Yao et al. EURASIP Journal on Image and Video Processing (2016) 2016:39
DOI 10.1186/s13640-016-0145-2

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Spatio-temporal information for human action recognition

Li Yao[1,2*], Yunjian Liu[3] and Shihui Huang[3]

Abstract

Human activity recognition in videos is important for content-based videos indexing, intelligent monitoring, human-machine interaction, and virtual reality. This paper uses the low-level feature-based framework for human activity recognition which includes feature extraction and descriptor computing, early multi-feature fusion, video representation, and classification. This paper improves the first two steps. We propose a spatio-temporal bigraph-based multi-feature fusion algorithm to capture the useful visual information for recognition. Meanwhile, we introduce a compressed spatio-temporal video representation to bag of words representation. Our experiments on two popular datasets show efficient performance.

Keywords: Spatio-temporal, Video representation, Multi-feature fusion, Human action recognition

1 Introduction

Automatic recognition of human actions in video automatically is a promising technology in computer vision. Application scenarios include content-based video retrieval, intelligent video surveillance, and human-computer interaction. Although many researchers have done a long-term study in this work, it remains challenging to recognize human actions in videos not only because of geometric variations between intra-class objects or actions, but also because of changes in scale, rotation, viewpoint, illumination, and occlusion [1].

In general, one of the most popular frameworks for human action recognition includes four steps: feature extraction, video representation, multi-feature fusion, and classification. In this paper, we mainly focus on improving two steps: video representation and multi-feature fusion.

BoW (bag of words) is one of the most popular methods for video representation. Much research is based on the classical BoW representation [2–5]. The classical BoW representation firstly clusters the features to several *visual vocabulary* (e.g., the KMEANS method), then encodes a video clip to the histogram of its features occurrences. The BoW model has shown good generalization capability and robustness on many works [3–5]. However, BoW has many drawbacks, such as the time-consuming clustering procedure, the supervised parameter k for KMEANS, and the well-known limitation of losing spatial and temporal cues for recognition. To make up the lack of spatio-temporal information of BoW, many researchers have proposed several extension of BoW representation [2, 3, 6–9]. But these extensions are too complicated and time-consuming for large-scale dataset, or reduce the time complexity with dropped recognition accuracy [3, 8, 9]. To reduce the computational cost with nearly no accuracy lost, we propose a simple spatio-temporal visual information retained representation for videos. We capture the spatio-temporal information between visual words by the spatio-temporal distance between features, and we compress the spatio-temporal cue to a compact representation.

As single feature cannot contain all the useful information for human action recognition, the researchers usually combine multiple features for better accuracy. Under the assumption that different features are independent, we can simply connect vectors of different features to a new vector. However, different features are not always independent. Researches have proposed several approaches to make further use of the different information in videos. Jiang et al. [10] introduced an audio-visual atom as joint audio-visual feature for video

* Correspondence: Yao.li@seu.edu.cn
[1]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing, China
[2]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
Full list of author information is available at the end of the article

concept recognition. Jiang and Loui [11] used the temporal relationship between audio feature and visual feature to group clusters up, and then construct new features from the groups. Fernando et al. [12] captured video-wide temporal information for action recognition. Jhou et al. [13] proposed to use the concurrent statistical information to construct a bipartite graph for feature fusion. In fact, these methods use the temporal relationship between audio feature and visual feature for early fusion. However, when combining two visual features, the spatial relationships between different features are ignored. In this paper, we improved Jhou et al.s' method by using the spatio-temporal relationship between two different visual features explicitly for early fusion.

Our contributions are as follows: (1) We proposed a bigraph multi-feature fusion method to model the spatio-temporal cue between visual words. (2) We proposed a spatio-temporal visual information retained representation method for the classical BoW video representation to reduce the computational cost with nearly no accuracy lost.

2 Related works

As mentioned in section 1, the classical BoW method ignores the spatio-temporal relationship between visual features. Several works have been proposed to capture spatio-temporal information to improve BoW. In this section, we divide these methods to two types: the absolute spatio-temporal information retaining method, and the relative spatio-temporal information retaining method.

1) The former [2, 6, 14–16] typically needs a global partition for the spatio-temporal volume which makes the representation sensitive to the absolute coordinate shift. Laptev et al. [2] split the spatio-temporal volume to grids, computed the histogram of visual word occurrence over each grid, and concatenated BoW vectors from different grids. In this way, each video was represented by the spatio-temporal information capturing BoW vector. However, it needed to figure out the best grid combination by cross-validation, which is time-consuming. The concatenated long vector made it even worse. Sun et al. [6] modeled the spatio-temporal context information in a hierarchical way which included three levels of context.
2) The relative spatio-temporal information retaining methods [3, 7, 9, 17] typically captures the relative distance between visual words and local features. Grauman et al. [7] formed new features composed of the neighborhoods around the raw initially detected interest points, taking into account the visual words to which the neighboring features correspond and their orientation with respect to the central interest point. However, this work built a hierarchy visual word which is complex and time-consuming. Wang et al. [3] exploited the contextual interactions between interest points by the density of all features observed in multi-scale spatio-temporal contextual domain of each interest point. And Zhou et al. [9] proposed a novel structured codebook construction method to encode rich spatial and temporal contextual information for human action recognition.

Although these researches have achieved some performance improvement in their experiments, the procedures they detailed are relatively complex. Rather than using the several predefined grids, such as the hierarchy information or the multiple spatio-temporal scales as some of these works did, a simple method is explored to model the spatio-temporal cue between visual words in our work.

To recognize complex human action, frameworks aligned on single feature are usually not good enough. Researchers have proposed several features to extract different information of videos, such as the dense trajectory feature, STIP, SIFT, and so on. Moreover, there are researchers trying to combine multiple features by well-designed models for feature fusion. Natarajan et al. [18] extracted dozens of features from videos, which is SIFT, SURF, D-SIFT, CHOG, and so on. And they took advantage of multiple kernel learning and late fusion technology to combine these features. Tang et al. [19] used two basic operators which is *and operator* and *or operator* to combine two feature vectors. *And operator* simply connects two vectors while *or operator* chooses one vector from two vectors as the combined vector. They tried to construct a *and-or tree* to compute the best combination for two features. However, as this method requires searching for the best structure of *and-or tree*, the time complexity of this method is too high.

Meanwhile, some people tried to design specific model to combine visual feature and audio feature. Jiang et al. [10] grouped visual and audio features together with their temporal relationship and computed combined features from these groups. Similarly, Jhou et al. [13] constructed a bigraph with temporal concurrency between visual words and employed a k-way segmentation algorithm to combine visual and audio features. In this work, we propose to construct a spatio-temporal bigraph and use the k-way segmentation algorithm to combine multiple features.

3 Approach

As Fig. 1 shows, we extract dense trajectory features from the videos and encode each feature to three different descriptors which are HOG, HOF, and MBH.

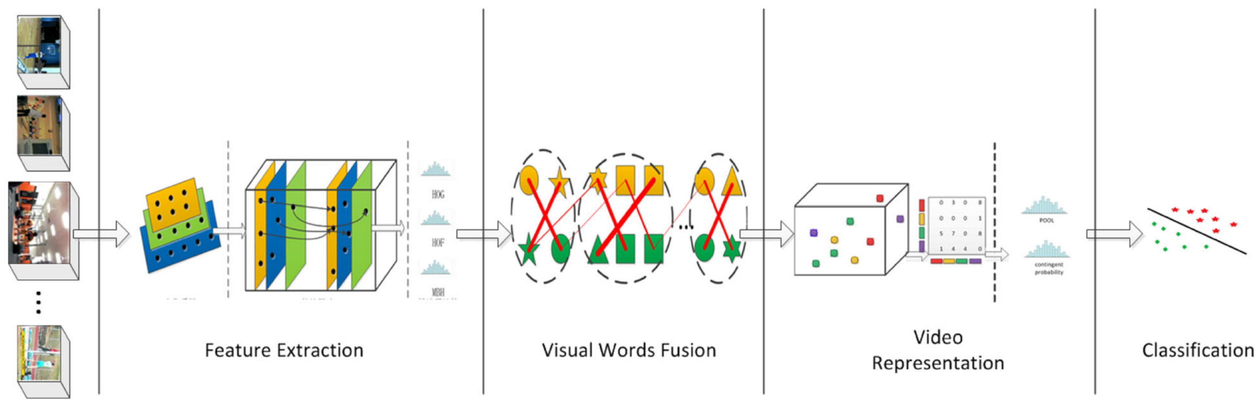

Fig. 1 The flowchart of this paper

Secondly, features are sampled and clustered into *k*-visual words. Then, we construct a spatio-temporal bigraph and employ an efficient *k*-way segmentation algorithm to segment the graph. Visual words with strong spatio-temporal relationship are fused while visual words with weak spatio-temporal relationship are segmented. Moreover, to further capture the spatio-temporal information between features, each video is represented with the algorithm detailed in section 3.3. Finally, we use a support vector machine (SVM) classifier for recognition.

3.1 Feature extraction

We employ the dense trajectory features. As dense sampling has shown improving results over sparse interest points, the dense trajectory firstly samples the points in different spatial scales densely, then performs tracking in a dense optical flow field. Finally, three different descriptors, namely, HOG, HOF, and MBH, are calculated along each trajectory. Different descriptors contain different information of features. HOG uses the distribution of grayscale images' gradient directions to describe the appearance and shape of objects in 3D world. HOF and MBH use the optical flow, so the motion information is captured. As a result, we can apply them as different information sources for early feature fusion.

3.2 Spatio-temporal bigraph-based feature fusion

We first sample a subset from each feature, and then cluster features to *k*-visual words by KMEANS algorithm. After that, we construct a spatio-temporal bigraph in which node is represented by visual words, and edge stands for the spatio-temporal relationship between visual words. Then, we employ a *k*-way segmentation algorithm to segment this bigraph. By this way, visual words with strong spatio-temporal relationship are fused while visual words with weak spatio-temporal relationship are segmented. Finally, a spatio-temporal information-based video representation is used to encode videos.

3.2.1 Spatio-temporal bigraph

Let's say two features named fea1 and fea2, and the visual words of these two features are $\left\{W_i^{fea1} \middle| 1 \le i \le k^{fea1}\right\}$ and $\left\{W_i^{fea2} \middle| 1 \le i \le k^{fea2}\right\}$, k^{fea1} and k^{fea2} stand for the amount of visual words in fea1 and fea2. We can construct a bigraph $G = (V, E)$ for these visual words. V is $\left\{W_i^{\mathrm{fea1}} \middle| 1 \le i \le k^{\mathrm{fea1}}\right\} \cup \left\{W_i^{\mathrm{fea2}} \middle| 1 \le i \le k^{\mathrm{fea2}}\right\}$, and E is the adjacency matrix of this bigraph:

$$E = \begin{bmatrix} 0 & S \\ S^{\mathrm{T}} & 0 \end{bmatrix} \tag{1}$$

where S is

$$S(i,j) = \sum\nolimits_V \mathrm{DM}_V(i,j) \quad = \quad \sum\nolimits_V \sum\nolimits_{p \in W_i^{\mathrm{fea1}},\, q \in W_j^{\mathrm{fea2}}} (d(p,q)) \tag{2}$$

and p, q are two feature descriptors from two features, $d(p, q)$ is the $L1$ distance of their spatio-temporal coordinates.

3.2.2 K-way segmentation algorithm

Given a bipartite graph $G = (V, E)$, a bipartitioning is to partition V to two subsets such that vertices in the same subset have strong relation, and vertices in different subset have weak relation. Formally, a graph bipartitioning aims to minimize the following objective function:

$$\mathrm{cut}(V_1, V_2) = \sum\nolimits_{i \in V_1, j \in V_2} s_{ij} \tag{3}$$

Actually, finding a bipartitioning of bigraph can be understood as classifying each point into two classes, e.g., +1 and −1. However, this may lead to a wrong solution that assigns all vertices to +1 or −1. In this paper, we are looking for a balanced partition whose objective function looks like below.

$$\mathrm{BalanceCut}(V_1, V_2) = \frac{\mathrm{cut}(V_1, V_2)}{\sum_{i\in V_1}\sum_j e_{ij}} + \frac{\mathrm{cut}(V_1, V_2)}{\sum_{i\in V_2}\sum_j e_{ij}} \tag{4}$$

This problem can be solved by spectral clustering, which firstly constructs a Laplace matrix L as below.

$$L(i,j) = \begin{cases} -e_{ij} & e_{ij}\in E \\ \sum_k e_{ik} & i=j \\ 0 & \text{else} \end{cases} \tag{5}$$

After that, a bipartitioning of G can be provided by the second smallest eigenvector of the generalized eigenvalue problem $Lz = \lambda Dz$ in which $D(i,j) = \Sigma_j e_{ij}$.

However, as an efficient solution proposed by Dhillon et al. [20], we can get an optimal bipartitioning with low computational complexity. Suppose we have a matrix L in which $D_1^{\mathrm{fea1}} = \sum_j e_{ij}$ and $D_2^{\mathrm{fea1}} = \sum_j e_{ij}$, as below:

$$\begin{aligned} L(i,\ j) &= \begin{bmatrix} D_1^{\mathrm{fea1}} & -S \\ -S^T & D_2^{\mathrm{fea2}} \end{bmatrix} \\ &= \begin{bmatrix} D_1^{\mathrm{fea1}} & 0 \\ 0 & D_2^{\mathrm{fea2}} \end{bmatrix} + \begin{bmatrix} 0 & -S \\ -S^T & 0 \end{bmatrix} \end{aligned} \tag{6}$$

Let $\mathbb{S} = D_1^{\mathrm{fea1}^{-1/2}} S D_2^{\mathrm{fea2}^{-1/2}}$, it can be proved that the second eigenvector of L can be expressed in terms of left and right singular vectors (say u_2 and v_2) of $\mathbb{S}$ as follows:

$$z_2 = \begin{bmatrix} D_1^{\mathrm{fea1}} - \frac{1}{2} u_2 \\ D_2^{\mathrm{fea2}} - \frac{1}{2} v_2 \end{bmatrix} \tag{7}$$

In a more general case, suppose we need to capture k new words containing relational information, the optimal k-way partitioning solution is provided by the $l = \lceil \log k \rceil$ singular vectors $U = (u_2, \ldots, u_{l+1})$ and $V = (v_2, \ldots, v_{l+1})$.

To be specific, let $\mathbb{Z} = \left[D_1^{\mathrm{fea1}^{-1/2}}\mathbb{U}, D_2^{fea2^{-1/2}}\mathbb{V}\right]^T$, we look for k clusters of row space in Z such that the sum of squares $\sum_{i=1}^k \sum_j \mathrm{distance}(i,j)$ is minimized.

Thus our bimodel-based clustering algorithm can be summarized as five basic steps as below:

1) Construct bipartite graph where each element of S is computed as:

$$S(i,j) = \sum_V \mathrm{DM}_V(i,j) \quad = \quad \sum_V \sum_{p\in W_i^{\mathrm{fea1}}, q\in W_j^{\mathrm{fea2}}} (d(p,q)).$$

2) Compute matrix D_1^{fea1}, D_2^{fea2}, and $\mathbb{S} = D_1^{\mathrm{fea1}^{-1/2}} S D_2^{\mathrm{fea2}^{-1/2}}$.
3) Apply SVD on $\mathbb{S}$, and compute U, V.
4) Compute matrix Z whose size is $(k^{fea1} + k^{fea2}) \times l$
5) Run k-means on matrix Z's row vectors to get k clusters.

With the k new clusters, each video can be further represented using a spatio-temporal information retaining representation which will be described in detail in section 3.3.

3.3 Spatio-temporal information-based video representation

Figure 2 is the flowchart of our spatio-temporal information retaining video representation. We first compute the distance matrix of visual words. After that, two different strategies are used to compress this matrix.

3.3.1 Distance matrix

Let's say video V has n feature points, then each video V can be represented as $V = (\langle x_1, y_1, t_1, b_1\rangle, \cdots, \langle x_n, y_n, t_n, b_n\rangle)$ where (x, y, t) are the spatio-temporal coordinates of a feature extracted from V, and b_i is the combined visual word this feature belongs to. We can use b_i to link the combined visual words' spatio-temporal information with the features.

Suppose DM_V is the distance matrix where each element represents the spatio-temporal distance between two combined visual words. Then, $\mathrm{DM}_V(i,j) = \Sigma_p\Sigma_q(d(p,q))$ where p, q are two feature descriptors from two features, $d(p, q)$ is the $L1$ distance of their spatio-temporal coordinates.

The $L1$ distance is

$$d(p,q) = |p.x - q.x| + |p.y - q.y| + |p.t - q.t| \tag{8}$$

3.3.2 Matrix compress

As the original distance matrix is too large to be applied to classifier (e.g., 500 visual words result in a 250000 dimension vector), we need to compress this matrix. In

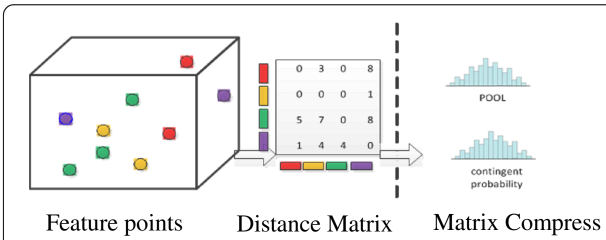

Fig. 2 Proposed spatio-temporal visual information retained video representation

this paper, we compare two different compress strategies with experiments which is *POOL compress* [21], *contingent probability-based representation* [22]. As experiment 2 shows, different compress strategies have different performances on different datasets.

3.3.2.1 POOL compress We compute the spatio-temporal distance from the i-th combined visual word of fea1 to all the other words of fea2 by

$$\mathcal{D}_i^{\text{fea1}} = \frac{\sum_j \text{DM}_v(i,j)}{\sum_i \sum_j \text{DM}_v(i,j)} \tag{9}$$

Symmetrically, we also compute the spatio-temporal distance from the i-th combined visual words of fea2 to all the other words of fea1 by

$$\mathcal{D}_j^{\text{fea2}} = \frac{\sum_i \text{DM}_v(i,j)}{\sum_i \sum_j \text{DM}_v(v,j)} \tag{10}$$

Then, each video can be represented as:

$$(x) = \begin{pmatrix} \mathcal{D}_1^{\text{fea1}}, \cdots, \mathcal{D}_k^{\text{fea1}}, \\ \mathcal{D}_1^{\text{fea2}}, \cdots, \mathcal{D}_k^{\text{fea2}} \end{pmatrix} \tag{11}$$

3.3.2.2 Contingent probability-based representation We discretize each value in DM_V to m sub-regions which are $\mathcal{L}_1, \cdots, \mathcal{L}_m$. Then the contingent probability that fea1's combined visual word W_i^{fea1} related to all the combined visual words of fea2 is

$$P\left(\mathcal{L}_s, : |W_i^{\text{fea1}}\right) = \frac{\text{His}\left(\text{W}_i^{\text{fea1}}, \mathcal{L}_s\right)}{\sum_s \text{His}\left(\text{W}_i^{\text{fea1}}, \mathcal{L}_s\right)} \tag{12}$$

where $\text{His}\left(W_i^{\text{fea1}}, \mathcal{L}_s\right)$ represents the frequency that combined visual word W_i^{fea1} is apart from all the fea1's combined visual words with $\mathcal{L}_s$.

Symmetrically, the contingent probability that fea2's combined visual word W_j^{fea2} related to all the combined visual words of fea1 is

$$P\left(\mathcal{L}_s, : |W_j^{\text{fea2}}\right) = \frac{\text{His}\left(W_j^{\text{fea2}}, \mathcal{L}_s\right)}{\sum_s \text{His}\left(W_j^{\text{fea2}}, \mathcal{L}_s\right)} \tag{13}$$

Then, video V can be represented as

$$(x) = \begin{pmatrix} P\left(\mathcal{L}_1, : |W_1^{\text{fea1}}\right) & \cdots & P\left(\mathcal{L}_m, : |W_k^{\text{fea1}}\right) \\ \vdots & \ddots & \vdots \\ P\left(\mathcal{L}_1, : |W_1^{\text{fea1}}\right) & \cdots & P\left(\mathcal{L}_m, : |W_k^{\text{fea1}}\right) \\ P\left(\mathcal{L}_1, : |W_1^{\text{fea2}}\right) & \cdots & P\left(\mathcal{L}_m, : |W_k^{\text{fea2}}\right) \\ \vdots & \ddots & \vdots \\ P\left(\mathcal{L}_1, : |W_1^{\text{fea2}}\right) & \cdots & P\left(\mathcal{L}_m, : |W_k^{\text{fea2}}\right) \end{pmatrix} \tag{14}$$

3.4 Classification

We use a multi-class non-linear support vector machine (SVM for short) [23–25] to classify videos. In general, multi-class SVM is built from two-class SVM. In this paper, a one-versus-one manner is used. Suppose there are N classes, we train $N(N\text{-}1)/2$ different two-class SVM classifiers on every possible pairs of classes, and then each test video is classified to the class that most classifiers vote this video to. Moreover, we use a widely used kernel function with a χ^2 distance function [2, 3, 7], which is

$$K(x,y) = \frac{1}{2} e^{-\gamma \chi^2(x,y)} \tag{15}$$

where $\chi^2(x,y) = \sum_i (x_i - y_i)^2/(x_i + y_i)$ and x, y are two videos' vector representations. The parameter γ is determined by cross-validation.

4 Results and discussion

4.1 Dataset and setup

In our experiments, we use two popular human action datasets which are KTH dataset [26] and Olympic dataset [4].

The KTH dataset consists of six human action classes. Each action class is performed by 25 people. And every person repeats one action four times under different scenarios. Figure 3 is some screenshots from this dataset. We follow the dataset partition as Schuldt et al. [26] did, which is widely used. This partition makes it possible to compare our results with other researchers' works directly.

The Olympic dataset is crawled from YouTube. There are 11 different actions. As these videos are shot under nearly no artificial constraints, there are many camera motions and noises in the videos. Figure 4 is some screenshots from this dataset.

As most of the research works [2, 27–30], we use the mean average precision (MAP) to measure our performance.

4.2 Experiment 1: spatio-temporal bigraph-based feature fusion

We evaluate proposed spatio-temporal bigraph-based feature fusion algorithm (STBi-fusion in Fig. 5). We first extract dense trajectory features from videos, and

Fig. 3 Screenshots of KTH dataset

then MBH and HOF descriptors are computed along each trajectory. After that, proposed spatio-temporal bigraph-based feature fusion algorithm is employed to combine the MBH and HOF feature's visual words. Finally, BoW model is used to compute the combined video representation. We compare proposed spatio-temporal feature fusion method with a widely used baseline algorithm, which combines two feature vectors by simply connecting two vectors. Moreover, the accuracy is also reported when single MBH or HOF feature is used.

As Fig. 5 shows, our proposed method can better take advantage of the useful information among MBH and HOF features, and get higher accuracy than the other methods.

4.3 Experiment 2: influence of contingent probability-based representation's parameter

We compare the results of contingent probability-based video representation's different parameter values on the KTH dataset as Fig. 6 shows. Abscissa refers to parameter m's change, and ordinate refers to accuracy.

On the one hand, we can see that MBH descriptor is always better than HOF descriptor while the HOG is worse. This is because HOG only capture the static information while ignoring the motion information in videos. Although the MBH descriptor and the HOF descriptor both capture the motion information in videos, MBH further removes the influence of camera motions which makes it better.

On the other hand, we can see that in most cases the cure are very stable, which means that the parameter m

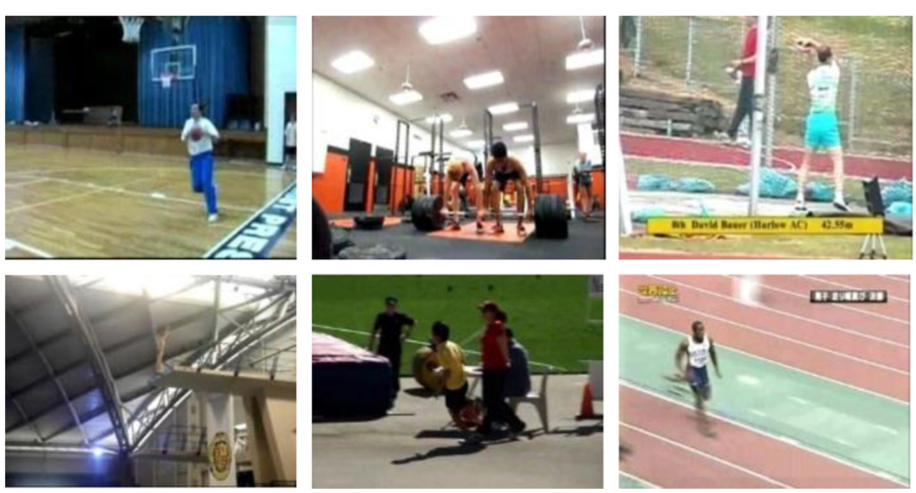

Fig. 4 Screenshots of Olympic dataset

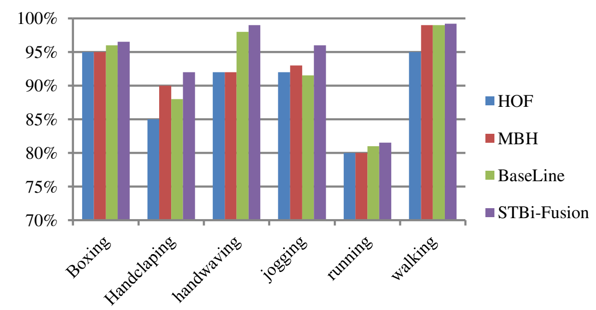

Fig. 5 Comparison of proposed feature fusion algorithm with baseline algorithm

has weak influence on the contingent probability-based representation. Moreover, the more k visual words we retain, the better the accuracy is.

4.4 Experiment 3: comparison of different compress strategies

Table 1 is the results of two different compress strategies on KTH and Olympic datasets. In this table, *POOL STRep* means the POOL compress strategies-based spatio-temporal video representation, and *CP STRep* means the contingent probability-based spatio-temporal video representation. It is shown that POOL-based representation outperform CP-based for 1% on KTH dataset, while CP is better than POOL for 1% on Olympic dataset.

4.5 Experiment 4: spatio-temporal video representation

As Fig. 7 shows, we compare our proposed spatio-temporal video representation with other BoW-based extensions for video representation. Among them, Laptev et al. [2] used a time-consuming per class cross-validation and greedy search to figure the best combination of channels for each video. Wang et al. [3] consider the spatio-temporal contextual information in multiple scales. And Zhou et al. [9] propose a novel structured codebook construction method to encode rich spatial and temporal contextual information for human action recognition.

Table 1 Comparison of different compress strategies

	POOL STRep	CP STRep
KTH	95.37%	94.91%
Olympic	72.90%	73.48%

It is shown that the proposed spatio-temporal video representations, including POOL STRep and CP STRepr, are better than other BoW-based extension methods for boxing, hand-waving, jogging, and walking actions. Meanwhile, we observe that the proposed methods perform relatively worse in hand-clapping class and running class. Because the running action looks similar to the jogging action except the speed, and the hand-clapping action looks similar to the hand-waving action, we need more specific information to distinguish them.

4.6 Experiment 5: combine proposed spatio-temporal video representation and spatio-temporal bigraph-based feature fusion algorithm, and compare with others' method

In this section, we compare the proposed method with the state of the art on both KTH and Olympic datasets.

Table 2 compares our method with the other approaches on their accuracy using the KTH dataset. Other BoW-based extensions are set in italics. On the one hand, we can see that by applying proposed POOL-based representation on MBH feature, we achieve an accuracy of 95.37% which outperforms Kovashka and Grauman [7] method for 1%. By applying proposed contingent probability on MBH feature, we achieve 94.91% which is comparable to Kovashka and Grauman [7] method. Most importantly, by combining the spatio-temporal bigraph-based feature fusion algorithm and POOL-based video representation, we achieve a much better accuracy of 95.83%.

Table 3 compares proposed method with the state of the art on Olympic dataset. We can see that by using STBi-fusion and BoW, we achieve a 71.48% which is

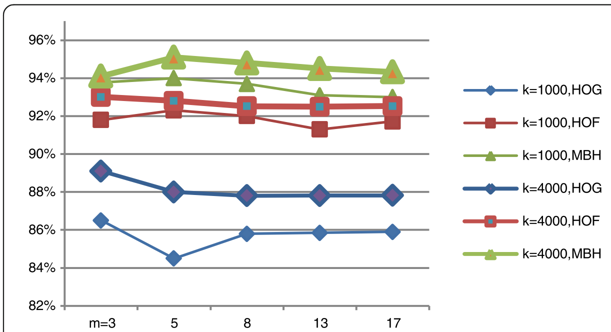

Fig. 6 Influence of evaluation parameters on the contingent probability-based representation

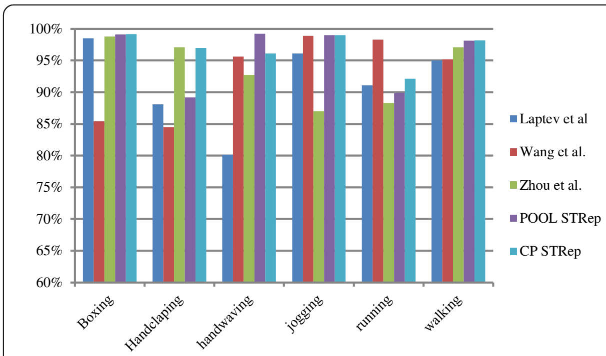

Fig. 7 Compare proposed spatial temporal video representation with other BoW-based method

Table 2 Comparison of proposed method with the state of the art on KTH dataset

Methods	KTH
Fathi et al. 2008 [33]	90.50%
Laptev et al. 2008 [2]	91.80%
Iosifidis et al. 2014 [34]	92.13%
Zhen and Shao 2013 [8]	92.20%
Bregonzio et al. 2009 [35]	93.17%
Yuan et al., 2009 [36]	93.30%
Liu et al. 2009 [4]	93.80%
Zhou et al., 2014 [9]	93.80%
Wang et al. 2011 [3]	93.80%
Wang et al, 2013 [29]	94.20%
Gilbert et al. 2009 [17]	94.50%
Kovashka and Grauman, 2010 [7]	94.53%
MBH + POOL STRep	95.37%
MBH + CP STRep	94.91%
STBi-fusion + BoW	94.44%
STBi-fusion + CP STRep	95.37%
STBi-fusion + POOL STRep	95.83%

comparable to Wang et al's [29] method. Moreover, STBi-fusion + POOL STRep is comparable to Liu et al's [30] 74.38% while STBi-fusion + CP STRep outperforms Liu et al [30] by 74.38% for 0.2%.

5 Conclusions

In this paper, we use the spatio-temporal information among videos to recognize human actions. First, we propose a spatio-temporal bigraph-based feature fusion to combine different features. Second, we introduce a spatio-temporal video representation which uses the spatio-temporal distance between features to measure the distances between visual words. Moreover, two compression strategies are compared experimentally. The experiments show the proposed method is better than other BoW-based extensions. The spatio-temporal bipartite graph-based early fusion technique can further improve the recognition accuracy.

Table 3 Comparison of proposed method with the state of the art on Olympic dataset

Methods	Olympic
Laptev et al, 2008 [2]	62.50%
Tang et al. 2012 [27]	66.80%
Niebles et al. 2010 [28]	72.10%
Wang et al. 2013 [29]	71.60%
Liu et al. 2011 [30]	74.38%
STBi-fusion + BoW	71.48%
STBi-fusion + POOL STRep	73.21%
STBi-fusion + CP STRep	74.40%

Distance matrix is calculated by pairs of all the features in the videos in this paper. For big datasets, this step is time-consuming. In the future, we need to find a new method to calculate the spatial and temporal relationship and reduce the complexity of computing distance matrix.

Although the early fusion of multiple features in the KTH and Olympic datasets have achieved a better average accuracy, the effect is not so good for some classes, such as hand-clapping and running. We plan to combine and-or tree [19] with the early fusion, by searching for an optimal and-or tree to achieve multi-feature fusion. Meanwhile, we plan to combine the low-level feature fusion with high-level feature-based deep-learning framework [31, 32] in the future.

Acknowledgements

This work is supported by National High-tech R&D Program of China (863 Program) (Grant No. 2015AA015904), China Postdoctoral Science Foundation funded project (2015 M571640), Special grade of China Postdoctoral Science Foundation funded project (2016 T90408), CCF-Tencent Open Fund (RAGR20150120), and Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase).

Authors' contributions

LY implemented the core algorithm and drafted the manuscript. YL participated in the video representation and the low-level feature-based framework. SH participated in SVM and helped draft the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interest.

Author details

[1]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing, China. [2]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China. [3]Computer Science and Engineering College, Southeast University, Dongnandaxue Road #2, Jiangning Distinct, Nanjing, People's Republic of China.

Received: 21 June 2016 Accepted: 15 November 2016
Published online: 24 November 2016

References

1. C Yuan et al., Multi-task sparse learning with beta process prior for action recognition, in *IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society*, 2013, pp. 423–429
2. I Laptev et al., Learning realistic human actions from movies, in *Conference on Computer Vision and Pattern Recognition IEEE Computer Society*, 2008, pp. 1–8
3. J Wang, Z Chen, Y Wu, Action recognition with multiscale spatio-temporal contexts, in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on IEEE*, 2011, pp. 3185–3192
4. J Liu, J Luo, M Shah, Recognizing realistic actions from videos "in the wild", in *IEEE Conference on Computer Vision & Pattern Recognition*, 2009, pp. 1996–2003
5. N Ikizlercinbis, S Sclaroff, Object, scene and actions: combining multiple features for human action recognition, in *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I*, 2010, pp. 494–507
6. J Sun et al., Hierarchical spatio-temporal context modeling for action recognition, in *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2004–2011

7. A Kovashka, K Grauman, Learning a hierarchy of discriminative space-time neighborhood features for human action recognition, in *CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2046–2053
8. X Zhen, L Shao, A local descriptor based on Laplacian pyramid coding for action recognition. Pattern Recogn. Lett. **34**(15), 1899–1905 (2013)
9. W Zhou et al., Action recognition via structured codebook construction. Signal Process. Image Commun. **29**(4), 546–555 (2014)
10. W Jiang et al., Short-term audio-visual atoms for generic video concept classification, in *Acm International Conference on Multimedia*, 2009, pp. 5–14
11. J Wei, AC Loui, Audio-visual grouplet: temporal audio-visual interactions for general video concept classification, in *International Conference on Multimedea 2011, Scottsdale, Az, Usa*, 2011, pp. 123–132
12. B Fernando et al., Modeling video evolution for action recognition, in *Cvpr IEEE*, 2015, pp. 5378–5387
13. IH Jhuo et al., Discovering joint audio–visual codewords for video event detection. Mach. Vis. Appl. **25**(1), 33–47 (2014)
14. M Marszalek, I Laptev, C Schmid, Actions in context, in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on IEEE*, 2009, pp. 2929–2936
15. G Zhu et al., Action recognition in broadcast tennis video, in *18th International Conference on Pattern Recognition (ICPR 2006), 20-24 August 2006, Hong Kong, China*, 2006, pp. 251–254
16. D Han, L Bo, C Sminchisescu, Selection and context for action recognition, in *IEEE International Conference on Computer Vision IEEE*, 2009, pp. 1933–1940
17. A Gilbert, J Illingworth, R Bowden, Fast realistic multi-action recognition using mined dense spatio-temporal features, in *IEEE International Conference on Computer Vision IEEE*, 2009, pp. 925–931
18. P Natarajan et al., Multimodal feature fusion for robust event detection in web videos, in *IEEE Conference on Computer Vision & Pattern Recognition*, 2012, pp. 1298–1305
19. K Tang et al., Combining the right features for complex event recognition, in *Proceedings of the 2013 IEEE International Conference on Computer Vision*, 2013, pp. 2696–2703
20. IS Dhillon, S Mallela, R Kumar, A divisive information theoretic feature clustering algorithm for text classification. J. Mach. Learn. Res. **3**(3), 1265–1287 (2003)
21. J Hervé et al., Aggregating local image descriptors into compact codes. IEEE Trans. Pattern Anal. Mach. Intell. **34**(9), 1704–1716 (2012)
22. Y Zhang, J Wu, J Cai, Compact representation for image classification: to choose or to compress? in *Computer Vision and Pattern Recognition IEEE*, 2014, pp. 907–914
23. J Schlenzig, E Hunter, R Jain, Recursive identification of gesture inputs using hidden Markov models, in *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, 1995, pp. 187–194
24. J Shawe-Taylor, S Sun, A review of optimization methodologies in support vector machines. Neurocomputing **74**(17), 3609–3618 (2011)
25. AM Andrew, An introduction to support vector machines and other kernel-based learning methods. AI Mag. **32**(8), 1–28 (2000)
26. C Schuldt, I Laptev, B Caputo, Recognizing Human Actions: A Local SVM Approach, in *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03 IEEE Computer Society*, 2004, pp. 32–36
27. K Tang, F-F Li, D Koller, Learning latent temporal structure for complex event detection, in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on IEEE*, 2012, pp. 1250–1257
28. JC Niebles, CW Chen, FF Li, *Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. Computer Vision – ECCV 2010* (Springer, Berlin Heidelberg, 2010), pp. 392–405
29. H Wang et al., Dense Trajectories and Motion Boundary Descriptors for Action Recognition. Int. J. Comput. Vis. **103**(1), 60–79 (2013)
30. J Liu, B Kuipers, S Savarese, Recognizing human actions by attributes, in *IEEE Conference on Computer Vision & Pattern Recognition*, 2011, pp. 3337–3344
31. CI Patel, S Garg, T Zaveri, et al., Human action recognition using fusion of features for unconstrained video sequences. Comput. Electr. Eng. (2016)
32. JJ Corso, Action bank: a high-level representation of activity in video, in *IEEE Conference on Computer Vision & Pattern Recognition*, 2012, pp. 1234–1241
33. A Fathi, G Mori, Action recognition by learning mid-level motion features, in *Cvpr, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8
34. A Iosifidis, A Tefas, I Pitas, Discriminant bag of words based representation for human action recognition. Pattern Recogn. Lett. **49**, 185–192 (2014)
35. M Bregonzio, S Gong, T Xiang, Recognising action as clouds of space-time interest points, in *IEEE Conference on Computer Vision & Pattern Recognition*, 2009, pp. 1948–1955
36. J Yuan, Z Liu, Y Wu, Discriminative subvolume search for efficient action detection, in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on IEEE*, 2009, pp. 2442–2449