, \nu' | x)}{\sum_{x \in \Lambda} \Gamma_a(\nu, \nu' | x)} \right). \tag{23}
\]

Therefore, only the temporal co-occurrence matrix \( \Gamma_a(x) \) and the spatial co-occurrence matrix \( \Gamma_a(x) \) are needed to measure the likelihood \( P_a(x) \).

4. Video Classification

4.1. Distance Measurement

After obtaining the motion models for each video clip, we need a distance measurement to estimate their distance. Here we use the Kullback-Leibler (KL) divergence to measure the distribution distance.

Given two video clips \( n_1 \) and \( n_2 \), let \( x^{n_1} \) and \( x^{n_2} \) be their corresponding motion quantities, and \( M^{n_1} \) and \( M^{n_2} \) be their motion models. Then, the KL divergence can be approximated by [6]

\[
KL(M^{n_1} \parallel M^{n_2}) = \frac{1}{K} \ln \left( \frac{P_{M^{n_1}}(x^{n_1}) P_{M^{n_2}}(x^{n_2})}{P_{M^{n_1}}(x^{n_2}) P_{M^{n_2}}(x^{n_1})} \right)
\]

\[
= \frac{1}{K} \ln \left( \frac{\sum_{\nu, \nu'} \Gamma_a(x^{n_1}, \nu, \nu')} {\sum_{\nu, \nu'} \Gamma_a(x^{n_2}, \nu, \nu')} \right). \tag{24}
\]

In order to have a symmetric measurement, we use \( D(n_1, n_2) \) to measure the distribution distance between the two videos \( n_1 \) and \( n_2 \).

\[D_{\text{KLD}}(n_1, n_2) = \frac{1}{2} [KL(M^{n_1} \parallel M^{n_2}) + KL(M^{n_2} \parallel M^{n_1})]. \tag{25}\]

4.2. K-means clustering

We employ the K-means algorithm with splitting technique to cluster all the video clips in a database. Initially, all the video clips are assumed to be within the same cluster. Next, we find out one video clip with the largest distance to its corresponding cluster center as a new cluster center and perform the K-means algorithm to decompose the video database into two clusters. Then, we iteratively perform the splitting step and the K-means algorithm until achieving a pre-defined number of clusters.

5. Experimental results

We conduct experiments over 100 video clips of length 1.5 seconds from the MPEG-7 test content. To construct the ground truth for performance evaluation, we manually classify the 100 clips into 4 classes: (I) videos containing one large object with low motion
activity, (II) videos containing several objects with middle motion activity, (III) videos containing multiple objects with high motion activity, and (IV) videos containing fragmental moving objects with very high motion activity. Fig. 2 shows some of our test videos.

In our experiments, we measure the locally smoothed normal flow [6,8] as the motion quantities. Classification results of using the temporal Gibbs model [6] and using our proposed model are shown in Table 1 and Table 2, respectively.

As shown in Tables 1 and 2, the classification results of the proposed model indeed outperform that of the temporal Gibbs model [6].

6. Conclusion

In this paper, we proposed a unified model to characterize the motion distributions for video content classification. The proposed approach represents the motion distributions in video clips by the spatial and temporal Gibbs models. The experimental results demonstrate that our approach indeed outperform the method of using only the temporal model [6].

7. References