Zhang *et al. SpringerPlus (2016) 5:1226*
DOI 10.1186/s40064-016-2876-z

SpringerPlus

RESEARCH Open Access

Multi-surface analysis for human action recognition in video

Hong-Bo Zhang*, Qing Lei, Bi-Neng Zhong, Ji-Xiang Du*, Jialin Peng, Tsung-Chih Hsiao and Duan-Sheng Chen

*Correspondence:
zhanghongbo@hqu.edu.cn;
jxdu@hqu.edu.cn
Department of Computer Science and Technology, Huaqiao University, Fujian, China

Abstract

The majority of methods for recognizing human actions are based on single-view video or multi-camera data. In this paper, we propose a novel multi-surface video analysis strategy. The video can be expressed as three-surface motion feature (3SMF) and spatio-temporal interest feature. 3SMF is extracted from the motion history image in three different video surfaces: horizontal–vertical, horizontal- and vertical-time surface. In contrast to several previous studies, the prior probability is estimated by 3SMF rather than using a uniform distribution. Finally, we model the relationship score between each video and action as a probability inference to bridge the feature descriptors and action categories. We demonstrate our methods by comparing them to several state-of-the-arts action recognition benchmarks.

Keywords: Human action recognition, Multi-view video analysis, Three surfaces motion feature, Probability inference

Background

Human action recognition in video sequences is a challenging research topic in computer vision (Aggarwal and Ryoo 2011; Poppe 2010) and serves as a fundamental component of several existing applications such as video surveillance human computer interaction, multimedia event detection and video retrieval. Extensive efforts have been devoted to action recognition, including: finding a robust, stable, discrimination feature to represent the action/video, such as the motion history image (MHI), motion trajectories of human bodies (Yoon et al. 2014), and spatio-temporal interest point (STIP) (Dawn and Shaikh 2015), and using effective machine learning or pattern recognition methods to identify human action, such as latent support vector machine (SVM) (Zhou et al. 2015), deep learning (Charalampous and Gasteratos 2014) and statistical methods. Facing complex scenes, action recognition in the depth video and multi-camera systems have gained increasing attention in recent years. However, in practical applications and real scenes, such models are not sufficient due to the variations in multiple facets and their high computational cost.

Modeling human actions in hybrid data, such as recognizing an action in RGB-depth data, multi-camera view data and mixed data, is one effective method for human action recognition in complex and dynamic environments. Many works have demonstrated the superior performance obtained when using hybrid data compared to a single data

source. For example, Luo et al. (2013) proposed a framework for the real-time realization of human action recognition in distributed camera networks. Liu et al. (2015) proposed the pyramid partwise bag of words (PPBoW) representation and regarded single/multi-view human action recognition as a multi-task learning problem penalized by the graph structure. Due to the limitations of devices, these methods have many restrictions in real scenes and high computational complexity.

In contrast to previous studies, we find that motion can be represented from multi-surfaces in a signal-view video. The video is expressed based on three surfaces: horizontal–vertical (XY surface), horizontal-time (XT surface) and vertical-time surface (YT surface), as shown in Fig. 1. From the different surfaces, the motion history is extracted and represented as a histogram of the orient gradient (HOG) features to model the holistic action, composing the three-surface motion feature (3SMF). Meanwhile, the STIP feature is extracted to represent the local motion.

However, the fusion of direct features is not sufficient or robust. To this end, we propose to integrate the holistic features and STIP features into an action classifier. A probability inference model is used to identify the action in the video. The proposed multi-surface video analysis method is shown in Fig. 2. In the training stage, a SVM classifier is trained by 3SMF to estimate the prior probability. The STIP feature is extracted, and a naïve Bayes nearest neighbor algorithm (NBNN) is used to estimate the posterior probability. In the testing stage, the test video is also represented as 3SMF and STIP. The action category is determined by probability inference using the prior probability and posterior probability.

The contributions of our work are threefold:

1. We propose a novel multi-surface video analysis strategy that is different from using multiple cameras.
2. We propose a probability method to combine the holistic and local features. In contrast to the majority of previous works, we use 3SMF for the prior probability rather than the uniform distribution.

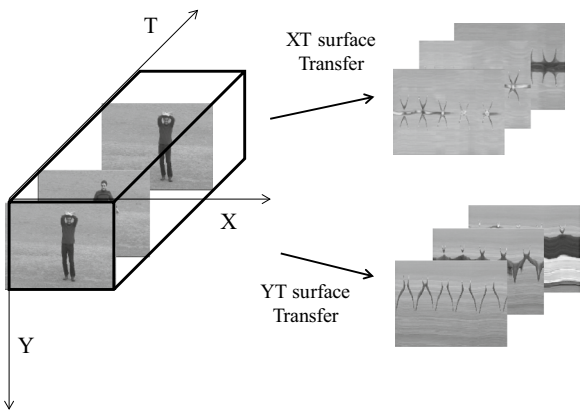

Fig. 1 Sketch of the multi-surface transfer of video. *X* indicates the *horizontal* direction, *Y* indicates the *verticality* direction and *T* indicates the time direction. The image sequence in the *top right* corner is the result of *XT* surface transfer. And the image sequence in the *lower right* corner is the result of *YT* surface transfer

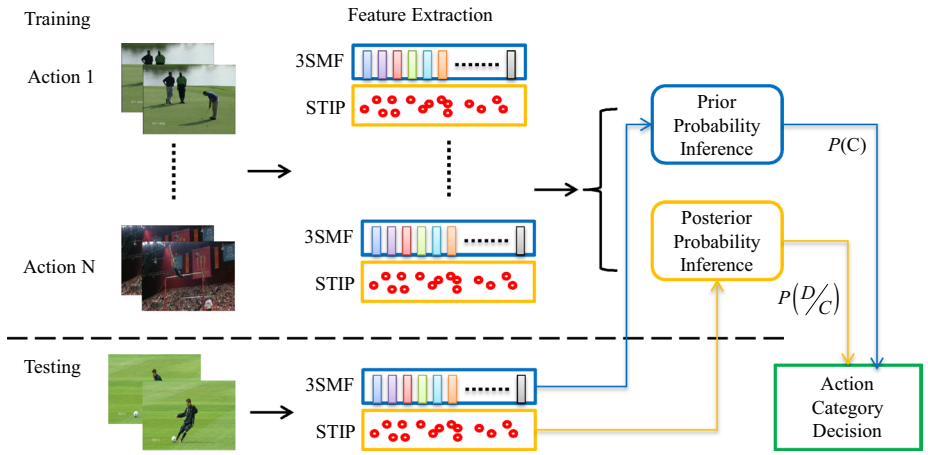

Fig. 2 Outline of the workflow of the proposed approach. The 3SMF and STIP feature have been extracted in the training data and testing data. In training process, prior probability inference model is trained by SVM, and posterior probability is estimated by NBNN algorithm

3. The experimental results show that the proposed method is effective, robust to scene motion and provides accurate results.

The remainder of this paper is organized as follows. "Related work" section introduces the related works of human action recognition. "Algorithm in the proposed method" section describes the algorithms on which the proposed method is based. "Experimental results and analysis" section presents and discusses the experimental results. "Conclusion" section concludes the work.

Related work

The numerous existing methods for recognizing human action from image sequences or video have been classified as template-based approaches, local-feature-based approaches (including appearance and motion features) and object/scene-context-based approaches. Methods of human action recognition in multi-view scenes have been proposed in many studies. A literature review (Aggarwal and Ryoo 2011; Poppe 2010; Dawn and Shaikh 2015; Nissi Paul and Jayanta 2016; Paul and Singh 2014) indicated that the work related to our method includes human action recognition approaches based on a STIP detector, human action recognition approaches with multi-view cameras and human action approaches using object/scene context information.

The STIP detector captures the 3D Harris interest points from a video in the spatio-temporal domain, which was extended from the Harris corner detection by Laptev (2005a). The STIP detector is widely used in human action recognition tasks due to its robustness and good performance. Chakraborty et al. (2012) proposed a novel action recognition algorithm using selective STIPs. Yu et al. (2012) developed a spatial–temporal implicit shape model (STISM) for characterizing the space–time structure of sparse local features. Yan and Luo (2012) proposed a new action descriptor, named the histogram of interest point locations, based on STIPs. Yuan et al. (2011a) proposed the naïve

Bayes mutual information maximization (NBMIM) algorithm based on STIPs for classifying actions. Zhang et al. (2013) proposed an improved version using the ε-NN probability estimation method and the variance filter for discriminative STIP selection. In the proposed method, the ε-NN probability estimation method is used in the NBNN algorithm for the posterior probability.

Multi-camera systems can provide more information for action recognition. Liu et al. (2015) proposed a unified single/multi-view human action recognition method via regularized multi-task learning. Gao et al. (2015) proposed a multi-view discriminative and structured dictionary learning method with group sparsity and a graph model to fuse different views and recognize human actions. Junejo et al. (2011) presented an action descriptor to capture the structure of the temporal similarities and dissimilarities in action sequences. The latent kernelized structural SVM was proposed by Wu and Jia (2012) for view-invariant action recognition. These methods of multi-views have good performance. However, the methods cannot be applied to real scenes due to the high computation complexity and difficultly in correlating information among different views.

The methods discussed above are independent of human action recognition. The concept of using context information for action recognition has been widely adopted in recent studies. Object detection and pose estimation play important roles in the process of recognizing human action. Yao and Fei-Fei (2012) proposed a mutual context model to jointly model objects and human poses in human–object interaction activities. Ikizler-Cinbis and Sclaroff (2010) proposed an approach for human action recognition that integrates multiple feature channels from several entities, such as objects, scenes, and humans. Burghouts et al. (2014) used object tracking trajectories as the context for improving threat recognition. Marszalek et al. (2009) proposed the context of natural dynamic scenes for action recognition. The scene information of the video was extracted from the movie script rather than from image sequences. Similarly, in the proposed method, 3SMF is regarded as the context information of the STIP to recognize human action. In the proposed method, 3SMF and STIP are extracted to model action. A probability inference algorithm is used to determine the action categories.

Algorithms in the proposed method

STIP and 3SMF features

In recent studies, many local features have been successfully used for human action recognition, such as STIP, dense sample, and dense trajectories (DTs). Numerous studies have demonstrated the good performance and robustness of STIP features. In the proposed method, the STIP is extracted by the 3D-Harris detector proposed by Laptev and Lindeberg (2006), and is described by concatenating the HOG and HOF features (162-dimensional feature vector).

To calculate STIP, the video is constructed a spatio-temporal scale-space representation L by convolution with spatio-temporal Gaussian kernel. The second-moment matrix μ of spatio-temporal scale-space representation L is calculated, which is 3-by-3 matrix composed by first order spatial and temporal derivatives. The response function H is defined by combining the determinant and the trace of μ as following:

$$\begin{aligned} H &= \det(\mu) - k \times trace^3(\mu) \\ &= \lambda_1 \times \lambda_2 \times \lambda_3 - k \times (\lambda_1 + \lambda_2 + \lambda_3) \end{aligned} \tag{1}$$

where $\lambda_1, \lambda_2, \lambda_3$ is the eigenvalue of matrix μ. The STIP is defined by searching the maxima of the point with H. In Laptev's work, the parameter k is set to .005 through experimental results.

To describe the action, we propose a new action feature named the 3SMF. The framework of 3SMF is shown in Fig. 3. The 3SMF is a fusion of the features of three different surfaces and is represented by the HOG feature of the MHI. In the proposed method, STIP feature is regarded as local feature, and 3SMF is holistic feature to represent action. And a probability inference model is used to combine STIP feature with 3SMF feature.

For the input video, the original image sequence appears as an XY surface image sequence $V_{xy} = f(x, y, t)$, where $x \in \{1, \ldots, N_x\}$, $y \in \{1, \ldots, N_y\}$, $t \in \{1, \ldots, N_x\}$. N_x is the width of the video, N_y is the height of the video, and N_t is the length of the video. The XT surface image sequence $V_{xt} = f(x', y', t')$ is regarded as the original image sequence rotated 90° along the X direction. Similarly, the YT surface image sequence $V_{yt} = f(x'', y'', t'')$ is regarded as the original image sequence rotated 90° along the Y direction. The XT and YT surface transfer are expressed by Eq. (2)

$$\begin{aligned} (x', y', t') &= (x, y, t) \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{bmatrix} \quad \alpha = \beta = 90^\circ \\ (x'', y'', t'') &= (x, y, t) \begin{bmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{bmatrix} \end{aligned} \tag{2}$$

For the XY, YT and XY surface image sequences, the MHI (Ahad et al. 2012) is extracted for the action description. The MHI approach is a view-based temporal template method that is simple yet robust in representing movements and has been widely employed by many researchers for human action recognition, motion analysis and other related applications. Video is expressed as XYMHI, the XT surface image sequence is expressed as XTMHI and the YT surface image sequence is expressed as YTMHI.

The HOG feature is computed to represent the MHI. HOG was first developed for use in human detection; the method divides an image into small spatial regions called cells. A local histogram of the gradient direction over the pixels in the cell is constructed. Figure 4 shows the framework of the HOG feature. To calculating HOG feature, it

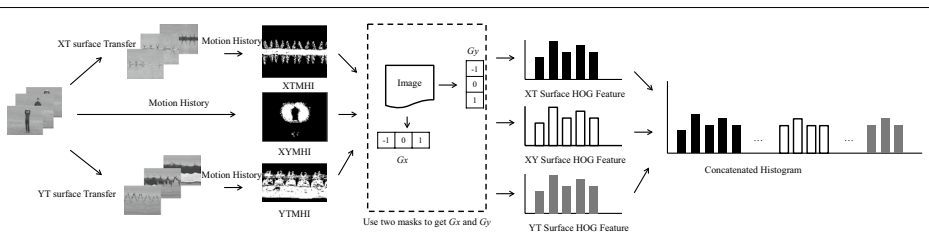

Fig. 3 Framework of the 3SMF. Firstly, video is regarded as three different surface image sequences. The MHI is calculated by frame difference. And HOG feature is detection for MHI

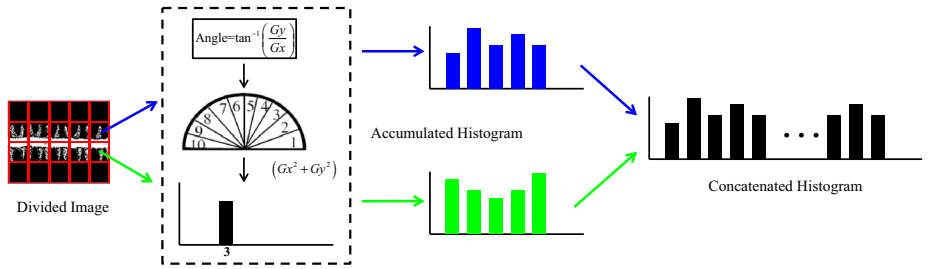

Fig. 4 Framework of the HOG feature. The MHI is divided to $M \times N$ grids. The gradient of pixel casts a weight vote for an orientation-based histogram

contained four steps: divided image into block by rectangle partitioning, calculated image gradient using two masks ([−1, 0, 1] and [1, 0, −1]), accumulated histogram for each blocks, and concatenated block histogram.

Rectangle partitioning is the most common method for representing small spatial regions in an image. An image can be divided into several rectangles of the same size. The ratio of the block size to the image size typically depends on the total number of blocks. In other words, if an image is divided into $M \times N$ blocks, the block size is $h/M \times w/N$, where h and w are the image height and image width, respectively. Traditional block-partition divides an entire image into a grid, in which all blocks are the same size. The block size is crucial because a large block may enclose a contiguous region and produce conspicuous features, whereas a small block cannot adequately represent object characteristics. In this work, both M and N are set to 9.

The most common gradient computation method is to apply a mask in both the horizontal and vertical directions. This study uses two masks to filter the intensity data of an image to obtain the orientation (or angle) of the current pixel.

Each pixel within a block then casts a weighted vote for an orientation-based histogram channel based on the values calculated by gradient computation. The histogram channels are evenly spread over 0°–180° or 0°–360°. In this work, angles of 0°–180° are divided into ten 18° intervals. To increase the tolerance for vertical and horizontal angles, angles of 0°–9° and 171°–180° are set to the same interval; the angles of 81°–99° form a new interval. After partitioning, feature extraction is applied to construct a local feature histogram for each block, which is concatenated to form the image representation. For a consistent measure, each value for bin i, $h(i)$, is normalized to $h'(i)$ within the range of 0–1 by the following equation:

$$h'(i) = \frac{h(i)}{\sum_{i=1}^{n} h(i)} \tag{3}$$

where n is the total number of bins, i.e., ten in this work. So, the HOG feature length of MHI is 810.

Finally, the HOG feature of each block is concatenated to build the 3SMF feature, and the length of the 3SMF feature is 2430.

The 3SMF feature detection algorithm is summarized in Algorithm 1.

Algorithm 1 Three-surface motion feature (3SMF) detection algorithm

Input: Video or Image sequence V_{xy}

Output: Feature Vector F

1. Image sequence transfer using Eq. (2): $V_{xy} \rightarrow V_{xt}, V_{xy} \rightarrow V_{yt}$
2. For each image sequence, calculate the motion history image (MHI) using frame difference method: $V_{xy} \rightarrow XYMHI, V_{xt} \rightarrow XTMHI, V_{yt} \rightarrow YTMHI$
3. For each MHI image I:
   (a) Divided into $M \times N$ blocks
   (b) Calculated the gradient of all pixel in I
   (c) Each pixel within a block casts a weighted vote for an orientation-based histogram: $h(i)$, $i = 1 \ldots M \times N$
   (d) Concatenated the histogram of blocks to represent MHI: $H_I = \{h(1), \ldots, h(i), \ldots, h(M \times N)\}$
4. Concatenated MHI feature to build 3SMF feature: $F = \{H_{xy}, H_{xt}, H_{yt}\}$

Action inference algorithm

To classify the test video V, the class of V is the class c^* that has the maximum probability score between V and a specific class c corresponding to the following equation:

$$c^* = \underset{c \in \{1,\ldots,N_c\}}{\arg\max}\ p(c, V) = \underset{c \in \{1,\ldots,N_c\}}{\arg\max}\ p(c)p(V|c) \tag{4}$$

where N_c is the number of action categories. Given the prior $p(c)$ and posterior $p(V|c)$, we can infer the best c^* by maximizing the joint distribution $p(c, V)$. Here, we train the SVM classifier to inference the prior $p(c)$ using the 3SMF feature. The posterior $p(V|c)$ is solved using the NBNN algorithm.

The SVM classifier is a binary classifier in a high-dimensional hyper plane and it is a decision function in high-dimensional space. For the problem of multiclass classification, one-versus-one strategy is used in SVM model training. We build the binary classifier with RBF kernel for every two actions [total of $\frac{N_c \times (N_c - 1)}{2}$ SVM classifiers]. For testing data, the target is to choose the class that is selected by most classifiers. In the training process, fivefold cross-validation is used to find the best parameters of RBF kernel.

To compute the posterior $p(V|c)$, the video is expressed as the set of STIPs $V = \{d_v | v = 1, \ldots, N_v\}$, where d_v is the STIP feature and N_v is the number of STIPs. The probability $p(V|c)$ is transformed to the probability $p(d_1, \ldots, d_{N_v}|c)$ between the STIP and action category. In Native Bayes algorithm, the joint probability $p(d_1, \ldots, d_{N_v}|c)$ is transformed to the product of each STIP based on independence assumption as following:

$$p\left(d_1, \ldots, d_{N_v}|c\right) = \prod_{N_v} p(d_v|c) \tag{5}$$

And to calculate the probability $p(d_v|c)$, Gauss probability distribution is used based on nearest neighbors.

For a special video, NBNN approximation is used to estimate the probability as follows:

$$
\begin{aligned}
p(V|c) &= \prod_{N_v} p(d_v|c) \\
p(d_v|c) &= \frac{1}{\left|NN_{\varepsilon}^{c_i}(d_v)\right|} \sum_{d_t \in T^{c_i}} K(d_v - d_t) \\
&\approx \frac{1}{\left|NN_{\varepsilon}^{c_i}(d_v)\right|} \exp\left[-\frac{1}{2\sigma^2}\left(\left\|d_v - d_{NN}(d_v)\right\|^2\right)\right] \\
\left\|d_v - d_{NN}(d_v)\right\| &= \min_{d_t \in NN_{\varepsilon}^{c_i}(d_v)} \left\|d_v - d_t\right\|,
\end{aligned}
\tag{6}
$$

where T^{c_i} is the STIP set of the training data with action label c_i. $NN_{\varepsilon}^{c_i}(d_v)$ denotes the set of samples d_t in the training videos T^{c_i}, with distances to d_v of less than ε. Furthermore, $d_{NN}(d_v)$ is the set of nearest neighbors of d_v in the set $NN_{\varepsilon}^{c_i}(d_v)$. Recognition performance is insensitive to the choice of ε. The experimental results of Yuan et al. (2011b) and Zhang et al. (2013) have shown that setting ε to 2.2 yields the highest accuracy. The same conclusion was obtained in this work; thus, we set ε to 2.2.

The proposed action recognition approach is summarized in Algorithm 2.

Algorithm 2: Action recognition through multi-surface analysis

Input: Video or Image sequence V_{xy}

Output: Action category c^*

Training:

1. Detection STIPs for training data: $T = \cup T^{c_i}$
2. Detection 3SMF using Algorithm 1 for training data
3. Using 3SMF feature to train SVM model

Testing:

1. Detection STIPs for testing data $D = \{d_v | v = 1, \ldots, N_V\}$
2. Detection 3SMF feature for testing data
3. For each feature d_v in D, searching nearest neighbors in the STIP set of training data, and calculate the probability $p(V|c)$
4. Using SVM model to calculate the prior probability $p(c)$
5. Inference action by Eq. (4)

Computational complexity

In the initial of the test process, the 3SMF and STIP features are extracted from the input video. The computation of these features need iterate over all of pixels in the video, so the computation complexity of the feature extraction is $N_x \times N_y \times N_t$. The SVM algorithm can calculate the classification probability of test feature in linear time. The intensive computation for NBNN is to search nearest neighbors from training data for all of the STIP features extracted from test video. The computation complexity of searching nearest neighbors depend on the size of training set. In the experiments, the number of STIP features extracted from training set is more than hundreds of millions. The computational complexity of the proposed is combination of feature detection and nearest neighbors searching in training data. However, due to the STIP set of training data is large, and the method of 3SMF detection, unfortunately, the proposed method is not suitable for real-time recognition.

Experimental results and analysis

Action dataset

This section describes the experiments used to verify the effectiveness of the proposed methods, as described in "Algorithms in the proposed method" section. All experimental results are obtained using the KTH dataset and UCF sport dataset (Rodriguez et al. 2008). Figure 5 shows some examples from the dataset.

The KTH dataset contains six (K = 6) actions (i.e., walking, jogging, running, boxing, hand-waving, and handclapping). Each action has 25 subjects in four environments (i.e., outdoors, outdoors with variable scales, outdoors with different clothes, and indoors with lighting conditions). Subjects are selected randomly, and their corresponding actions are collected as a training dataset; the remaining videos are used as the dataset test. In our experiment, we used 25-fold leave-one-out cross-validation to measure the performance of the proposed method.

The UCF Sports Action dataset consists of ten different types of sports actions (A = 10), i.e., 'swing-bench', 'swing-side', 'diving', 'kicking', 'lifting', 'riding horse', 'running', 'skateboarding', 'golf swing', and 'walking'. The dataset consists of 150 real videos. A horizontally flipped version of each video sequence was added to the dataset to increase the number of training samples. In our experiment, we used the leave-one-out strategy

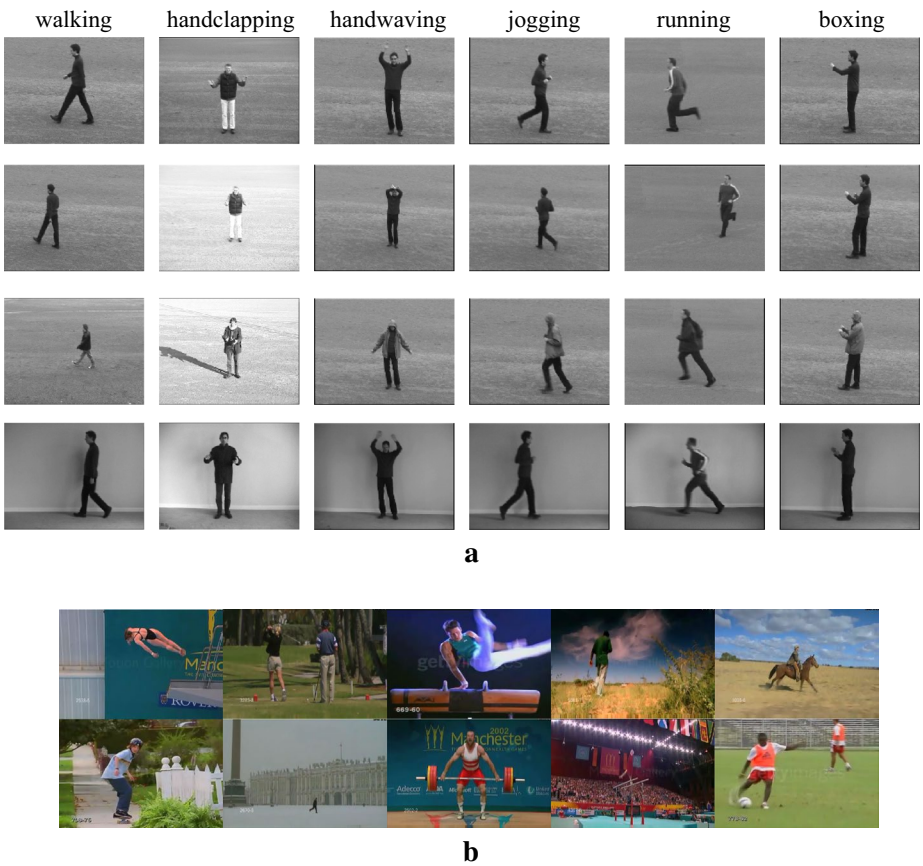

Fig. 5 Examples of action datasets. **a** KTH dataset, **b** UCF sport dataset

to test each original action sequence, whereas the remaining original video sequences, together with their flipped versions, were included in the training set.

In the experiment, certain parameters affect the accuracy of action recognition. The relevant parameters were set as follows:

1. The parameters in the STIP detection are the same as those used by Laptev (2005b).
2. To compute the 3SMF feature, the size of the MHI image of video is the same as the image in the video. The size of the MHI image of the XT surface image sequence is $N_x \times N_t$. The size of the MHI image of the YT surface image sequence is $N_y \times N_t$. Figure 6 shows the MHI of KTH actions.
3. Due to the difference of the length of the video, the sizes of the XTMHI and YTMHI are different, as shown in Fig. 7. The general approach to normalize the XTMHI is image scaling. However, for certain similar actions, image scaling may eliminate the classified information, such as walking, running and jogging. In our study, the length of the video is cut to a fixed length $N_c = 200$. The same setting is used for YTMHI.

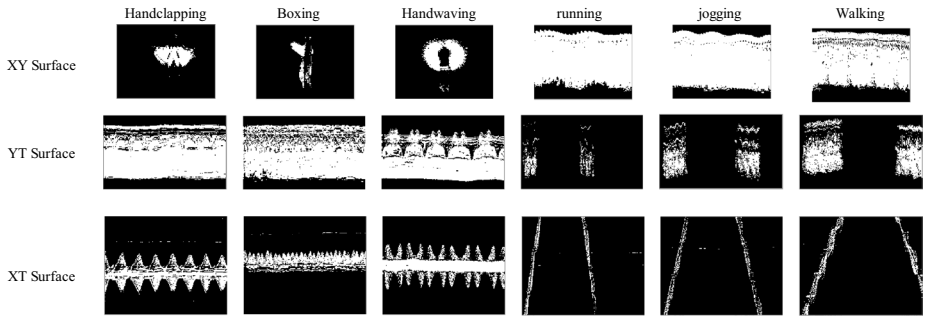

Fig. 6 Motion history image of six actions (handclapping, boxing, hand-waving, running, jogging and walking) from three different surfaces. Every MHI belonged to different actions has its own appearance, as shown in *each row*. The same as the above, the MHI of specific action has different appearance in the different surfaces

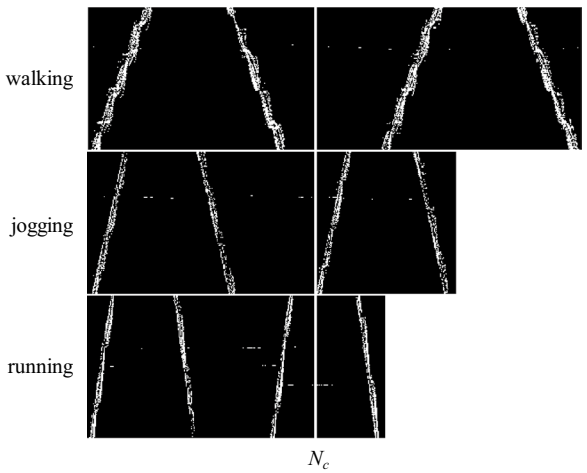

Fig. 7 MHI image of the XT surface image sequence (XTMHI)

The other parameters for detecting 3SMF are set as stated in "Action inference algorithm" section.

Performance evaluation of action recognition

The experimental results of the KTH dataset are shown in Table 1. The comparison results illustrate that the proposed method is effective for human action recognition. The recognition accuracy of the proposed method was the highest among all relevant methods. 3SMF improved the accuracy of action recognition by more than 3 % compared to the approach using STIP and the NBNN algorithm by Zhang et al. (2013).

Comparing with the approach using STIP and the NBNN algorithm in Table 1, we can find it is effective for action recognition by using the 3SMF feature with SVM model to estimate the prior probability instead of uniform distribution. And comparing with other methods on KTH dataset, our approach have the best performance. These results can verify the effective of the proposed method. The confusion matrix of the proposed method is shown in Table 2.

In confusion matrix, each column represents the instances in a predict class while each row represents the instances with ground truth. Confusion matrix summarize the classification results of test samples. For example, if there are N_i test samples with action c_i, N_i is the number of predict the test samples to action c_j (in KTH dataset, $i = 6$, $N_i = \sum_{j=1}^{6} N_{ij}$). The value of confusion matrix in first row can been computed as following:

$$value_{ij} = \frac{N_i}{N_{ij}} \tag{7}$$

Next, the proposed method was applied to the UCF sport dataset to verify the effectiveness of the proposed method in practical use. Table 3 compares the proposed

Table 1 Comparison of the proposed method with existing methods for the KTH dataset

Method	Accuracy (%)
3SMF + STIP + NBNN	*96.50*
STIP + NBNN algorithm (uniform distribution) (Zhang et al. 2013)	92.83
Yuan et al. (2011a)	94.00
Yan and Luo (2012)	93.98
Chakraborty et al. (2012)	96.35
Wang et al. (2014)	94.4
Weinland et al. (2010)	92.4

Italic value mean the best results

Table 2 Confusion matrix of the proposed method on the KTH dataset

	Walking	Running	Jogging	Handwaving	Handclapping	Boxing
Walking	.98	.01	.01	.00	.00	.00
Running	.00	.86	.14	.00	.00	.00
Jogging	.00	.02	.98	.00	.00	.00
Handwaving	.00	.00	.00	.99	.00	.01
Handclapping	.00	.00	.00	.02	.98	.00
Boxing	.00	.00	.00	.00	.00	1.0

Table 3 Comparison of the proposed method with existing methods for the UCF sports dataset

Methods	Accuracy (%)
3SMF + STIP + NBNN	*94.39*
Wang et al. (2009)	85.60
Yan and Luo (2012)	90.67
Le et al. (2011)	86.50
Shao et al. (2014)	93.4
Zhang et al. (2015)	88.0

Italic value mean the best results

method to the existing methods. Based on the results, the proposed method has the best accuracy of 94.39 %. Comparing with the best accuracies of the state-of-art methods for the UCF dataset, the improvement of the proposed method is .99 %. Table 4 shows the confusion matrix of the proposed method on the UCF sport dataset.

Based on these performance evaluation, the accuracy for KTH dataset is close to other methods. For the UCF dataset, the differences are higher. There are two main reasons. Firstly, the proposed 3SMF feature is a holistic feature to representation video. The background of KTH dataset is simple, monotonous and uniformity. And from the confusion matrix in Table 3, we find the classification error occurred mainly in "running" category. This action is very similar with "jogging". Different with KTH dataset, the special background of UCF dataset is related to respective action category. So the discriminative power of 3SMF for KTH dataset is weaker compared with UCF dataset. Therefore, the improvement of our method of UCF dataset is better than KTH dataset. On the other hand, the accuracy of the existing algorithm for KTH dataset has exceeded 96 %, while only 94 % for UCF dataset. Further improvement has greater challenge in the case of higher accuracy.

Conclusion

In this paper, we propose a novel multi-surface feature named 3SMF. The prior probability is estimated by an SVM, and the posterior probability is computed by the NBNN algorithm with STIP. We model the relationship score between each video and action as

Table 4 Confusion matrix of the proposed method on the UCF sport dataset

	Diving	Golf	High-swinging	Kicking	Lifting	Riding	Running	Skating	Swing	Walking
Diving	1.0	.00	.00	.00	.00	.00	.00	.00	.00	.00
Golf	.00	.90	.00	.00	.00	.04	.00	.00	.00	.06
High-swinging	.00	.00	.89	.00	.00	.02	.00	.00	.09	.00
Kicking	.00	.00	.00	1.0	.00	.00	.00	.00	.00	.00
Lifting	.00	.00	.00	.00	1.0	.00	.00	.00	.00	.00
Riding	.00	.00	.00	.00	.00	1.0	.00	.00	.00	.00
Running	.00	.00	.00	.01	.00	.00	.93	.00	.00	.06
Skating	.00	.00	.00	.00	.00	.00	.09	.86	.05	.00
Swing	.00	.00	.00	.00	.00	.00	.05	.00	.95	.00
Walking	.00	.04	.00	.00	.00	.05	.00	.00	.00	.91

a probability inference to bridge the feature descriptors and action categories. The main contributions of our study is that a new holistic feature (3SMF) is proposed to represent video. 3SMF can reflect the difference of the action in different surfaces. The results of the comparisons with the state-of-the-art action recognition benchmarks demonstrate the effectiveness of the proposed method. However, it also has some limitations in this study. Due to the computation complexity of feature detection and NBNN algorithm, the proposed method is not suitable for real-time recognition. And our method works only in the case that the videos contain one action category. Therefore, in future, we will address the following topics: real-time action recognition and multiple action events recognition in videos.

Authors' contributions

HZ and QL carried out three-surface motion feature and drafted the manuscript. BZ carried out the STIP feature. JD carried out the SVM. HZ and JP carried out NBNN algorithm. TH and DC assisted with experimental results analysis. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank the anonymous reviewers for the valuable and insightful comments on the earlier version of this manuscript. This work is supported by the Natural Science Foundation of China (Nos. 61502182, 61175121, 11401231, and 61572205), the Natural Science Foundation of Fujian Province of China (Nos. 2015J01253, 2015J01257, and 2013J06014), the Pilot Project of Fujian Province of China (No. 2015H0025) and the Promotion Program for Young and Middle-aged Teachers in Science and Technology Research of Huaqiao University (No. ZQN-YX108).

Competing interests

The authors declare that they have no competing interests.

Received: 22 March 2016 Accepted: 19 July 2016
Published online: 02 August 2016

References

Aggarwal JK, Ryoo MS (2011) Human activity analysis: a review. ACM Comput Surv 43(3):194–218

Ahad MA, Tan JK, Kim H, Ishikawa S (2012) Motion history image: its variants and applications. Mach Vis Appl 23(2):255–281

Burghouts GJ, Schutte K, ten Hove RJM, van den Broek SP, Baan J, Rajadell O, van Huis JR, van Rest J, Hanckmann P, Bouma H, Sanroma G, Evans M, Ferryman J (2014) Instantaneous threat detection based on a semantic representation of activities, zones and trajectories. Signal Image Video Process 8(1):191–200

Chakraborty B, Holte MB, Moeslund TB, Gonzalez J (2012) Selective spatio-temporal interest points. Comput Vis Image Underst 116(3):396–410

Charalampous K, Gasteratos A (2014) On-line deep learning method for action recognition. Form Pattern Anal Appl 19(2):337–354

Dawn D, Shaikh S (2015) A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. Vis Comput 32(3):289–306

Gao Z, Zhang H, Xu GP, Xue YB, Hauptmann AG (2015) Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition. Signal Process 112:83–97

Ikizler-Cinbis N, Sclaroff S (2010) Object, scene and actions: Combining multiple features for human action recognition. In: 11th European conference on computer vision (ECCV 2010), September 5–11. Springer, Heraklion, Crete, Greece

Junejo IN, Dexter E, Laptev I, Perez P (2011) View-independent action recognition from temporal self-similarities. IEEE Trans Pattern Anal Mach Intell 33(1):172–185

KTH. http://www.nada.kth.se/cvap/actions/. Retrieved on date 07/2016

Laptev I (2005a) On space-time interest points. Int J Comput Vis 64(2-3):107–123

Laptev I (2005b) On space-time interest points. Int J Comput Vis 64(2-3):107–123

Laptev I, Lindeberg T (2006) Local descriptors for spatio-temporal recognition. Spat Coherence Vis Motion Anal 3667:91–103

Le QV, Zou WY, Yeung SY, Ng AY (2011) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: IEEE conference on computer vision and pattern recognition (CVPR 2011), June 20–25. IEEE Computer Society, Colorado Springs, CO, USA

Liu AA, Xu N, Su YT, Lin H, Hao T, Yang ZX (2015) Single/multi-view human action recognition via regularized multi-task learning. Neurocomputing 151:544–553

Luo JJ, Wang W, Qi HR (2013) Feature extraction and representation for distributed multi-view human action recognition. IEEE J Emerg Sel Top Circuits Syst 3(2):145–154

Marszalek M, Laptev I, Schmid C (2009) Actions in context. In: IEEE computer society conference on computer vision and pattern recognition workshops, CVPR workshops, June 20–25. IEEE Computer Society, Miami, FL, USA

Nissi Paul S, Jayanta Y (2016) Tri-level unified framework for human gait analysis. ADBU J Eng Technol 4:28–41
Paul SN, Singh YJ (2014) Survey on video analysis of human walking motion. Int J Signal Process Image Process Pattern Recognit 7:99–122
Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28(6):976–990
Rodriguez MD, Ahmed J, Shah M (2008) Action MACH: a spatio-temporal maximum average correlation height filter for action recognition. In: 26th IEEE conference on computer vision and pattern recognition (CVPR), June 23–28, 2008. Institute of Elec. and Elec. Eng. Computer Society, Anchorage, AK, USA
Shao L, Zhen XT, Tao DC, Li XL (2014) Spatio-temporal Laplacian pyramid coding for action recognition. IEEE Trans Cybern 44(6):817–827
Wang H, Ullah MM, Klaser A, Laptev I, Schmid C (2009) Evaluation of local spatio-temporal features for action recognition. In: 20th British machine vision conference (BMVC 2009), September 7–10. British Machine Vision Association (BMVA), London, UK
Wang T, Wang S, Ding X (2014) Detecting human action as the spatio-temporal tube of maximum mutual information. IEEE Trans Circuits Syst Video Technol 24(2):277–290
Weinland D, Özuysal M, Fua P (2010) Making action recognition robust to occlusions and viewpoint changes. In: European conference on computer vision conference on computer vision
Wu X, Jia Y (2012) View-invariant action recognition using latent kernelized structural SVM. In: Fitzgibbon A et al (eds) Computer vision—ECCV 2012. Springer, Berlin, pp 411–424
Yan XS, Luo YP (2012) Recognizing human actions using a new descriptor based on spatial–temporal interest points and weighted-output classifier. Neurocomputing 87:51–61
Yao B, Fei-Fei L (2012) Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. IEEE Trans Pattern Anal Mach Intell 34(9):1691–1703
Yoon H, Kim KS, Kim D, Bresler Y, Ye JC (2014) Motion adaptive patch-based low-rank approach for compressed sensing cardiac cine MRI. IEEE Trans Med Imaging 33(11):2069–2085
Yu G, Yuan J, Liu Z (2012) Predicting human activities using spatio-temporal structure of interest points. In: Proceedings of the 20th ACM international conference on multimedia, ACM, Nara, Japan, pp 1049–1052
Yuan JS, Liu ZC, Wu Y (2011a) Discriminative video pattern search for efficient action detection. IEEE Trans Pattern Anal Mach Intell 33(9):1728–1743
Yuan JS, Liu ZC, Wu Y (2011b) Discriminative video pattern search for efficient action detection. IEEE Trans Pattern Anal 33(9):1728–1743
Zhang HB, Li SZ, Su SZ, Chen SY (2013) Selecting effective and discriminative spatio-temporal interest points for recognizing human action. IEICE Trans Inf Syst 96(8):1783–1792
Zhang Z, Liu S, Liu S, Han L, Shao Y, Zhou W (2015) Human action recognition using salient region detection in complex scenes. In: Mu J et al (eds) Proceedings of the third international conference on communications, signal processing, and systems. Springer, New York, pp 565–572
Zhou Z, Shi F, Wu W (2015) Learning spatial and temporal extents of human actions for action detection. IEEE Trans Multimed 17(4):512–525