Research on Behavior Recognition Algorithm Based on Compact Representation of Low-level Feature

To cite this article: Yun Liu et al 2019 J. Phys.: Conf. Ser. 1176 032018

View the article online for updates and enhancements.
Research on Behavior Recognition Algorithm Based on Compact Representation of Low-level Feature

Yun Liu 1,*, Jianbin Yang 1, b and Chuanxu Wang 1, c

1 No.99 Songling Rd, Qingdao, Shandong, China

*Corresponding author e-mail: lyun-1027@163.com
b 2252128507@qq.com, c wangchuanxu_qd@163.com

Abstract. Recently, the bag of visual words (BOVW) model based on spatial-temporal interest points (STIPs) is used more and more widely in the field of behavior recognition. However, the model ignores temporal order between frames and intra frame position information of interest points. In the paper, an algorithm is proposed to acquire the geometrical and temporal distribution of STIPs. Firstly, the STIPs mutual information (STIPsMI) algorithm based on Co-occurrence matrix is proposed to describe the spatial-temporal relationship of STIPs between different visual words. Then the descriptor is concatenated with the BOVW histogram as the final descriptor. The two authoritative human action datasets were used to test the algorithm in paper: the KTH and the UCF sports. Experimental results verify the robustness of the algorithm and better than BOVW model and other mainstream methods.

1 Introduction

In the field of computer vision, behavior recognition and detection are very important modules in the past ten years. With the continuous development and maturity of Internet technology and the continuous expansion of video surveillance applications, more and more applications involve the automatic recognition of video events. Among this, behavior recognition is a hot topic, and widely used in surveillance, virtual reality, content-based video retrieval, human-machine interaction[1].

The mainstream methods of behavior recognition are relying on low-level feature extraction, which can describe the human motion from the video effectively; further to learn behavior patterns from these low-level features and then classify behavior categories[2]. However, human behaviors are complex and diverse, which results in no feature models can be used universally.

Proverbially, STIPs-based BOVW model for behavior recognition is effective and easy realized for neither need of moving object detection and tracking, nor background modeling. It provides an approach to understand and analyze the interesting points from the video sequences directly. So, the BOVW model is widely used in the research of behavior recognition[3-12]. However, the BOVW model ignores the spatial location information and temporal structure knowledge of STIPs inside a video, which has proved to be essential for boosting behavior recognition discrimination[4].

Due to these shortcomings of the BOVW model, this paper improves the BOVW model and proposes a behavior recognition algorithm based on the STIPs mutual information. Firstly, the Dollar et al.[12] method is used to extract the interesting points and described by HOG3D feature. Afterward, the K-means algorithm is used to generate the bag of visual words by clustering these HOG3D
features. Then, the spatio-temporal relationship between pair-wise visual words is represented by STIPs mutual information descriptor based on the co-occurrence matrix. Finally, the two kinds of information: BOVW histogram and STIPs mutual information are fused as the descriptors of a video sequence, SVM classifier is used for recognition and Classification. The overall algorithm flow is shown in Figure 1.

The paper is arranged as follows. Section 2 is related work about the improved BOVW methods. Section 3 introduces the way of extraction and description of spatial-temporal interest points. Section 4 gives the detailed procedures to compute STIPs mutual information. Section 5 shows the results on two challenge datasets. Section 6 analyzes and summaries the paper.

Fig.1. The flowchart of the paper method framework

2 Related Work

In recent year, The algorithms for integrating geometric information into BOVW have been proposed. The most common method is to use spatial-temporal pyramid(STP). In the method, descriptor level statistics are pooled and the video sequence is divided into a set of spatial and temporal cells repeatedly and equably. The approach can capture the spatial layout and temporal order of an roughly action sequence. However, directly using the BOVW histogram to represent video increases the feature scale while greatly increasing the cost of learning and storage[4].

Later, some scholars put forward the modeling methods of spatial-temporal interest point. Wu X et al.[5] propose to use multiple Gaussian Mixture Models (GMMs) at different time and space scales to represent the distribution of local spatial-temporal background between points of interest as a feature representation. Q Hu et al.[6] propose the spatial-temporal context in spatial-temporal domain. The Spatial-temporal context is a representative video collection including video words and context words. Actually, the context can obtain and understand the semantics of video words.Wang et al.[7] propose to use dense trajectories and motion boundary information to describe video. X Yang et al.[4] propose an effective coding scheme to merge spatial-temporal information by aggregating low-level descriptors into the super descriptor vector(SDV). Kovashka et al.[8] propose to form candidate neighborhoods by learning the shapes of the neighborhood of space-time features, and then composing descriptors with words associated with nearby points and their orientation relative to the central point of interest. Wong and Cipolla[9] propose that the non-negative matrix factorization(NNMF) is applied to the entire video sink to achieve the purpose of adding global information to the point of interest detection. Bregenzio et al.[10] extract to model the global space and time distribution by extracting the overall features of the point cloud of interest accumulated on multiple time scales. C Yuan et al.[11] use the 3D discrete Radon transform, also called R transform, which is extended by 2D Radon transform to capture the detailed geometric distribution of points of interest to obtain global characterization. However, these methods are too complex and have not been received extensive attention. Therefore, it is still a challenging work to datamine these STIPs of a video for better behavior recognition.

3 STIPs extraction and description

In the paper, We use the interest point detector proposed by Dollar et al.[12]which can determine the spatial-temporal salience by calculating its response value at each pixel in the video sequence with two-dimensional Gauss functions and one temporal Gabor wavelet function. The detected STIPs are shown as shown in Fig.2. The response function is defined as equation (1).
\[ R = (S \ast g \ast h_{ev})^2 + (S \ast g \ast h_{od})^2 \]  \hspace{1cm} (1)

Where, \( S \) as the video sequence, \( g(x, y, \sigma) \) as two dimensional spatial Gauss smoothing filter, \( h_{ev} \) and \( h_{od} \) as one dimensional Gabor filter in time domain. The two functions are defined as equations (2) and (3) respectively.

\[ h_{ev}(t, \tau, \omega) = -\cos(2\pi t \omega)e^{-r^2/\sigma^2} \]  \hspace{1cm} (2)

\[ h_{od}(t, \tau, \omega) = -\sin(2\pi t \omega)e^{-r^2/\sigma^2} \]  \hspace{1cm} (3)

\( \sigma \) and \( \tau \) are two parameters of the spatial scale and temporal scale, respectively. \( \omega = 4 / \tau \).

It is usually necessary to describe the spatial-temporal information of the cuboid centered at each interest point, i.e. structuring the spatial-temporal feature descriptor. We employ the HOG3D feature to build this descriptor for each interest point.

4 STIPs Co-occurrence matrix and mutual information

4.1 STIPs Co-occurrence matrix

Inspired by spatial co-occurrence matrix for 2-D interest points in an image[13], this paper extends the concept and applies it to the action recognition task. STIPs co-occurrence matrix (STIPsCM) is to obtain the spatial-temporal correlation information of interest points between different visual words. The set of spatial-temporal interest points \( V \) in a video sequence \( S \) is quantified by BOVW model into \( N \) classes, expressed as \( w_1, w_2, \ldots, w_n \), here, \( w_i \in V, i = 1, 2, \ldots, n \). For any pairs of STIPs \( p_i(x_i, y_i, t_i) \in w_p, p_j(x_j, y_j, t_j) \in w_q \), if their spatial-temporal distance is no longer than the threshold \( r \) defined as equation (4), their relation is called 3-D co-occurrence. This co-occurrence statistics among any two words \( w_q \) and \( w_p \) is defined as equation (5).

\[ \text{Fig.3. STIPs Co-occurrence matrix} \]
\begin{equation}
\begin{aligned}
G &= \begin{cases} 
1, & \text{if } \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (t_i - t_j)^2} \leq r \\
0, & \text{otherwise} 
\end{cases} \\
CM(w_q, w_p) &= \sum (p_i \in w_p) \land (p_j \in w_q) \land (G) 
\end{aligned}
\end{equation}

Where \( \land \) represents logic AND. In fact, the item \( CM(w_q, w_p) \) represents the total STIPs number of visual word \( w_q \) co-occurred with visual word \( w_p \) measured by this spatial-temporal radius \( r \). So the obtained STIPs co-occurrence matrix for each pair-wise visual words is an \( n \times n \) matrix, as shown in Fig.3. Their detailed co-occurrence relation is shown in Fig.4. Finally, in order to get the co-occurrence probability distribution, each row in Fig.3. can be normalized by dividing the sum of its row elements.

Fig.4. The structure of STIPs Co-occurrence statistics among any two words

### 4.2 STIPs mutual information

Although the STIPs co-occurrence matrix can represent the spatial-temporal distribution of interest points among different words, the \( n \times n \) dimension of the co-occurrence matrix is high, which results in costly computation and more storage memory[14,15]. In the paper, we propose the concept of spatial-temporal interest points mutual information based on the co-occurrence matrix, which not only reduces the dimension of feature vector to \( N \), but also conserves the spatial-temporal information.

The concept of mutual information was proposed originally in the theory of information, which means that the amount of information provided by event \( Y \) for another event \( X \), is defined as equation (6).

\begin{equation}
I(X; Y) = \log_2 \frac{p(X \mid Y)}{p(X)}
\end{equation}

The STIPs co-occurrence matrix represents the spatial-temporal distribution of all pairs of spatial-temporal interest points. Each element in the \( p \)-th row in the co-occurrence matrix represents the probability of the visual word \( w_p \) matching with other visual words. The paper gives the definition of mutual information of visual word \( w_p \) for \( w_q \) as equation (7).

\begin{equation}
I(w_q; w_p) = \log_2 \frac{p(w_q \mid w_p)}{p(w_q)} = \log_2 \frac{\gamma_{w_q, w_p}}{p(w_q)}
\end{equation}

\begin{equation}
\gamma_{w_q, w_p} = \frac{CM(w_q, w_p)}{\sum_{q=1}^{n} CM(w_q, w_p)}
\end{equation}

\( \gamma_{w_q, w_p} \) is a way to calculate conditional probability of \( p(w_q \mid w_p) \). In this paper, the average mutual information provided by visual word \( w_p \) for all of other words is obtained by equation (9).
\[ I(W; w_p) = \sum_{j=1}^{n} I(w_j; w_p) p(w_j | w_p) = \sum_{j=1}^{n} (\log \frac{\gamma_{w_j,w_p}}{p(w_j)} \gamma_{w_j,w_p} ) \] (9)

Where, \( p(w_j) \) is the frequency of the \( j^{th} \) word in BOVW histogram descriptor. By this transformation, the \( n \times n \) dimensional co-occurrence matrix of spatial-temporal interest points in section 4.1 can be reduced as a \( n \)-dimensional mutual information descriptor, which is called STIPs mutual information, shortened as STIPsMI.

5 Algorithm validation and data analysis

In this section, we will evaluate the robustness of the proposed algorithm by experimental results on authoritative datasets: KTH and UCF sports. The output is identified by a leave-cross-validation (LOOCV) method of linear SVM. The KTH dataset includes 6 kinds of behaviors. Each video data was collected from 25 individuals in 4 different scenarios, as shown in figure 2. The UCF dataset consists of 150 video sequences including of 10 kinds of behavior, as shown in figure 5. The dataset has been used in many applications widely, such as: action recognition, motion positioning and saliency detection.

The main parameters of the algorithm include temporal scale \( t \), spatial scale \( \sigma \), the number \( n \) of BOVW, the scale \( r \) of the location operator. For KTH dataset, according to previous work and a priori knowledge[11], the spatial scale can be set as 1.5, the temporal scale as 1.5, the number of BOVW is 500. Similarly, according to previous work and a priori knowledge[4], initializations for UCF sports database go as follows, the spatial scale is 2, the time scale is 2.5, the number of BOVW is 800. Therefore, in the paper, we mainly test the performance of parameter \( r \) and two kinds of low-level descriptors (BOVW, STIPsMI) on the recognition accuracy.

5.1 Performance of three low-level descriptors (BOVW, STIPsMI) on recognition accuracy

To verify the feasibility of the algorithm, the two basic descriptors of BOVW and STIPsMI, as well as their different combinations were tested on KTH and UCF sports datasets respectively.

| Method | KTH | BoHandclap | Handwave | Jog | Run | Walk | Average |
|--------|-----|------------|----------|-----|-----|-----|---------|
| BOVW   | 0.9 | 0.92       | 0.97     | 0.81| 0.78| 0.93| 0.894   |
| STIPsMI| 0.96| 0.92       | 0.97     | 0.82| 0.78| 0.93| 0.8942  |
| BOVW+BOVW| 0.96| 0.94       | 0.97     | 0.79| 0.74| 0.94| 0.8900  |
| STIPsMI+STIPsMI| 0.96| 0.93       | 0.97     | 0.78| 0.73| 0.93| 0.8846  |
| BOVW+STIPsMI| 0.96| 0.97       | 1.0      | 0.94| 0.92| 0.95| 0.9529  |

Fig.5. Samples of the UCF sports dataset
Table 2. Comparison of ten methods on the UCF dataset.

|             | Dive | Golf | Kick | Lift | Ride | Run | Skete | Swing1 | Swing2 | Walk | Average |
|-------------|------|------|------|------|------|-----|-------|--------|--------|------|---------|
| BOVW        | 0.93 | 0.83 | 0.80 | 1.0  | 0.58 | 0.64| 0.50  | 0.80   | 0.85   | 0.95 | 0.7933  |
| STIPsMI     | 1.0  | 0.83 | 0.75 | 1.0  | 0.67 | 0.79| 0.67  | 0.85   | 0.92   | 0.86 | 0.8267  |
| BOVW+BOVW   | 1.0  | 0.83 | 0.80 | 1.0  | 0.58 | 0.64| 0.50  | 0.85   | 0.92   | 1.0  | 0.90 8133 |
| STIPsMI+STIPsMI | 1.0 | 0.83 | 0.75 | 1.0  | 0.67 | 0.79| 0.67  | 0.90   | 0.92   | 0.86 | 0.8333  |
| BOVW+STIPsMI| 1.0  | 0.78 | 0.85 | 1.0  | 0.67 | 0.79| 0.92  | 0.95   | 1.0    | 0.86 | 0.8733  |

Table 1 and Table 2 show that recognition accuracy based on these three kinds of descriptors (BOVW+BOVW, STIPsMI+STIPsMI) are not improved significantly or even declined. However, recognition accuracy has been improved significantly when different descriptors are cascaded. The performance of descriptor (BOVW+STIPsMI) improve by 4.5%-6.0% compared with that of BOVW descriptor tested on the KTH and UCF datasets. This shows that local features are important in the descriptors of the video. The results also demonstrates that the STIPsMI and BOVW descriptors are complementary. That is, STIPsMI describes the local relationships between different visual words, which make up for BOVW descriptor.

In terms of KTH dataset, recognition accuracy of the three descriptor fusions are up to 95.29%. For jog and run behaviors themselves have large similarities, it is easy to produce some confusion, so the recognition accuracy is relatively low; the other 4 kinds of behavior can be identified accurately. As to UCF database, the highest recognition accuracy is 87.33%. STIPsMI stands for local distribution of STIPs, and STIPsMI performs well whether the video contains multi person behavior or single behavior.

5.2 Performance of parameter \( r \) on recognition accuracy

Changes of \( r \) can exert remarkable impacts on recognition accuracy, as shown in Figure 6, which are experimental results on KTH and UCF sports datasets. The optimal values of parameter \( r \) is about [20,25].

![Fig.6. Influence of parameter \( r \)](image)

6 Conclusion

Aiming at the limitation of BOVW model in behavior recognition, the paper proposes a behavior recognition algorithm based on STIPs mutual information. The STIPs mutual information is used to represent the spatial-temporal distribution of interest points between different visual words in a small neighborhood. Table 3 presents a comparison of our results with state-of-the-art results, which are based on the BOVW model. The citations of papers are over 350, and two of them are cited more than 1000 times, so they can be compared as a benchmark.
Table 3. Evaluation results on the KTH dataset and the UCF dataset. (%)

| Methods              | KTH  | UCF  |
|----------------------|------|------|
| Our work             | 95.29| 87.33|
| Dollar et al.[12]    | 81.2 | -    |
| Bregonzio et al.[10] | 93.1 | -    |
| Kovashka et al.[8]   | 94.5 | 87.2 |
| Le et al.[3]         | 93.9 | 86.5 |
| Wang et al.[7]       | 92.1 | 85.6 |

In addition, STIPsMI based on the co-occurrence matrix can also be considered as a dimension reduction method from 2-D to 1-D, and does not change the sparsity of data. That is to say, the feature is still sparse, if the matrix is sparse. That is one of the reasons why we use STIPsMI instead of 2DPCA or (2D)2PCA. More sophisticated parameter optimizing methods will be explored later, for instance, assign different weights to BOVW and STIPsMI features.

Acknowledgement

Our work is supported by China Nature Science Fund Project 61472196 and 61672305; also I would like to express our thanks to Shandong Nature Science Fund Project ZR2015FM012.

References

[1] Herath, Samitha, et al. Going Deeper into Action Recognition[J]. Image and Vision Computing, 2017 vol. 60, 4-21

[2] Zhen, Xiantong, and Ling Shao. Action Recognition via Spatio-Temporal Local Features[J]. Image and Vision Computing, 2016 50(50) 1-13

[3] Le Q V, Zou W Y, Yeung S Y, et al. Learning hierarchical invariant spatial-temporal features for action recognition with independent subspace analysis [C]. IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2011. 3361-3368

[4] Yang X, Tian Y L. Action Recognition Using Super Sparse Coding Vector with Spatial-temporal Awareness[C]. ECCV. Berlin: Springer, 2014.727-741

[5] Wu X, Xu D, Duan L, et al. Action recognition using context and appearance distribution features[C]. IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE,2011.489-496

[6] Hu Q, Qin L, Huang Q, et al. Action Recognition Using Spatial-Temporal Context[C]. International Conference on Pattern Recognition. New York: IEEE, 2010.1521-1524

[7] Wang H, Kläser A, Schmid C, et al. Dense Trajectories and Motion Boundary Descriptors for Action Recognition[J]. International Journal of Computer Vision, 2013 103(1) 60-79

[8] Kovashka A, Grauman K. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE,2010 2046-2053

[9] Wong S F, Cipolla R. Extracting Spatial temporal Interest Points using Global Information[C]. International Conference on Computer Vision. New York: IEEE,2015.1-8

[10] Bregonzio M, Gong S, Xiang T. Recognising action as clouds of space-time interest points[C].

7
IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2013.1948-1955

[11] Yuan C, Li X, Hu W, et al. 3D R Transform on Spatial-temporal Interest Points for Action Recognition[C]. Computer Vision and Pattern Recognition. New York: IEEE, 2013.724-730

[12] Dollar P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatial-temporal features[C]. Joint IEEE International Workshop on Visual Surveillance and PERFORMANCE Evaluation of Tracking and Surveillance. New York: IEEE, 2006.65-72

[13] Coelho F, Braga A P, Verleysen M. A Mutual Information estimator for continuous and discrete variables applied to Feature Selection and Classification problems[J]. International Journal of Computational Intelligence Systems, 2016 9(4) 726-733

[14] Gecer, Baris, et al. Color-Blob-Based COSFIRE Filters for Object Recognition[J]. Image and Vision Computing, 2017 57(1) 165-174

[15] Gao S, Zhu Z, Liu S, et al. Estimating the spatial distribution of soil moisture based on Bayesian maximum entropy method with auxiliary data from remote sensing[J]. International Journal of Applied Earth Observation & Geoinformation, 2014 32(10) 54-66