Journal of Electronic Imaging

JElectronicImaging.org

Weighted score-level feature fusion based on Dempster–Shafer evidence theory for action recognition

Guoliang Zhang
Songmin Jia
Xiuzhi Li
Xiangyin Zhang

SPIE. IS&T imaging.org

Guoliang Zhang, Songmin Jia, Xiuzhi Li, Xiangyin Zhang, "Weighted score-level feature fusion based on Dempster–Shafer evidence theory for action recognition," *J. Electron. Imaging* **27**(1), 013021 (2018), doi: 10.1117/1.JEI.27.1.013021.

Weighted score-level feature fusion based on Dempster–Shafer evidence theory for action recognition

Guoliang Zhang,* Songmin Jia, Xiuzhi Li, and Xiangyin Zhang
Beijing University of Technology, Faculty of Information Technology, Beijing, China

Abstract. The majority of human action recognition methods use multifeature fusion strategy to improve the classification performance, where the contribution of different features for specific action has not been paid enough attention. We present an extendible and universal weighted score-level feature fusion method using the Dempster–Shafer (DS) evidence theory based on the pipeline of bag-of-visual-words. First, the partially distinctive samples in the training set are selected to construct the validation set. Then, local spatiotemporal features and pose features are extracted from these samples to obtain evidence information. The DS evidence theory and the proposed rule of survival of the fittest are employed to achieve evidence combination and calculate optimal weight vectors of every feature type belonging to each action class. Finally, the recognition results are deduced via the weighted summation strategy. The performance of the established recognition framework is evaluated on Penn Action dataset and a subset of the joint-annotated human metabolome database (sub-JHMDB). The experiment results demonstrate that the proposed feature fusion method can adequately exploit the complementarity among multiple features and improve upon most of the state-of-the-art algorithms on Penn Action and sub-JHMDB datasets. [DOI: 10.1117/1.JEI.27.1.013021]

Keywords: action recognition; weighted score-level feature fusion; Dempster–Shafer evidence theory; bag-of-visual-words.

Paper 170699 received Aug. 18, 2017; accepted for publication Jan. 16, 2018; published online Feb. 16, 2018.

1 Introduction

Human action recognition for videos has been applied extensively in man–machine interaction systems, video surveillance, virtual reality, and patient monitoring, which are still challenging problems in computer vision due to the complex backgrounds, changeable movement speeds, and different shooting scales with multiperspectives. To improve the robustness and accuracy of the recognition algorithm, many state-of-the-art methods have been proposed.

Recently, local spatiotemporal features[1–4] applied to describe human movements by treating the action volume as a rigid three-dimensional (3-D)-object have achieved promising performance on many datasets.[5] The low-level features are extracted from local regions where the temporal and spatial characteristics change observably or are obtained by dense sampling strategy in videos to represent the patterns of each 3-D volume. These spatiotemporal features usually combine with the pipeline of bag-of-visual-words (BoVW) and its improved variants[6–9] to model human behaviors, which do not require any human detection procedures and have strong robustness to illumination and background. Then, the global representation, which is constructed from a set of local features, is fed into support vector machines (SVMs) to achieve action classification.[9,10] As the two critical steps of this classic and effective process, ample research progress has been made on the methods of local features extraction and features encoding. Laptev and Lindeberg[1] proposed the detector of space-time interest points (STIPs), which is extended from two-dimensional (2-D)-Harris corner detection, and employed histogram of oriented gradients (HOG)[11] and histogram of oriented flow (HOF)[12] to describe the extracted regions. Because the STIPs are usually sparse and more abundant information about human movement cannot be mined, many improved algorithms were put forward.[13,14] Wang et al.[15] demonstrated that dense sampling for video local blocks is more efficient than sparse corner detection. The dense trajectories (DTs)[16] and improved dense trajectories (IDTs),[5] which obtained good performance in various experiments, are presented based on the dense sampling strategy. In the feature encoding stage, several methods can be used to produce a suitable dictionary, such as voting-based encoding,[16–18] Fisher vectors (FV),[3,8,19] and sparse coding techniques.[20,21] As a super vector encoding method, FVs were applied to large-scale image classification by Perronnin et al.[22] A vector of locally aggregated descriptors (VLAD)[23] is an improved algorithm for FVs, where the nearest cluster centers and the per-dimension values of feature points are considered. Although the recognition accuracy of VLAD is slightly lower than FVs, it is more efficient to execute.

The above research for action recognition bypasses body poses and achieves promising results using local spatiotemporal features. Despite their different goals, the two types of features are not only highly coupled but complementary, and it is desirable to study them in a common framework.[24] The prevailing methods for pose estimation[24–26] from still images adopt a pictorial structure model, which resembles the human skeleton and allows for efficient inference based on tree structures.[27] Jhuang et al.[28] used various types of descriptors containing joint position, translation information, and direction of the translational vector, all of which are derived from joint annotations to represent human postural characteristics by employing the pose estimation algorithm

*Address all correspondence to: Guoliang Zhang, E-mail: zhangglmxy@foxmail.com

from Ref. 25. Pishchulin et al.[29] revealed the potential complementarity between holistic methods and pose-based methods by analyzing two kinds of fusion, namely feature- and classifier-level fusions. Meanwhile, Yao et al.[30] proposed a method that requires the videos of training set are from multiple angles and utilizes pose information to optimize the manifold of each action category, then conducts the two tasks iteratively. Nie et al.[24] presented a spatial–temporal and-or graph (AOG) model adopted latent structure-SVM for learning to describe actions at three scales, where coarse-level features are regarded as *a priori* knowledge of pose estimation, and the two tasks benefit from each other in experiments.

Action characteristics in videos ordinarily have many attributes, which describe various categories in different aspects, such as appearance, trajectory of motion, moving boundary, and pose information. The reasonable fusion algorithms[31–33] can utilize the extracted features efficiently and adequately and then boost the performance of constructed system. There are generally three typical methods of combination in the field of action recognition:[8] descriptor-,[2,34] kernel-,[3,35] and score-level fusions.[36,37] Wang et al.[34] integrated multiple descriptors into a new descriptor for subsequent processes of the BoVW framework using a simple strategy for feature weighting. Jain et al.[35] presented an innovative motion descriptor named divergence–curl–shear (DCS), where a linear combination of kernel matrices belonging to each local descriptor is concatenated directly by the method of kernel average and then fed into the linear SVM. For score-level fusion, Myers et al.[36] presented a method that uses cross validation on a training set to obtain the weights of each descriptor, which will combine the scores from multiple classifiers to get the final recognition results.

The core purpose of feature fusion is to enhance the accuracy of recognition using the complementarity among multiple features adequately. In general, each fusion method has its own pros and cons under different circumstances when the action features have fewer types.[8] However, with research going deep, description forms for actions are increasingly numerous. To establish an extensible and universal fusion framework, we focus on the score-level fusion, which does not cause the curse of dimensionality that is prevalent in descriptor- and kernel-level fusion methods. In many cases, although there are some typical dimensionality reduction algorithms, including principal component analysis (PCA),[38] locally linear embedding,[33] and linear discriminant analysis,[39] feature reduction will lose some motion information, which leads to a decrease in recognition accuracy.

In this paper, contrary to the aforementioned approaches, many score-level fusion methods obtain the weights of different features by a learning step, in which the randomness and incompleteness of training data are usually neglected. The Dempster–Shafer (DS) evidence theory can narrow the scope of assumptions continually by the accumulation of evidence and resolve the problem of uncertainty of information. The decision results, which conform to objective condition, are then inferred without the prior probability. In view of the above advantages, the evidence theory is employed to our weighted score-level fusion method, which has not received enough attention in the previous research of action recognition. Concretely, the local spatiotemporal features and the optimized pose features will be extracted from the validation samples, which are selected from the training set, to obtain the credible evidence information. Second, the evidence combination strategy is utilized to calculate weight vectors of all feature types for each action class, which will be optimized by the proposed rule of survival of the fittest. Subsequently, the classification results are deduced by the weighted summation strategy. The main contributions of this paper are as follows:

- According to the characteristics of local features and pose features, the corresponding encoding methods and SVM classifiers are employed. Moreover, to describe the human joints in videos more reasonably, the translation matrix and the angle matrix are constructed to obtain the optimized pose features. The effectiveness of the extendible and universal score-level feature fusion method for action recognition is then demonstrated on Penn Action[40] and a subset of the joint-annotated human metabolome database (sub-JHMDB) datasets.[28]
- Considering the randomness and incompleteness of training data, the weighted score-level feature fusion method based on DS evidence theory (WSF-DS) is proposed, in which the validation set of a dataset is constructed, and weight vectors of multiple features belonging to each action are achieved by evidence combination.
- The rule of survival of the fittest and weighted summation strategy are, respectively, proposed to eliminate the components of weight vectors, which are inefficient and adverse for the recognition of particular action category, and calculate the results of classification.

The rest of this paper is organized as follows: Sec. 2 presents the overview of the proposed action recognition framework. Section 3 elaborates the local spatiotemporal features extraction, optimized pose features extraction, the pipeline of BoVW frameworks for different features, and the proposed weighted score-level feature fusion method. Section 4 evaluates the performance of the proposed action recognition framework on the Penn Action and sub-JHMDB datasets and provides the comparisons with other methods. Finally, we conclude this paper in Sec. 5.

2 System Overview

The framework of the proposed action recognition method, which consists of two stages, is shown in Fig. 1. In the first stage, to obtain the optimized weight vectors of every feature for each action, the original videos in training set are divided into two parts called 3/4 training videos and validation videos (i.e., the remaining 1/4 training videos). For the validation videos, the scenes and human bodies in video clips are more distinctive than the other 3/4 training videos to ensure the validity of evidence. Then, the different BoVW frameworks are adopted for modeling human behaviors based on the local features and pose features.

In the local spatiotemporal features thread, we sample features that include trajectory shape, HOG, HOF, and motion boundary histogram (MBH)[12] based on the IDT method from each video clip in the 3/4 training set. Because the primary low-level local features are usually high dimensional and strongly correlated, the PCA with whitening[38] is used to

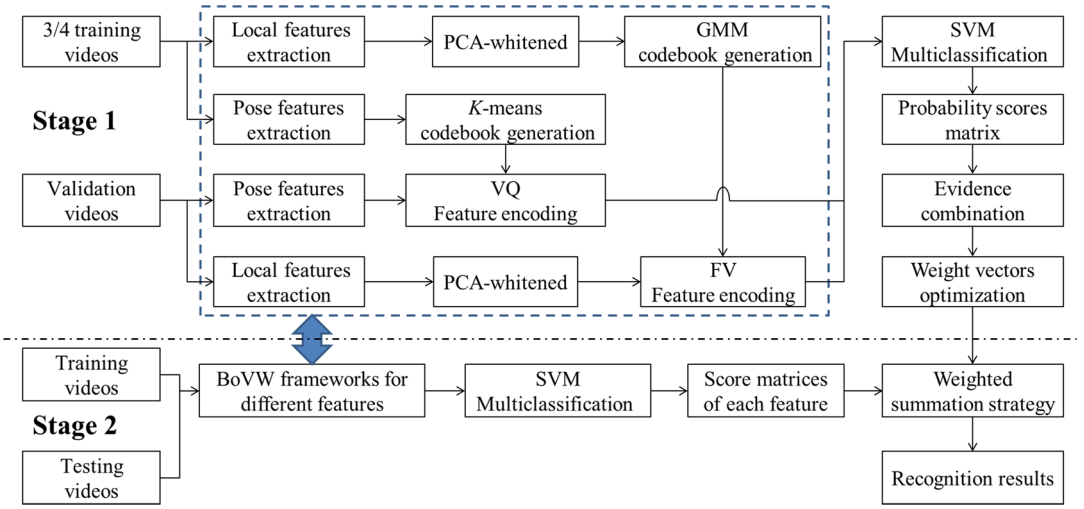

Fig. 1 Framework of the proposed action recognition method.

reduce the dimensionality and weaken the correlation of features. Then, we randomly sample a subset of features from the 3/4 training set to estimate the Gaussian mixture model (GMM), which is learned through maximum likelihood estimation and regarded as a codebook. Unlike the encoding of vector quantization (VQ)[8,16] used in the work of DT,[16] we employ FV[3,22] to encode features and obtain video descriptors. In the pose features thread, the full body joints of every frame in video clips are estimated via the tree-based models of part mixtures.[25] The descriptors for the pose are extracted from two hierarchies to represent different attributes, in which the translation matrix and the angle matrix are constructed to optimize the descriptors of the time hierarchy. All the data of each descriptor type in the 3/4 training set are utilized to generate a codebook by k-means clustering[41] independently and then concatenated as the pose features for each video after the VQ[16,17] encoding method. The implementation of the library for support vector machines (LIBSVM)[42] is used to train multiclassifiers for the two threads.

The probability values of validation videos that belong to each action category are obtained by feeding them to the SVM classifiers. The average probability values of the correct classification for positive and negative samples are utilized as two sources of evidence for DS evidence theory, respectively. Then, the weight vectors are calculated by evidence combination strategy and optimized via the proposed rule of survival of the fittest.

In the second stage, the same BoVW frameworks described above for different features are utilized. Unlike the first stage, the input of the BoVW frameworks is replaced by all the training videos. During the testing stage, when the score matrices of each feature are obtained, the final score matrix of testing videos is calculated by the weighted summation strategy, in which the optimized weight vectors are regarded as *a priori* knowledge. Subsequently, the recognition results (labels) are efficiently inferred by calculating the row maximum of the score matrix.

3 Action Recognition Framework Based on Weighted Score-Level Feature Fusion

In this section, details of the local spatiotemporal features extraction, optimized pose features extraction, the pipeline of BoVW frameworks for different features, and the proposed WSF-DS are presented.

3.1 *Local Spatiotemporal Features Extraction*

This section presents the dense sampling strategy[15] and multiple-features extraction principle. When the trajectory shape, HOG, HOF, and MBH are extracted, a feature preprocessing method is employed to make features have the same variance, which is beneficial for training GMM.

3.1.1 *Multiple-features extraction*

The feature extraction method based on IDT,[5] which conforms to the visual attention mechanism of human eye, is insensitive to the background and motion speed and can describe the apparent information of motion perfectly. The multiple features, including HOG,[11] HOF, and MBH,[12] are extracted around densely sampled points. These points are tracked on each image scale individually. The optical flow field formed by frame t and frame $t+1$ on a certain scale is defined as $\omega_t = (u_t, v_t)$, where u_t and v_t represent the horizontal and vertical components of optical flow, respectively. When a point $P_t = (x_t, y_t)$ in frame t is given, the $P_{t+1} = (x_{t+1}, y_{t+1})$ can be obtained via median filtering in a dense optical flow field

$$P_{t+1} = (x_t, y_t) + (M_{\mathrm{f}} * \omega_t)|_{(\overline{x}_t, \overline{y}_t)}, \tag{1}$$

where M_f is a 3×3 median filter and $(\overline{x}_t, \overline{y}_t)$ represents the rounded position of (x_t, y_t). To reduce the accumulated error produced from trajectory tracking process, trajectory length L is commonly set to 15 frames. The shape of a trajectory is represented by a sequence $T_s = (\Delta P_t, \Delta P_{t+1}, \ldots, \Delta P_{t+L-1})$ for displacement vectors $\Delta P_t = (P_{t+1} - P_t)$. The final trajectory shape feature T_{traj} is expressed as a normalized vector sequence

$$T_{\mathrm{traj}} = \frac{(\Delta P_t, \Delta P_{t+1}, \ldots, \Delta P_{t+L-1})}{\sum_{j=1}^{t+L-1} \|\Delta P_j\|}. \tag{2}$$

For the MBH feature, it is defined as the gradient values for horizontal and vertical components of optical flow field,

and therefore, two histograms (i.e., MBHx and MBHy) can be calculated.[16] To remove the influence of camera motion on recognition accuracy and processing speed, the descriptor for speeded up robust features (SURF)[13] is used to implement frame matching in view of its strong robustness to motion blur, and then the motion vectors are reserved. The IDT method also extracts motion vectors from dense optical flow by employing dense matching strategy among frames. Finally, the DTs are corrected through the global motion vector, which is estimated from the homography matrix calculated by the random sample consensus[43] algorithm.

3.1.2 *Feature preprocessing*

The low-level local features are usually high-dimensional, and there is a strong correlation among different dimensions. To enhance the clustering accuracy of GMM and k-means,[41] PCA is used to eliminate the correlation of feature vectors and reduce the dimensionality of features. Peng et al.[8] proved that combining whitening technology and PCA[38] can effectively boost the recognition accuracy in the BoVW framework. After the above steps, each dimension of features will have the same variance. The mathematical expression of PCA-whiten is as follows:

$$\mathbf{x} = \Lambda D^T \mathbf{f}, \tag{3}$$

where $\mathbf{f} \in R^H$ is the primary feature vector. Λ is a whitening matrix with diagonal elements that can be formulated as $\mathrm{diag}(\Lambda) = [1/\sqrt{\lambda_1}, \ldots, 1/\sqrt{\lambda_i}, \ldots, 1/\sqrt{\lambda_F}]$, where λ_i represents the i'th eigenvalue of feature covariance matrix. $D \in R^{H\times F}$ is a transition matrix used by PCA dimension reduction. $\mathbf{x} \in R^F$ is the processed feature vector.

3.1.3 *Codebook generation*

For high-dimensional local features, the voting-based encoding method (e.g., VQ encoding method) only expresses the subordinate relationship between feature vectors and visual words (i.e., clustering centers), which will produce the quantization errors. In comparison, the FV encodes both first- and second-order statistics between the feature vectors and a GMM. So, we randomly sample a subset of features from the training data to estimate the GMM with K components, which will be regarded as a codebook and employed to calculate the FV. The parameters set of GMM is $\theta = \{\omega_k, \boldsymbol{\mu}_k, \sum_k, k = 1, \ldots, K\}$, where ω_k is the mixed weight of the k'th Gaussian, $\boldsymbol{\mu}_k$ is the mean vector, and $\sum_k$ is the covariance matrix. The probability distribution model $p(\mathbf{x}|\theta)$ of GMM is defined as follows:

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} \omega_k \phi(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k), \tag{4}$$

where $\phi(\mathbf{x}|\mu_k, \sum_k)$ is F-dimensional Gaussian distribution and F is the dimension of feature $\mathbf{x}$, which has been processed by PCA-whiten. When the features set is $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, we learn the parameters of GMM via the maximum likelihood estimation arg $\max_\theta \ln p(X|\theta)$, which is solved by the iterative EM algorithm.[44]

3.2 *Optimized Pose Features Extraction*

IDT features can extract appearance and motion information from videos and achieve a global representation for the action. However, high-level pose features focus on describing the distribution and coupling relationship of human joints. The two types of features are strongly complementary.[29] In this section, the procedure of the optimized pose features extraction will be presented in detail.

3.2.1 *Pose estimation*

The popular methods based on the pictorial structure framework for human pose estimation[24–26] from 2-D video frames imitate the human skeleton and enable systems to efficiently infer the position of human joints in case of tree structures.[27] We follow the framework of Ref. 25 to achieve pose estimation, because it is clear and representative in the principle. It is worth noting that our proposed score-level feature fusion method for action recognition is a universal framework, and completing the pose estimation task is not restricted to one method.

For each image I, the pixel location of the human joint $i \in \{1, \ldots, J\}$ is represented as $p_i = (x, y)$. t_i is the type of joint i, which is defined by the position relation between i and its parent. The number of t_i is equivalent to the number of k-means cluster centers. A J-node tree-structured graph $G = (V, E)$ is constructed, where V is the joint points set and E is the edges set used to represent the parent–child relationships among whole joints. Then, a generalized support function can be defined as

$$S(t) = \sum_{i\in V} b_i^{t_i} + \sum_{i,j\in E} b_{ij}^{t_i,t_j}, \tag{5}$$

where $b_i^{t_i}$ denotes the support of joint i belonging to a designated type and $b_{ij}^{t_i,t_j}$ represents the support of particular co-occurrences of joint types. Subsequently, the full score associated with a set of fixed joint types and positions in image I can be calculated by the following equation:

$$S(I, p, t) = S(t) + \sum_{i\in V} w_i^{t_i} \cdot \boldsymbol{\varphi}(I, p_i) + \sum_{i,j\in E} w_{i,j}^{t_i,t_j} \cdot \psi(p_i - p_j), \tag{6}$$

where $\boldsymbol{\varphi}(I, p_i)$ is an HOG feature vector obtained from joint position p_i.$\psi(p_i - p_j) = [dx\ dx^2\ dy\ dy^2]$ is employed to calculate the relative position between joint i and its parent j. Furthermore, $w_i^{t_i}$ is the filter template for joint i with type t_i, and $w_{i,j}^{t_i,t_j}$ is the filter template for a pair of joints with the type configuration of t_i and t_j.

Then, the latent SVM framework is used to train the detection model by the coordinate-descent solver,[27] where the types of joints are treated as latent variables. In the detection stage, due to the relational graph being a tree structure, human joints in all video frames can be estimated by dynamic programming and nonmaximum suppression. The details can be found in Ref. 25.

3.2.2 *Human joints description*

When the human pose in the video frame is estimated, the joint data are mined to obtain various descriptors, which are

then fed into the BoVW framework. Due to the action in a video clip being decomposed into a series of poses that might change over time, the descriptors for pose need to be carried out from two hierarchies (i.e., space and time) and concatenated into an entire one as the final pose feature. Therefore, we follow Ref. 28 to denote full body pose through 15 joints. For the space hierarchy, joint coordinates are split into x and y, which have proved to be more effective,[28] so 30 descriptor types can be obtained from one frame. For the time hierarchy, the translation of joint coordinates on the time axis (i.e., dx and dy) and the angle of the space-time displacement vector [i.e., $\arctan(dy/dx)$] are calculated according to a certain frame step s.

Note that the configuration of descriptors is different from Ref. 28 here. We find that the translation of joint position in the starting and ending frames of a video clip is usually not salient, and the translation in the middle is more representative for movement tendency. Accordingly, to improve the distinguishing ability of the pose feature, the weakening factor is set as R, and then the translation matrix P_{tr} of joint coordinates for a video with T frames can be written as follows:

$$P_{\mathrm{tr}} = \begin{bmatrix} f_1 - f_{1+s} & f_2 - f_{2+s} & \cdots & f_{T-Rs} - f_{T-(R-1)s} \\ f_{1+s} - f_{1+2s} & f_{2+s} - f_{2+2s} & \cdots & f_{T-(R-1)s} - f_{T-(R-2)s} \\ \vdots & \vdots & \ddots & \vdots \\ f_{1+(R-1)s} - f_{1+Rs} & f_{2+(R-1)s} - f_{2+Rs} & \cdots & f_{T-s} - f_T \end{bmatrix}, \tag{7}$$

$$P_{\mathrm{an}} = \begin{bmatrix} (f_1, f_{1+s}) & (f_2, f_{2+s}) & \cdots & (f_{T-Rs}, f_{T-(R-1)s}) \\ (f_{1+s}, f_{1+2s}) & (f_{2+s}, f_{2+2\,s}) & \cdots & (f_{T-(R-1)s}, f_{T-(R-2)s}) \\ \vdots & \vdots & \ddots & \vdots \\ (f_{1+(R-1)s}, f_{1+Rs}) & (f_{2+(R-1)s}, f_{2+Rs}) & \cdots & (f_{T-s}, f_T) \end{bmatrix}, \tag{8}$$

where f is the data of joint coordinates in a frame, and the indices represent video frame number. The format of the angle matrix P_{an} is same as that of P_{tr}, where the element (f_1, f_{1+s}) represents the angles of space-time displacement vectors of the joint points between frame 1 and frame $1+s$. Then, all elements in P_{tr} and P_{an} are assembled as the holistic description for a video clip, individually.

Finally, these 75 descriptor types (30 for joints coordinates, 30 for translations, and 15 for angle of space-time displacement vector) are separately fed into the subsequent clustering algorithm to generate codebooks and then encoded by VQ.[17]

3.3 *Bag-of-Visual-Words Frameworks for Different Features*

In this work, the BoVW pipeline is employed to build a model for each video clip via the extracted low-level local features and high-level pose features. However, algorithms used in each subunit are different in view of the fact that dimensionality of two categories of features has a clear distinction. In the local feature thread, the global feature vector obtained from each training video is processed by PCA-whiten and then encoded by the FV, where the parameters of GMM are learned by the subset of features. In the pose feature thread, all training data are utilized to construct a codebook with a few vocabularies for each descriptor type by k-means algorithm[41] and there is not any preprocessing, because every descriptor for pose is a one-dimensional vector.

In the classification stage, the two threads are both categorized by SVM use, the implementation of LIBSVM.[42] Due to the encoding methods being different, the linear SVM is chosen as the classifier for local features because it has been proven to be more efficient in combination with FV,[8] and the SVM with radial basis function kernel (RBF-SVM) is similarly selected for pose features, where the optimal parameters are obtained by fivefold cross validation.

3.4 *Weighted Score-Level Feature Fusion Method Based on Dempster–Shafer Evidence Theory*

Multiple features represent actions in different emphases. For example, local spatiotemporal features are used to describe the state of structural and motion around a sampling point, and pose features focus on expressing the joint position and tree structure of a moving human body. There is a strong complementarity among these features.[29] We find that the accuracy of identification is discrepant for an action class when different features are adopted, which means that the strength among each feature is different for a specified action. Therefore, we propose a weighted score-level feature fusion framework, where weight vectors of all feature types for each action are achieved by DS evidence theory[45] through the constructed validation set.

3.4.1 *Evidence theory*

The concept of lower and upper bounds of probability distribution proposed by Dempster[46] is used to solve the problem of multivalued mapping, which is the original work for evidence theory. Shafer[47] proposed a mathematical technique to deal with uncertainty reasoning via a series of rules of evidence combination and introduced the belief function to consummate the evidence theory.

We set Θ as the frame of discernment. The basic belief assignment function m (i.e., mass function) is a mapping from set 2^{Θ} to [0, 1]. $A \subseteq \Theta$ is an arbitrary subset of Θ, which satisfies the following equation set:

$$\begin{cases} m(\varnothing) = 0 \\ \sum_{A \subseteq \Theta} m(A) = 1 \cdot \end{cases} \tag{9}$$

Let $m_1, m_2, \ldots, m_N$ be a set of mass functions on the same frame of discernment Θ. When the focal elements are defined as A_j, the combination rules of DS evidence theory can be utilized to implement information fusion as follows:

$$m(A) = (m_1 \oplus \ldots \oplus m_N) = \frac{1}{1-K} \sum_{\cap A_j = A} \prod_{1 \le i \le N} m_i(A_j)$$
$$A \neq \varnothing. \tag{10}$$

In the case of two evidence combination, the rules can be formulated as

$$\begin{cases} m(A) = \frac{1}{1-K} \sum_{A_i \cap B_j = A} m_1(A_i) m_2(B_j) \; A \neq \varnothing \\ K = \sum_{A_i \cap B_j = \varnothing} m_1(A_i) m_2(B_j) \end{cases}, \tag{11}$$

where K reflects the extent of conflict among evidences and coefficient $\frac{1}{1-K}$ is the regularization factor. When $K \geq 1$, the orthogonal sum $m_1 \oplus m_2$ does not exist, which means the evidence provided by different features from actions cannot be combined.

3.4.2 *Weighted score-level feature fusion method*

To obtain convincing evidence that can reflect the difference in effectiveness between different features in the recognition process of a particular action class, we create a validation set for original dataset (i.e., Penn Action dataset and sub-JHMDB dataset) based on its training samples. More specifically, when the training set is divided into several equal parts, one part is treated as the validation set, in which the scenes and human body in video clips are more distinctive than the other parts to ensure the validity of evidence.

In stage 1, the samples of training sets except validation videos are used to train multiclass classifier belonging to each feature through the BoVW framework presented in Sec. 3.3. For multiclassification, we adopt a one-versus-all cross-validation[5] training scheme and obtain the prediction with probability scores of each sample in the validation set. Then, the probability scores matrix S is defined as

$$S = [s_{ij}^1, s_{ij}^2, \ldots, s_{ij}^h, \ldots, s_{ij}^C], \tag{12}$$

where s_{ij}^h is the probability score of sample i belonging to class j achieved by the multiclass classifier h (i.e., feature h). C is the number of feature types, and the total number of action categories is M.

In stage 2, the 3-D scores matrix S is split into M 2-D score matrices defined as S_j, in which its elements s_{ih} denote the probability score of sample i predicted by classifier h. Assuming that the number of samples belonging to class j is T_r and the number of samples not belonging to class j is F_r, the effectiveness of feature h for a particular action class can be reflected by S_{Tavg}^h and S_{Favg}^h, which are as follows:

$$S_{Tavg}^h = \frac{1}{T_r} \sum_{i=1}^{T_r} s_{ih}, \tag{13}$$

$$S_{Favg}^h = \frac{1}{F_r} \sum_{i=1}^{F_r} (1 - s_{ih}), \tag{14}$$

where S_{Tavg}^h and S_{Favg}^h represent the average probability values of correct classification for the positive samples and negative samples, respectively. We define two average probability vectors of action class j for all feature types as $\mathbf{S_{Tavg}} = [S_{Tavg}^1, S_{Tavg}^2, \ldots, S_{Tavg}^h, \ldots, S_{Tavg}^C]$ and $\mathbf{S_{Favg}} = [S_{Favg}^1, S_{Favg}^2, \ldots, S_{Favg}^h, \ldots, S_{Favg}^C]$, which will be normalized via Eqs. (15) and (16) as two sources of evidence for DS evidence theory

$$P_j = [P_{j1}, P_{j2}, \ldots, P_{jC}] = \frac{\mathbf{S_{Tavg}}}{\sum_{h=1}^{C} S_{Tavg}^h}, \tag{15}$$

$$Q_j = [Q_{j1}, Q_{j2}, \ldots, Q_{jC}] = \frac{\mathbf{S_{Favg}}}{\sum_{h=1}^{C} S_{Favg}^h}. \tag{16}$$

In stage 3, we define a set of focal elements as $\{H_1, H_2, \ldots, H_h, \ldots, H_C\}$ for the two evidences, where H_h means that the positive role of feature h in the recognition process of action class j. The mass functions m_1 and m_2 for the evidence can be assigned as follows:

$$[m_1(H_1), m_1(H_2), \ldots, m_1(H_C), m_2(H_1), m_2(H_2), \ldots, m_2(H_C)]$$
$$= [P_{j1}, P_{j2}, \ldots, P_{jC}, Q_{j1}, Q_{j2}, \ldots, Q_{jC}]. \tag{17}$$

The weight vector of sensitivity for different features belonging to action j is calculated by the strategy of evidence combination [i.e., Eq. (11)] and expressed as follows:

$$\mathbf{W}_j = [m(H_1), m(H_2), \ldots, m(H_h), \ldots, m(H_C)]$$
$$= [w_{j1}, w_{j2}, \ldots, w_{jh}, \ldots, w_{jC}]. \tag{18}$$

In stage 4, inspired by the idea of survival of the fittest, $\mathbf{W}_j$ is optimized during the experiment. Specifically speaking, the features with low weight are not only inefficient for the recognition of a specific action class but also adverse for the final classification score, which should be penalized by the penalty thresholds α and β. For the six action feature types (i.e., trajectory shape, HOG, HOF, MBHx, MBHy, and pose features) in this paper, the rule of elimination is formulated as follows:

- Given the weight vector $\mathbf{W}_j$, the values of its components, which are greater than α, are defined as Ng. When $Ng \geq 3$, the weights less than α will be reset to 0.
- Let $V_{\min}$ be the minimum value of components in $\mathbf{W}_j$. When $Ng \leq 2$ and $V_{\min} \leq \beta$, the smallest value in $\mathbf{W}_j$ will be reset to 0.

The corrected weight vectors of every feature type for each action class are obtained and then utilized for the subsequent classification of testing samples.

In the final stage, the scores matrix Z for all samples in the testing set is calculated by summing the weighted score matrices of each feature, which can be written as

$$Z = \left[\sum_{h=1}^{C} w_{1h} s_{ij}^{h}, \sum_{h=1}^{C} w_{2h} s_{ij}^{h}, \ldots, \sum_{h=1}^{C} w_{jh} s_{ij}^{h}, \ldots, \sum_{h=1}^{C} w_{Mh} s_{ij}^{h}\right], \tag{19}$$

where w_{jh} is the h'th component in the weight vector $\mathbf{W}_j$. We retrieve the maximum values belonging to each row of every scores matrix Z_j (i.e., $Z_j = \sum_{h=1}^{C} w_{jh} s_{ij}^{h}$), which are constituted as a matrix F with M column vectors. Then, the decision function of action recognition can be defined as follows:

$$l_i = \arg \max_{j=1,2,\ldots,M} F_{ij}, \tag{20}$$

where l_i is the column index for the maximum score of sample i. Finally, the action label of testing sample i is inferred efficiently by tracking back to the Z_j, which corresponds to l_i, and retrieving the action label belonging to the maximum value of i'th row.

Note that the proposed pipeline of action recognition is a universal and extendible framework, which has the following characteristics:

- Our weighted score-level feature fusion method can be embedded in different versions of the BoVW pipeline combined with SVM classifier, which only needs to establish a validation set for the corresponding dataset to obtain the weight vectors in the process of method transplantation.
- When an innovative feature needs to be applied in our recognition framework, its effectiveness for different action categories can be analyzed by the WSF-DS method, and the weight vectors will be updated simultaneously. Furthermore, the extendibility of the framework is also reflected in that the local spatiotemporal features and pose features can be replaced by the parallel algorithms for feature extraction to improve overall performance. For instance, the state-of-the-art works for the pose estimation proposed in Refs. 24, 26, and 48 can be employed to replace the algorithm in Ref. 25.
- The strongly targeted weight vectors of each action, which can not only effectively enhance the efficiency of discriminative features but also restrain interference from relatively ineffective features, are calculated by evidence combination and then optimized via the proposed rule of elimination, which combines the idea of survival of the fittest. The feature with low weight for specific action is eliminated in the experiments, which has been found to be effective in improving the accuracy of action recognition.

4 Experiments

The performance of our method is evaluated on two publicly available datasets: Penn Action dataset[40] and sub-JHMDB.[28] Both datasets are proposed for the purposes of action recognition and pose estimation for the full body, and the annotations of human joints and activity labels for each video clip are provided. The experimental results are presented, including the difference in the effectiveness of different features for each action, the evaluation of our proposed weighted score-level feature fusion, a comparison between WSF-DS and multiple-feature fusion baselines, the performance analysis for the proposed rule of survival of the fittest, and a comparison with state-of-the-art action recognition methods.

4.1 *Datasets*

The Penn Action dataset[40] consists of 15 different actions, 13 human joints for each frame, and 2326 video clips collected from the internet, which have the challenges of larger scale and appearance variations, low-resolution images, and obscured human body. The list of action categories is as follows: baseball pitch, baseball swing, bench press, bowling, clean and jerk, golf swing, jump rope, jumping jacks, pull up, push up, sit up, squats, strumming guitar, tennis forehand, and tennis serve. We follow Ref. 24 to discard the class "strumming guitar" and several video clips where most of the human body is invisible and difficult to achieve pose estimation for the full body.

The sub-JHMDB dataset[28] is a subset of JHMDB that contains 15 human joints inside a frame and 316 video clips. The dataset comprises 12 action categories, including catching, climbing stairs, golfing, jumping, kicking ball, picking, pulling up, pushing, running, shooting ball, swinging baseball, and walking. The threefold cross-validation configure presented in Ref. 28 is adopted for testing on the sub-JHMDB dataset. Each split contains on average 229 training samples and 87 testing samples, and the experimental results reported in this paper are the average accuracy of three splits. For the Penn Action dataset, we follow the train/test split released in Ref. 40 (which has been pruned and includes 1206 samples for training and 1017 samples for testing) and report the average accuracy. The sample frames of Penn Action dataset and sub-JHMDB dataset are shown in Fig. 2.

The numbers of validation samples in Penn Action and sub-JHMDB are about 302 and 57, respectively, which are 1/3 of the numbers of 3/4 training samples.

4.2 *Implementation Details*

The proposed action recognition framework is performed on an Intel Core i7-5930K processor with 64-GB RAM and 3.50-GHz frequency. The MATLAB® R2015a with 64-bit is used as the software configuration of code execution.

For the local spatiotemporal features, we use the same settings in Ref. 5, where the size of space-temporal grid is $2 \times 2 \times 3$ and the gradient direction is quantized in 8 directions so that the dimension of HOG is 96. Since the HOF has a stationary state, its dimension is 108. The gradients for horizontal and vertical components of optical flow are defined as MBHx and MBHy, respectively, and the dimension of both features is 96. In addition, the dimension of trajectory shape is 30 when the trajectory length L is set to 15 frames. In the remaining experiments, PCA-whiten[38] is employed to achieve feature reduction and eliminate correlation. For each feature, the dimensions of HOG, MBHx, and MBHy are reduced to 48. Trajectory shape is reduced to 15. HOF is reduced to 54. In the stage of codebook generation, the 256,000 features randomly sampled from each features set are utilized to train the GMM, respectively, which contains 256 Gaussian

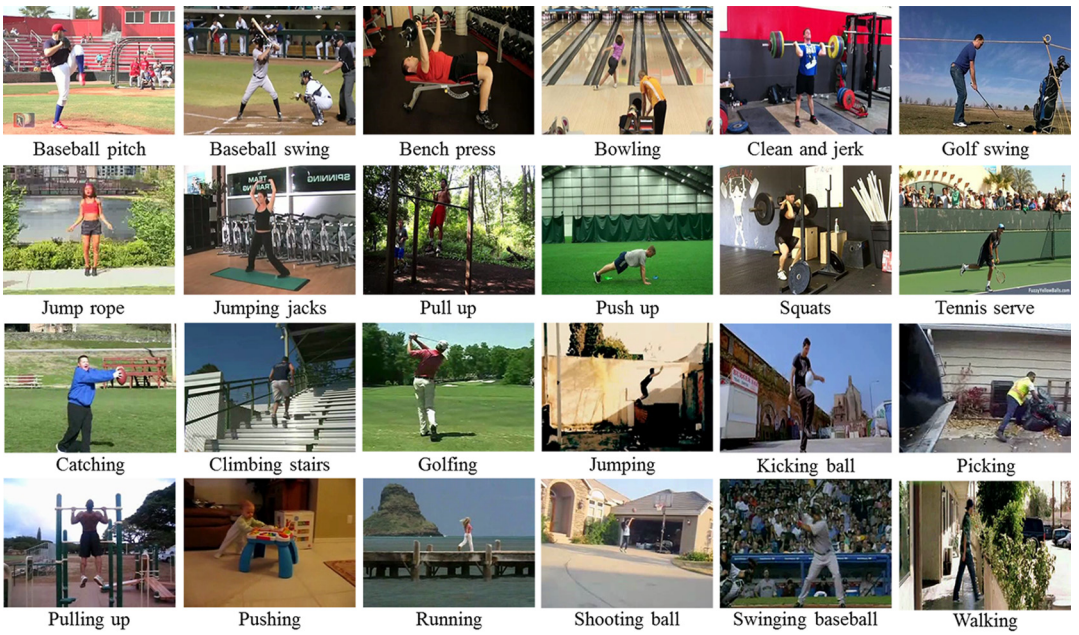

Fig. 2 Sample frames from Penn Action dataset and sub-JHMDB dataset. The frames in the first and second rows are from Penn Action, the third and last rows are from sub-JHMDB.

components. For the FV encoding, the VLFeat Toolbox[49] is employed, and the L2 and power normalization[50] are utilized to perform normalization for FV of each feature.

For the pose features, the model of 26 human joints with 6 types (which is learned by the pose estimation algorithm and proved having good efficiency and performance)[25] is trained to detect human pose in each video frame. The human pose can be described better with dense joints, but it will reduce the distinguishing ability of joints. Because the translation of joint coordinates for some points is less obvious (e.g., joints on the torso), which are meaningless and inefficient for action recognition. Therefore, 15 key points generated from 26 joints are used as the data of pose, which is similar to the work of Jhuang et al.[28] For the descriptors of pose, the frame step size s and the weakening factor R are both set to 3, which have been proven to have good performance. It is worth noting that the 3225 descriptor types proposed in Ref. 28, including a set of relational features, perform better than only using normalized joint coordinates. However, its running time for an ordinary video clip with 42 frames is about 6.17 s when the spatial resolution is 320×240 pixels. In contrast, the running time of the 75 descriptor types optimized in this work is about 0.0058 s under the same conditions, and its recognition accuracy on the sub-JHMDB dataset is basically equal to 52.9%, which is achieved by the 3225 descriptor types. For each descriptor type, all the training samples are utilized to generate an exclusive codebook by k-means algorithm with 20 clustering centers.

For multiclass classification, a one-against-rest approach is adopted to select the prediction with the highest score.

4.3 *Experimental Results*

4.3.1 *Evaluation of effectiveness on different features for each action*

The performance of the six features, including five local spatiotemporal features extracted by IDT method and pose

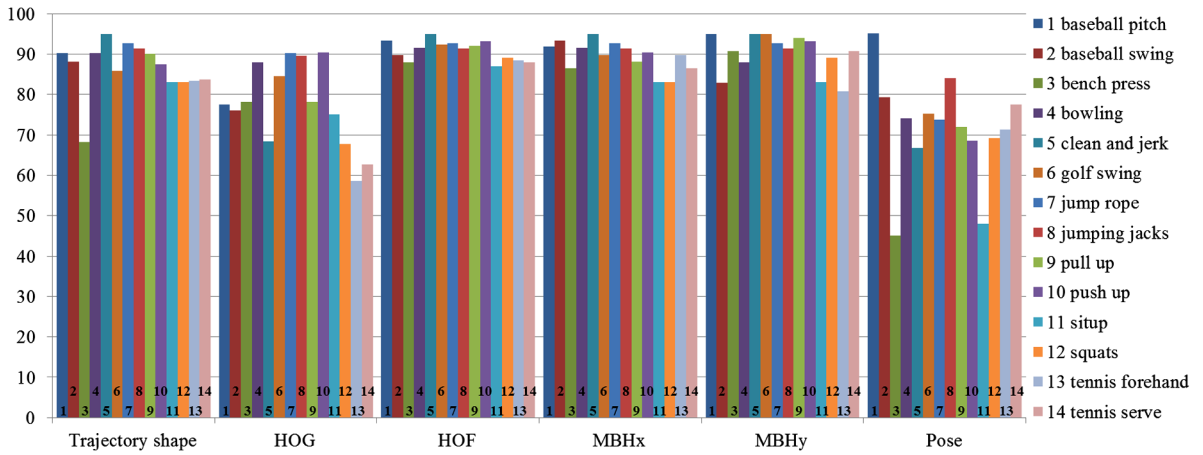

Fig. 3 Classification accuracy comparison of every feature for each action on Penn Action dataset.

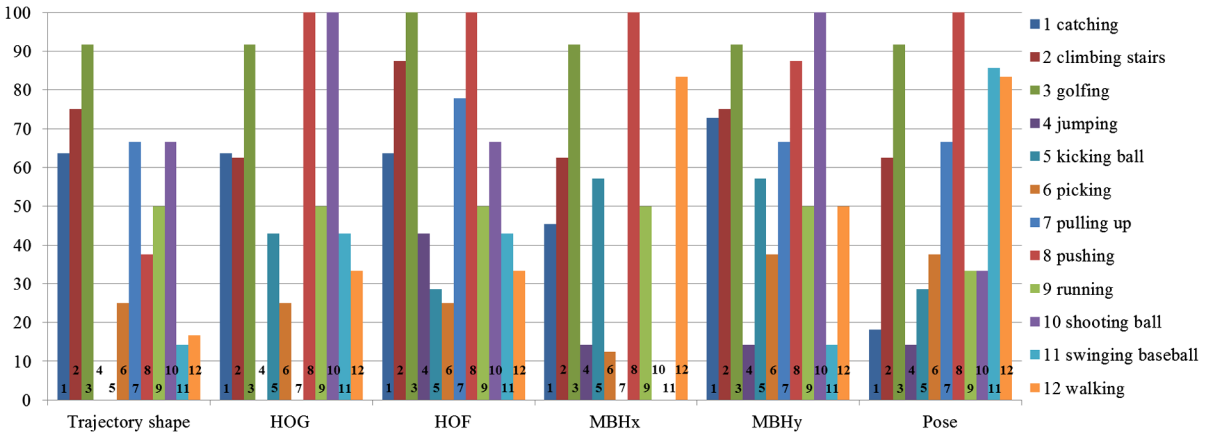

Fig. 4 Classification accuracy comparison of every feature for each action on sub-JHMDB dataset.

features, is evaluated separately on the two public datasets based on the BoVW frameworks presented in Sec. 3.3. The classification accuracies of different features for each action are shown in Figs. 3 and 4. For notational convenience, we only provide the classification accuracy comparison on split-3 of the sub-JHMDB dataset.

The results indicate that there is a great difference between the recognition accuracies of specific features for various actions. For instance, HOG and pose features demonstrate the highest accuracies (90.33% and 95.2%) for individual actions, whereas the lowest accuracies for some actions are only 58.64% and 45.1% as shown in Fig. 3, where the phenomenon is more pronounced in Fig. 4. Furthermore, a specific action category is shown to be more sensitive to several feature types, which is the basis for designing the weighted score-level feature fusion approach. From Fig. 3, the recognition accuracy of action "squats" achieved by HOF or MBHy feature is 89.02%, which outperforms the other four features by 13.29% on average. From Fig. 4, the best classification accuracy for the action "swinging baseball" achieved by pose features is 85.71%. However, HOF demonstrates the highest accuracy among the other five features, which is only just up to 42.86%. From Figs. 3 and 4, due to low resolution and large intracategory discriminations, the overall recognition efficiency of six features on sub-JHMDB is much lower than it on Penn Action.

Based on the above results, we find that the weight vectors of every feature for each action are necessary for improving the overall classification accuracy in the decision-making stage. It is worth noting that the sensitivities of a particular action category to the same set of features between different datasets have a great disparity because of the influences of image resolution, human scale, and various viewpoints, so the corresponding weight vectors for different datasets are required to calculate.

4.3.2 *Evaluation of weighted score-level feature fusion based on Dempster–Shafer evidence theory*

The effectiveness of our proposed WSF-DS method is demonstrated by testing it on two public datasets for human action recognition. To obtain the weight vectors of all feature types for a specific action category, the samples extracted from the training set are assembled as a validation set, and then the evidence used for DS evidence theory is computed by the approach presented in Sec. 3.4. Specifically, to obtain the robust sensitivity information of an action category about every feature type, about 1/4 of training samples that are significantly different from the other 3/4 samples in motion scenes and body appearance are chosen as the validation set.

The influence of different parameters α and β on classification results will be elaborated in Sec. 4.3.3. Here, we report the best performance of our recognition framework. Note that although the average accuracy is reported both for the evaluation of two datasets, we follow Ref. 28 to calculate the per-video accuracy for the sub-JHMDB, which does differ much from the per-class accuracy adopted in Penn Action.[40] The confusion matrices computed by the proposed WSF-DS method for Penn Action and the sub-JHMDB have three splits that are shown in Fig. 5, respectively. Table 1 presents the comparison of average accuracies on the two datasets achieved by different feature types and our WSF-DS method, where the five local spatiotemporal features are extracted by IDT.

From Table 1, when a single feature is employed, the classification accuracies achieved by MBHy are 60.6% and 90.1% on sub-JHMDB and Penn Action individually, which are close to the results achieved by HOF (i.e., 58.1% and 90.8%) but significantly outperform other feature types. This suggests that the optic flow field and motion boundary of the image are more effective than image appearance, motion trajectory, and human pose in the process of action recognition on the two datasets. It should be noted that the estimated joint positions are not precise compared to the ground truth. We leave such pose estimation problem as significant future work, which has been confirmed to be effective in raising the accuracy of action recognition in Ref. 28. Moreover, we observe that the proposed WSF-DS method improves the performance of each single feature type following the order in Table 1 by 7.9%, 17.0%, 3.7%, 5.0%, 4.4%, and 23.1% on Penn Action and 27.3%, 24.1%, 12.9%, 21.0%, 10.4%, and 18.2% on sub-JHMDB. These results also demonstrate that the proposed score-level feature fusion approach can adequately exploit the complementarity among multiple features and be applied on different datasets robustly.

As shown in Fig. 5, our WSF-DS performs well on the actions such as "golfing," "pulling up," "pushing," and

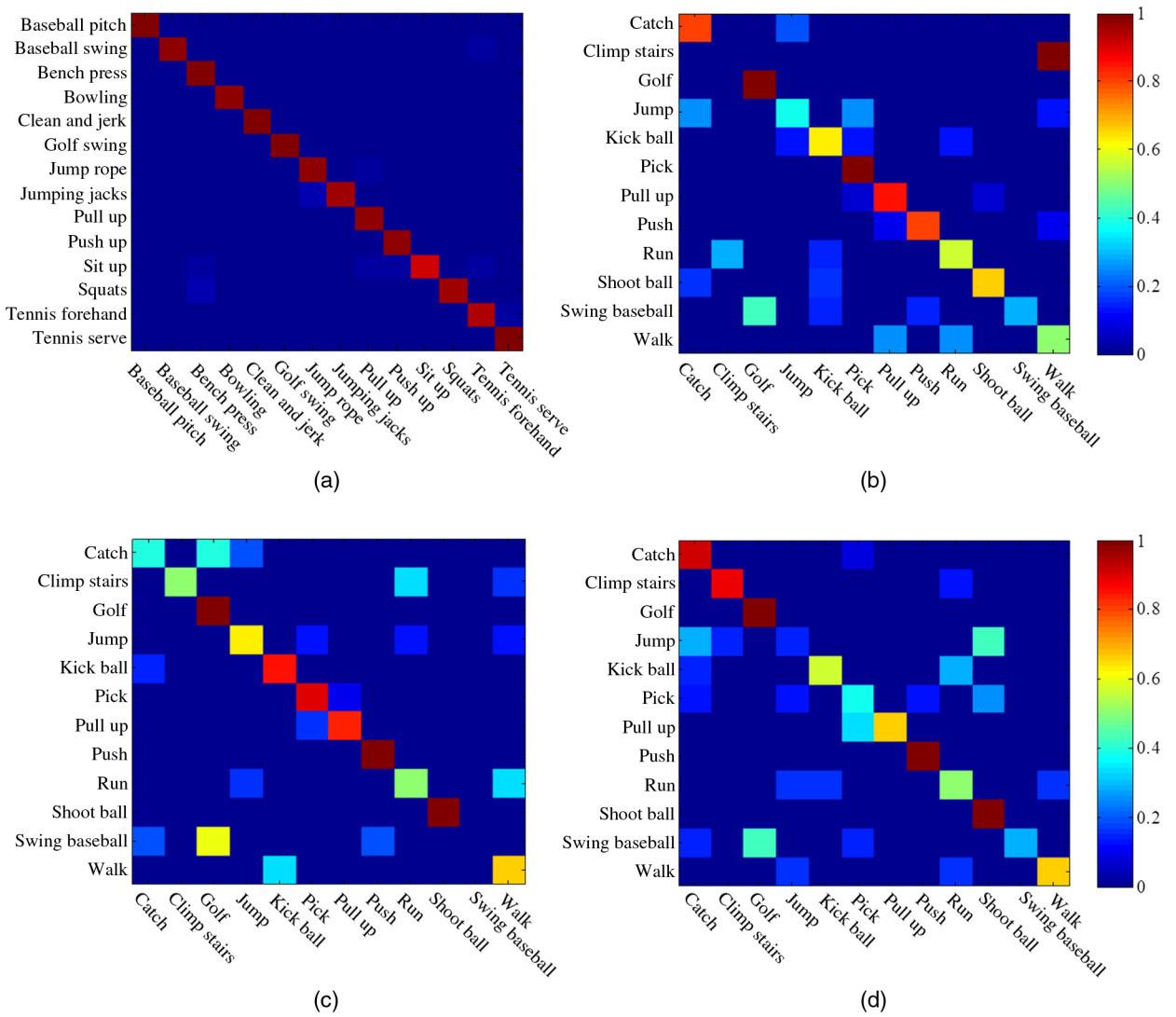

Fig. 5 Confusion matrices for two datasets: (a) the confusion matrix for Penn Action dataset and (b), (c), and (d) the confusion matrices for three splits of sub-JHMDB dataset, respectively.

Table 1 Comparison of the performance for WSF-DS and different feature types on datasets.

	Sub-JHMDB				
Methods	Split 1	Split 2	Split 3	Average	Penn Action
Trajectory shape	40.5	45.0	45.7	43.7	86.6
HOG	38.2	51.3	51.1	46.9	77.5
HOF	53.9	57.5	63.0	58.1	90.8
MBHx	48.3	55.0	46.7	50.0	89.5
MBHy	58.4	62.5	60.9	60.6	90.1
Pose	49.4	52.5	56.5	52.8	71.4
WSF-DS	70.8	73.8	68.5	71.0	94.5

"shooting ball" belonging to sub-JHMDB. However, we achieve low accuracies on several actions, for instance, "swinging baseball" is easy to confuse with "golfing," because the motion patterns between the two actions are similar. For the Penn Action, only the accuracy about "sit up" is significantly lower than other actions because of large intracategory discriminations and varied shooting angles. The proposed method could accurately classify the vast majority of action categories, which demonstrates the effectiveness of our proposed method.

4.3.3 *Comparison with multiple-feature fusion baselines*

This section demonstrates the advantage of the proposed WSF-DS method by comparing our results with two baseline methods of combination in the field of action recognition (i.e., kernel- and score-level fusions). For the descriptor-level fusion mentioned in Sec. 1, we concatenate multiple features extracted from a local cuboid of video into an integrated whole as the input of the BoVW framework, and pose is a holistic feature that describes the distribution of human joints

Table 2 Comparison with multiple-feature fusion baselines.

Methods	Penn Action	Sub-JHMDB
Kernel-level fusion[8]	92.7	65.6
Score-level fusion[8]	93.2	66.3
Score-level fusion[36]	93.4	65.8
WSF-DS (no survival of the fittest)	94.1	69.2
WSF-DS	94.5	71.0

in the entire image. The local sampling is meaningless for it, which causes the fusion method to be unavailable in this work. For kernel-level fusion, each feature is fed into the BoVW individually to obtain different descriptions of action video, which represent various aspects of the motion characteristic, and then fused as a single one to implement action classification by SVM.[8] For score-level fusion, the process is similar to kernel-level fusion. However, the feature fusion operation is executed in the stage of processing scores, where every multiclass classifier, which is trained by different features independently, achieves the scores. We compare our WSF-DS with two typical score-level fusion methods. Specifically, the geometrical mean is employed to combine the scores, which is presented in Ref. 8. A single set of fixed weights for different features is learned by cross validation on the training set and then utilized to obtain the final recognition score, which is presented in Ref. 36. The experimental results on the two datasets are shown in Table 2.

From Table 2, we observe that the WSF-DS method demonstrates higher accuracies than other fusion methods on both Penn Action and sub-JHMDB datasets, which outperforms the best results by 1.1% and 4.7%, respectively. For the fusion strategies of local spatiotemporal features and pose features, the score-level fusion is proved to be more effective. The best accuracies of our weighted score-level fusion increased by almost 1.8% and 5.4% compared to kernel-level fusion. Furthermore, the differences of accuracy rates between two typical score-level fusion methods are <1%, but the method in Ref. 36 has a surplus learning step.

Table 3 Comparison of our WSF-DS with the state-of-the-art methods.

Methods	Year	Penn Action	Sub-JHMDB
Dense[28]	2013	—	46.0
IDT-FV[5]	2013	92.0	60.9
Pose[28]	2013	—	54.1
Pose[26]	2017	79.0	61.5
Pose[51]	2017	—	55.4
Dense + pose[28]	2013	—	52.9
IDT-FV + pose[26]	2017	92.9	74.6
MST[52]	2014	74.0	45.3
AOG[24]	2015	85.5	61.2
P-CNN[53]	2015	—	66.8
WSF-DS (ours)	—	94.5	71.0

We also compare our WSF-DS with the WSF-DS without using the rule of survival of the fittest. The former is 0.4% and 1.8% higher than the latter in the two datasets, which demonstrated the effectiveness of the proposed survival of the fittest. The effect of varying α and β on the accuracy of action recognition on the two datasets is considered in Fig. 6, where $\beta = 0.05$, 0.10, and 0.15 are compared. The idea is that the penalty threshold β should not be larger than the average weight of six feature types (i.e., $\beta = 0.167$).

Figure 6 shows that increasing the value of β will decrease performance, due to the fact that some valuable features are removed in the decision-making stage. It also shows that the values of α corresponding to the optimal accuracies for two datasets are both <0.24 and larger values can cause failure of the proposed survival of the fittest. We report the performance of $\alpha = 0.18$ and $\beta = 0.05$ for Penn Action and $\alpha = 0.22$ and $\beta = 0.05$ for sub-JHMDB in this work.

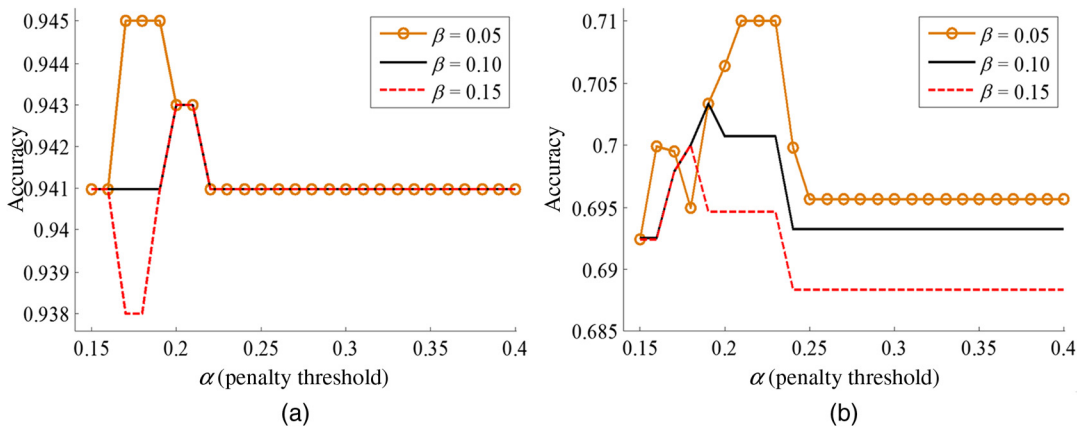

Fig. 6 Performance of WSF-DS as a function of the penalty thresholds α and β on (a) Penn Action and (b) sub-JHMDB datasets.

4.3.4 *Comparison with the state-of-the-art*

The recognition accuracies obtained by our WSF-DS are compared with state-of-the-art methods on Penn Action and sub-JHMDB datasets, and the results are shown in Table 3.

For Penn Action, our WSF-DS has improved the state-of-the-art methods in recent years. For the sub-JHMDB dataset, only the recent work in Ref. 26, which combines improved pose and IDT with FV encoding (IDT-FV),[5] achieves better result than our method, because the more advanced pose estimation algorithm is employed. Moreover, the accuracy achieved by a single local spatiotemporal feature or pose feature is lower than their combination in general. In our experiments, the proposed WSF-DS method achieves better recognition accuracy than most of the recently proposed methods based on the ideas of feature fusion using dendrogram and convolutional neural network features, such as MST,[52] AOG,[24] and P-CNN.[53]

5 Conclusion

In this paper, we proposed an extendible and universal weighted score-level feature fusion method for human action recognition using DS evidence theory. Concretely, the BoVW pipeline is employed to build a model for each video clip via the extracted local spatiotemporal features and pose features. The DS evidence theory and the proposed rule of survival of the fittest are utilized to complete evidence combination and calculate optimal weight vectors of every feature type belonging to each action class. The recognition accuracies of WSF-DS on Penn Action and sub-JHMDB datasets are obtained by the weighted summation strategy, and the experimental results revealed that WSF-DS can achieve promising performance, which outperforms other state-of-the-art methods on Penn Action and sub-JHMDB datasets.

The proposed WSF-DS can enhance the accuracy of classification by utilizing the complementarity among multiple features adequately and perform the task of action recognition efficiently. However, to a certain extent, the overall recognition accuracy of multifeature fusion framework depends to the performances of various features. For example, the more advanced pose estimation algorithm can effectively improve the action recognition performance of pose features[28] and then improve the efficiency of feature fusion.[26]

In the future, the method of obtaining the distribution of human joints and the structure information for the incomplete body will be researched to expand the applied range of the pose estimation. Furthermore, the two types of features will be optimized to excavate more abundant information of appearance and structure for human action and further improve the recognition accuracy and the efficiency of the proposed WSF-DS method.

Disclosures

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was financially supported by the 2017 Beijing University of Technology United Grand Scientific Research Program on Intelligent Manufacturing (No. 040000546317552) and the National Natural Science Foundation of China (Nos. 61175087 and 61703012). The detailed splits and instructions of the 3/4 training set and the validation set for Penn Action and sub-JHMDB can be obtained by contacting us (zhangglmxy@foxmail.com).

References

1. I. Laptev and T. Lindeberg, "On space-time interest points," *Int. J. Comput. Vision* **64**(2), 107–123 (2005).
2. I. Laptev et al., "Learning realistic human actions from movies," in *Proc. of Computer Vision and Pattern Recognition*, pp. 1–8 (2008).
3. H. Wang et al., "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vision* **103**(1), 60–79 (2013).
4. L. Wang, Y. Qiao, and X. Tang, "Latent hierarchical model of temporal structure for complex activity classification," *IEEE Trans. Image Process.* **23**(2), 810–822 (2014).
5. H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. of Int. Conf. on Computer Vision*, pp. 3551–3558 (2013).
6. J. Wu, Y. Zhang, and W. Lin, "Towards good practices for action video encoding," in *Proc. of Computer Vision and Pattern Recognition*, pp. 2577–2584 (2014).
7. Z. Lan et al., "Beyond Gaussian pyramid: multi-skip feature stacking for action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 204–212 (2015).
8. X. J. Peng et al., "Bag of visual words and fusion methods for action recognition: comprehensive study and good practice," *Comput. Vision Image Understanding* **150**, 109–125 (2016).
9. H. Luo et al., "Human action recognition with group lasso regularized-support vector machine," *J. Electron. Imaging* **25**(3), 033015 (2016).
10. G. L. Zhang et al., "Action recognition based on adaptive mutation particle swarm optimization for SVM," *Opt. Precis. Eng.* **25**(6), 1669–1678 (2017).
11. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of Computer Vision and Pattern Recognition*, Vol. 1, pp. 886–893 (2005).
12. N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," *Lect. Notes Comput. Sci.* **3952**, 428–441 (2006).
13. G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," *Lect. Notes Comput. Sci.* **5303**, 650–663 (2008).
14. M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *Proc. of Computer Vision and Pattern Recognition*, pp. 1948–1955 (2009).
15. H. Wang et al., "Evaluation of local spatio-temporal features for action recognition," in *Proc. of British Machine Vision Conf.*, pp. 124.1–124.11 (2009).
16. H. Wang et al., "Action recognition by dense trajectories," in *Proc. of Computer Vision and Pattern Recognition*, pp. 3169–3176 (2011).
17. J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *Proc. of Int. Conf. on Computer Vision*, Vol. 2, pp. 1470–1477 (2003).
18. Z. Wu et al., "Group encoding of local features in image classification," in *Proc. of Int. Conf. on Pattern Recognition*, pp. 1505–1508 (2012).
19. D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with Fisher vectors on a compact feature set," in *Proc. of Int. Conf. on Computer Vision*, pp. 1817–1824 (2013).
20. T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(8), 1576–1588 (2012).
21. A. Castrodad and G. Sapiro, "Sparse modeling of human actions from motion imagery," *Int. J. Comput. Vision* **100**(1), 1–15 (2012).
22. F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," *Lect. Notes Comput. Sci.* **6314**, 143–156 (2010).
23. H. Jegou et al., "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(9), 1704–1716 (2012).
24. X. H. Nie, C. Xiong, and S. C. Zhu, "Joint action recognition and pose estimation from video," in *Proc. of Computer Vision and Pattern Recognition*, pp. 1293–1301 (2015).
25. Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(12), 2878–2890 (2013).
26. U. Iqbal, M. Garbade, and J. Gall, "Pose for action-action for pose," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG) Workshops*, pp. 438–445 (2017).
27. P. F. Felzenszwalb et al., "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(9), 1627–1645 (2010).
28. H. Jhuang et al., "Towards understanding action recognition," in *Proc. of Int. Conf. on Computer Vision*, pp. 3192–3199 (2013).
29. L. Pishchulin, M. Andriluka, and B. Schiele, "Fine-grained activity recognition with holistic and pose based features," in *German Conf. on Pattern Recognition*, pp. 678–689 (2014).

30. A. Yao, J. Gall, and L. Van Gool, "Coupled action recognition and pose estimation from multiple views," *Int. J. Comput. Vision* **100**(1), 16–37 (2012).
31. H. Kuang et al., "Fruit classification based on weighted score-level feature fusion," *J. Electron. Imaging* **25**(1), 013009 (2016).
32. L. Ma et al., "Multiple feature fusion via weighted entropy for visual tracking," in *Proc. of Int. Conf. on Computer Vision*, pp. 3128–3136 (2015).
33. B. Y. Sun et al., "Feature fusion using locally linear embedding for classification," *IEEE Trans. Neural Networks* **21**(1), 163–168 (2010).
34. X. Wang, L. M. Wang, and Y. Qiao, "A comparative study of encoding, pooling and normalization methods for action recognition," *Lect. Notes Comput. Sci.* **7726**, 572–585 (2012).
35. M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *Proc. of Computer Vision and Pattern Recognition*, pp. 2555–2562 (2013).
36. G. K. Myers et al., "Evaluating multimedia features and fusion for example-based event detection," *Mach. Vision Appl.* **25**(1), 17–32 (2014).
37. K. Tang et al., "Combining the right features for complex event recognition," in *Proc. of Int. Conf. on Computer Vision*, pp. 2696–2703 (2013).
38. H. Gou and O. Chum, "Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening," *Lect. Notes Comput. Sci.* **7573**, 774–787 (2012).
39. E. I. Altman, G. Marco, and F. Varetto, "Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks (the Italian experience)," *J. Banking Finance* **18**(3), 505–529 (1994).
40. W. Zhang, M. Zhu, and K. G. Derpanis, "From actemes to action: a strongly-supervised representation for detailed action understanding," in *Proc. of Int. Conf. on Computer Vision*, pp. 2248–2255 (2013).
41. S. Mathavan et al., "Fast segmentation of industrial quality pavement images using laws texture energy measures and k-means clustering," *J. Electron. Imaging* **25**(5), 053010 (2016).
42. C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intell. Syst. Technol.* **2**(3), 1–27 (2011).
43. S. M. Jia et al., "A novel improved probability-guided RANSAC algorithm for robot 3D map building," *J. Sens.* **2016**(1), 1–18 (2016).
44. T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Process Mag.* **13**(6), 47–60 (1996).
45. S. M. Jia and G. L. Zhang, "Evaluation method of module granularity partition for intelligent service robot based on DS evidence theory," in *IEEE Int. Conf. on Information and Automation*, pp. 870–875 (2016).
46. A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Ann. Math. Stat.* **38**(2), 325–339 (1967).
47. G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton (1976).
48. X. Chu et al., "Structured feature learning for pose estimation," in *Proc. of Computer Vision and Pattern Recognition*, pp. 4715–4723 (2016).
49. A. Vedaldi and B. Fulkerson, "VLFeat: an open and portable library of computer vision algorithms," in *Proc. of ACM Int. Conf. on Multimedia*, pp. 1469–1472 (2010).
50. J. Sánchez et al., "Image classification with the Fisher vector: theory and practice," *Int. J. Comput. Vision* **105**(3), 222–245 (2013).
51. S. Cao, K. Chen, and R. Nevatia, "Activity recognition and prediction with pose based discriminative patch model," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pp. 1–9 (2016).
52. J. Wang et al., "Cross-view action modeling, learning and recognition," in *Proc. of Computer Vision and Pattern Recognition*, pp. 2649–2656 (2014).
53. G. Chéron, I. Laptev, and C. Schmid, "P-CNN: pose-based CNN features for action recognition," in *Proc. of Int. Conf. on Computer Vision*, pp. 3218–3226 (2015).

Guoliang Zhang is currently a PhD candidate at the Faculty of Information Technology, Beijing University of Technology (BJUT), China. His research interests include action recognition, machine learning, and man–machine interaction system of robots.

Songmin Jia received her PhD from the University of Electro-Communications, Japan, in 2002. Currently, she is a professor at the Faculty of Information Technology, BJUT. Her research interests include distributed robotics, machine learning, visual computation, and image processing.

Xiuzhi Li received his PhD from Beihang University, China, in 2008. Currently, he is an associate professor at the Faculty of Information Technology, BJUT. His research interests include computer vision, three-dimensional image reconstruction, and mobile robot control and navigation.

Xiangyin Zhang received his PhD from Beihang University, China, in 2016. Currently, he is a lecturer at the Faculty of Information Technology, BJUT. His research interests include bionic intelligent computing theory and application, machine vision, and robot control theory.