TIME-ORDERED SPATIAL-TEMPORAL INTEREST POINTS FOR HUMAN ACTION CLASSIFICATION

Mengyuan Liu†, Chen Chen‡ and Hong Liu†∗

†Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, China
‡Center for Research in Computer Vision, University of Central Florida, USA

liumengyuan@pku.edu.cn chenchen870713@gmail.com hongliu@pku.edu.cn

ABSTRACT

Human action classification, which is vital for content-based video retrieval and human-machine interaction, finds problem in distinguishing similar actions. Previous works typically detect spatial-temporal interest points (STIPs) from action sequences and then adopt bag-of-visual words (BoVW) model to describe actions as numerical statistics of STIPs. Despite the robustness of BoVW, this model ignores the spatial-temporal layout of STIPs, leading to misclassification among different types of actions with similar numerical statistics of STIPs. Motivated by this, a time-ordered feature is designed to describe the temporal distribution of STIPs, which contains complementary structural information to traditional BoVW model. Moreover, a temporal refinement method is used to eliminate intra-variations among time-ordered features caused by performers’ habits. Then a time-ordered BoVW model is built to represent actions, which encodes both numerical statistics and temporal distribution of STIPs. Extensive experiments on three challenging datasets, i.e., KTH, Rochester and UT-Interaction, validate the effectiveness of our method in distinguishing similar actions.

Index Terms—Human action classification, Spatial-temporal interest point

1. INTRODUCTION

Human action classification based on action sequences has been a core topic in content-based video analysis and intelligent surveillance for decades, while it is still challenging due to problems like inter-ambiguities and clustered backgrounds. Early methods based on local features and bag-of-visual words (BoVW) model have achieved promising results [1, 2, 3]. These methods firstly extract spatial-temporal interest points (STIPs) from training videos and cluster STIPs into words. Then, BoVW model is adopted to describe original video by a histogram of words. Despite the robustness of BoVW model against disturbing motions and clustered backgrounds, this model only involves the numerical statistics of STIPs and ignores the spatial-temporal distribution of STIPs which is vital to distinguish similar actions.

To solve this problem, spatial-temporal relationships among STIPs have been explored to make up for the BoVW model. Latent topic models such as the probabilistic Latent Semantic Analysis (pLSA) model was introduced in [4] to learn the probability distributions of the whole words. Considering words in pairs, a spatial-temporal relationship matching method was proposed in [5] which explored temporal relationships (e.g., before and during) as well as spatial relationships (e.g., near and far) among pairwise words.

Despite modeling spatial-temporal relationships, researches [6, 7, 8] proved the efficiency and discriminative power of using sole temporal ordering of movements to model actions without referring any spatial relationships. Nowozin
et al. [6] used arbitrarily fixed number of uniformly sized temporal bins to assign a sequence of individual local features. Tang et al. [7] also adopted fixed temporal bins divided from a video to discover discriminative segments of video and to learn the transitions between and durations of these segments with a chain structured model. An improved method in [8] decomposed a video sequence into many temporal segments of variable length, and trained a discriminative model that encodes a temporal decomposition of video sequences. However, these methods cannot capture the temporal information in each segment, since STIPs from the same segment are supposed to share the same temporal information.

This work proposes a time-ordered feature for temporal modeling without directly referring any spatial distribution information (shown in Fig. 1). To describe the temporal distribution of STIPs, we firstly assign STIPs to multi-channel temporal distributions. Then, a temporal distribution refinement method is adopted to suppress the effect of performers’ habits on temporal distributions. Since the STIPs are usually sparse, we further estimate movement distributions to enhance the temporal distributions of STIPs. Finally, a time-ordered BoVW model is built to encode both numerical statistics and temporal distributions of STIPs.

Our main contributions are two-fold. (a) A time-ordered feature is proposed to describe the temporal distribution of STIPs. This feature suffers less from the effect of performers’ habits and encode more abundant temporal relationships among STIPs when compared with previous temporal modeling methods [6, 7, 8]. (b) A time-ordered BoVW model is built upon the time-ordered feature. This model captures the complementary property between numerical statistics and temporal distributions of STIPs, therefore it outperforms BoVW model and most related works.

2. TIME-ORDERED BOVW MODEL

Human actions can be featured by sequences of basic movements, which are produced by moving body parts in time order. To describe an action, a new time-ordered local feature is designed in this section, which captures the temporal distribution of the whole movements by local features. The pipeline of extracting the proposed feature is shown in Fig. 2. Noting that STIPs are denoted by colored dots, and multi-channels refer to different types of labels. Specifically speaking, basic movements are firstly described by local features and clustered into labels. Then, local features are split into multi-channels based on their labels. In each channel, a refinement criterion is applied to refine the temporal distribution, and the refined temporal distribution is converted to a basic movement distribution by weighted histogram estimation method. Finally, a time-ordered feature is formed for the action by combining multi-channel movement distributions.

2.1. Multi-channel Temporal Distribution

In Fig. 2, an action is denoted by a cloud of labeled STIPs. To construct the multi-channel temporal distribution, an action sequence V is firstly represented by a set of labeled STIPs \( \{S_k\}_{k=1}^{K} \), where \( k \) refers the \( k \)-th channel. \( S_k \) contains all STIPs labeled \( k \) extracted from \( V \). \( S_k \) is denoted as \( S_k = \{P_n\}_{n=1}^{N_k} \), where \( N_k \) means the whole number of STIPs labeled \( k \). Let \((x_{k,n}^i, y_{k,n}^i, t_{k,n}^i)\) denote the coordinates and time label of a STIP \( P_{n,k}^i \).

Let \( M \) denote the maximum number of frame in \( V \) which involves at least one STIP:

\[
M = \max \left( \{t_{n,k}^{N_k}\}_{k=1}^{K} \right). \tag{1}
\]

The temporal distribution for channel \( k \) is denoted as \( \mathcal{T}_{k} = \{T_{k,n}\}_{n=1}^{M} \), where \( T_{k,n} \) denotes the number of STIPs labeled \( k \) in the \( n \)-th frame. \( T_{k,n} \) is formulated as:

\[
T_{k,n} = \sum_{n=1}^{N_k} \delta(i, t_{k,n}^i) \quad \text{where}
\]

\[
\delta(i, t_{k,n}^i) = \begin{cases} 
1, & \text{if } i = t_{k,n}^i \\
0, & \text{otherwise}.
\end{cases}
\tag{2}
\]

We treat \( \mathcal{T}_{k} \) as the temporal distribution of STIPs labeled
A variable $R$ extracted from $T_k(k = 1, ..., K)$ is determined as $R = \{R^i\}_{i=1}^{M}$, where $R^i$, indicating whether there exists STIPs in frame $i$ or not, is defined as:

$$R^i = \begin{cases} 1, & \text{if } \sum_{k=1}^{K} T^i_k > 0 \\ 0, & \text{otherwise} \end{cases}.$$  

(3)

Given $R$ and $T_k$, the refined time-ordered feature $\tilde{T}_k$ in channel $k$ is formulated as a subset of $T_k$,

$$\tilde{T}_k = \bigcup_{i = 1, ..., M} T^i_k.$$  

(4)

For an action sequence $V$, the refined multi-channel temporal distribution feature is $\{\tilde{T}_k\}_{k=1}^{K}$. The pipeline of extracting refined temporal distribution in the $k_{th}$ channel is illustrated in Fig. 3, where the intervals in the third bin of $\tilde{T}_k$ is removed by $R$ which is extracted from Fig. 3 (b), and the rest bins of $\tilde{T}_k$ are concatenated to form $\tilde{T}_k$.

### 2.3. Movement Distribution Estimation

In channel $k$, $\tilde{T}_k$ represents the temporal distribution of STIPs which are labeled $k$. Since limited STIPs are detected for each frame, movements which are featured by these STIPs are not fully captured. In this section, we treat detected STIPs as sampling points and estimate the real temporal distribution of movements using weighted histogram estimation method.

Histogram estimation [13] is one simple density estimation method, which is defined as:

$$\hat{f}(x) = \sum_{i=1}^{L} \left( \frac{1}{N} I(x \in B_i) \right),$$  

(5)

where $L$ is the number of bins, $N$ is the number of sampling points, $i$ means the observed point number in $i_{th}$ bin, $I$ equals one when $x \in B_i$ and equals zero when $x \notin B_i$, and all bins is denoted as:

$$B_1 = [0, \frac{1}{L}), B_2 = [\frac{1}{L}, \frac{2}{L}), ..., B_L = [\frac{L-1}{L}, 1).$$  

Despite the efficiency of histogram estimation, this method neglect the relationships among neighborhoods, when applied to estimate the temporal distribution of movements. For example, when a movement appears in the second frame of an action sequence, there exists high probability for this movement appearing in the first frame. While the probability estimated by counting numbers in bins is either the same as the second frame or zero, which cannot reflect real distribution.

To solve this problem, a weighted histogram estimation method is proposed to estimate the movement distribution, where each sampling point is assigned a weight and one point affects all other sampling point in this method.

We change $\tilde{T}_k$ into another form $\{l_i\}_{i=1}^{N_k}$, where $l_i$ means the $i_{th}$ sampling point, and $N_k$ is the total number of sampling points. A weight function is defined as a gaussian distribution with parameter $\delta$, which controls the shape of the distribution.

Note that $\delta$ is set to 3 in our experiments. Let $|l_i - l_j|$ denote the distance between $l_i$ and $l_j$. Final value for sample point $l_i$ is defined as:

$$w(l_i) = \sum_{j=1}^{N_k} \left( \frac{1}{\sqrt{2\pi}\delta} \exp\left(-\frac{|l_i - l_j|^2}{2\delta^2}\right) \right).$$  

(7)

Let $\text{index}(l_i)$ refer the index of $l_i$ in $\tilde{T}_k$, and $\text{Length}$ denote the length of vector $\tilde{T}_k$. Then, $\tilde{v}_i$ is re-defined as:

$$\tilde{v}_i \leftarrow \sum_{j=1}^{N_k} \left( w(l_j) \cdot I(\text{index}(l_j) \leq \text{Length}) \right).$$  

(8)

In the $k_{th}$ channel, a basic temporal distribution of movements $Q_k$ with $L$ dimension is estimated from $\tilde{T}_k$ using Formula 5.

The time-ordered feature combing all channels $\{1, ..., K\}$ is formulated as $Q = \{Q_k\}_{k=1}^{K}$. There are two reasons for
Algorithm 1 Time-Ordered BoVW Model

Require: Video \( V \), parameters \( K, L, D, \alpha \)
Ensure: Video representation \( \mathcal{H} \)
1: STIPs from \( V \) : \( S = \{ S_k = \{ P_k^i = (x_k^i, y_k^i, t_k^i, u_k^i) \}_{i=1}^{N_k} \}_{k=1}^K \)
2: calculate \( M \) by Formula 1
3: for \( k = 1 \) to \( K \) do
4: for \( i = 1 \) to \( M \) do
5: calculate \( T_k^i \) by Formula 2
6: end for
7: \( T_k \leftarrow \{ T_k^i \}_{i=1}^M \)
8: end for
9: calculate \( R \) by Formula 3
10: for \( k = 1 \) to \( K \) do
11: calculate \( \bar{T}_k \) by Formula 4
12: \( \delta \leftarrow 3, L \leftarrow 3 \)
13: for \( i = 1 \) to \( N_k \) do
14: calculate \( w(l_i) \) by Formula 7
15: end for
16: for \( i = 1 \) to \( L \) do
17: calculate \( \hat{v}_i \) by Formula 8
18: end for
19: calculate \( Q_k \) by Formula 5
20: end for
21: \( \tilde{Q} \leftarrow \{ Q_k \}_{k=1}^K \)
22: \( D \leftarrow \min \{ \text{num}, L \cdot K \} \)
23: calculate \( \hat{Q} \) by Formula 9
24: calculate \( B \) by BoVW model
25: \( \beta \leftarrow 1 - \alpha \)
26: return \( \mathcal{H} = \{ \tilde{Q} \cdot \alpha; B \cdot \beta \} \)

further reducing the feature dimension of \( Q \). For one reason, the dimension of \( Q \) is \( L \cdot K \), which is still high and not time efficient for classification. For another reason, there exist redundant temporal information in \( Q \), which may cause misclassification. The reduction step is given as:

\[
\tilde{Q} = \mathbb{P}(Q, D), \tag{9}
\]

where \( \mathbb{P} \) stands for standard principal component analysis (PCA) method, and \( D \) means the length of new feature \( \tilde{Q} \). Noting that \( D \) is no larger than neither the number of training samples for PCA nor the dimension of \( Q \).

Time-ordered feature and the numerical statistics of STIPs are naturally complementary, since the former encodes extra temporal distribution of STIPs which is ignored by the later one. In time-ordered BoVW model, we combine both features and form final action representation as:

\[
\mathcal{H} = \{ \tilde{Q} \cdot \alpha; B \cdot \beta \} \text{ s.t. } \alpha + \beta = 1, \tag{10}
\]

where \( B \) denote the BoVW representation, \( \alpha \) and \( \beta \) are weights for \( Q \) and \( B \) respectively. The algorithm of the proposed time-ordered BoVW model is detailed in Algorithm 1, where \( S, S_k, T_k, \tilde{T}_k, Q_k, \tilde{Q}_k, B, \mathcal{H} \) are illustrated in Fig. 2. In line 22, \( \text{num} \) denotes the number of training samples. Main time cost lies in applying PCA method to \( Q \). A fast PCA method [14] is utilized in this work with computational complexity of \( O(L^2D + L^2\beta) \), where \( \beta \) is the number of training samples used. When dimension \( L \) is far larger than \( D \) and \( \beta \), the computational complexity can be simplified as \( O(L^2D + L^2\beta) \).

3. EXPERIMENTS AND ANALYSES

3.1. Benchmark Datasets

Our method is evaluated on three benchmark datasets: KTH [15], Rochester [16] and UT-Interaction [17]. KTH dataset contains 600 videos of 25 persons performing 6 actions for 4 times with homogeneous indoor/outdoor backgrounds. Rochester dataset contains 150 videos of 5 actors performing 10 actions for 3 times. UT-Interaction (UT) dataset contains 6 types of actions repeated 10 times in two scenes resulting in 120 videos. Scene-1 is taken in a parking lot with little camera jitter and slightly zoom rates. In scene-2 (UT2), the backgrounds are cluttered with moving trees, camera jitter- s and passerby. Actions like “walking”, “jogging” and “running” are similar in KTH dataset, and actions like “answer a phone” and “dial a phone” are alike in Rochester dataset. In UT-Interaction dataset, actions like “kick” and “punch” share similar movements. Besides, the complex filming scenes in UT2 also brings difficulty for classification. Some action s- naps are shown in Fig. 4.

3.2. Experimental Settings

This work applies Laptevs detector in [15] obeying original parameter setting to detect STIPs and uses HOG/HOF in [1] to generate 162 dimension descriptors with 90 dimension for HOG and 72 dimension for HOF. After extracting STIPs from each video, K-means clustering is applied to generate visual words. Recognition is conducted using a linear SVM [18]. To evaluate the proposed model, we set default parameters \( L \) to 3 and set \( D \) to the number of training videos. Default values of parameter \( K \) and \( \alpha \) are shown in Table I. To ensure fair comparison with state-of-the-art works on KTH dataset, we use the dense trajectory feature [19] instead of STIPs to implement our model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>KTH</th>
<th>Rochester</th>
<th>UT</th>
<th>UT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K )</td>
<td>900</td>
<td>2300</td>
<td>2100</td>
<td>1900</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>0.6</td>
<td>0.5</td>
<td>0.7</td>
<td>0.6</td>
</tr>
</tbody>
</table>
3.3. Evaluation of Parameters

Fig. 5 evaluates effect of parameters, i.e., $K$, $L$, $D$, and $\alpha$, on the performance of our method. In Fig. 5 (a) and (d), more than 85% recognition accuracies are gained on KTH dataset when $K$ or $\alpha$ changes in a large range. In Fig. 5 (b), recognition precision grows in the beginning and then drops from the middle to the end, when parameter $L$ changes from 1 to 10. This phenomenon denotes that much information is ignored when $L$ is too small, and that the effect of quantization drops when $L$ is too big. Recognition precision grows when the parameter $D$ changes from 10% to 100% of the number of training videos, which is shown in Fig. 5 (c). Obviously, more information is retained to achieve higher recognition results when $D$ grows.

3.4. Comparison with related works

In Table II, our method achieves better performances on all four datasets than related works. Confusion matrices of our method are shown in Fig. 7. To ensure fair comparison with traditional BoVW model, we implement BoVW with parameter $K$ ranging from 100 to 4000 at 100 interval, and the highest accuracy is selected as the performance of BoVW. Our “Time-Ordered BoVW” method achieves higher accuracies than BoVW on all benchmark datasets. Specifically, “Time-Ordered BoVW” achieves 5.83% higher than BoVW on KTH dataset. Fig. 6 shows that the ambiguities between similar actions like “jogging” and “running” are eliminated by our method. Our method also performs better than related works using trajectory-based feature, e.g. Relative dense tracklets [20], or deep learning method, e.g. dRNN + HOG3D [21].

4. CONCLUSIONS AND FUTURE WORK

In this paper, we evaluate the significance of temporal information for human action classification without relying on any spatial clues. We propose a new time-ordered BoVW model which explores both temporal distributions and numerical statistics of STIPs. In general, our model shows superior classification precisions than traditional BoVW model and some recent works, which verify the robustness and effectiveness of our time-ordered feature. Future work focuses on combining extra information from spatial-temporal distribution of STIPs for more challenging tasks, such as human action detection and tackling with clustered backgrounds. Fusing other types of data, e.g., inertial data [28, 29, 30], depth data [31, 32], skeleton data [33], is also a new way for action classification.

5. REFERENCES


