ACTION RECOGNITION WITH NOVEL HIGH-LEVEL POSE FEATURES

*Jiayi Fan, Zhengjun Zha, Xinmei Tian**

University of Science and Technology of China
CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System
jyfan91@mail.ustc.edu.cn, zhazj@ustc.edu.cn, xinmei@ustc.edu.cn

ABSTRACT

Recently high-level pose features (HLPF) have been shown to be efficient for action recognition in joint-annotated tasks. However, the relative positions between pairs of joints in actual situations and the spatio-temporal information are not considered in constructing HLPF. To tackle their problems, we propose a set of novel high-level pose features (NHLPF). Specifically, considering that the distances between adjacent pairs of joints usually remain unchanged, we propose a horizontally relative position feature and a vertically relative position feature. In addition, a joint inner product feature is proposed to code the spatial information among each triplet of joints. To code temporal information, we calculate the trajectories of the above-mentioned three types of features as corresponding trajectory features. Furthermore, to combine the spatial and temporal information, we present a joint energy change feature, which is designed using observations of the magnitude and direction of the force between joints. We evaluate our NHLPF on a benchmark dataset. The results show that NHPLF are superior features for action recognition.

Index Terms— Action recognition, high-level pose features

1. INTRODUCTION

Action recognition and pose estimation are both of great significance in the field of computer vision. They have a multitude of potential applications, such as intelligent surveillance and human-computer interaction. In spite of the different goals of these tasks, many existing methods use pose estimation as an input to recognize actions. Unfortunately, the performance of pose estimation is far from perfect due to large pose variations and complex backgrounds. To examine the performance of features under the condition that pose estimation is perfect, some researchers begin to use annotated joints to study action recognition.

*This work is supported by the NSFC under the contracts No.61201413 and No.61572451, Youth Innovation Promotion Association CAS CX2100060016, Fok Ying Tung Education Foundation, the Specialized Research Fund for the Doctoral Program of Higher Education No.WJ2100060003, the Fundamental Research Funds for the Central Universities WK2100060011 and WK2100100021.

High-level pose features (HLPF), which are a combination of nine human pose-based features, proposed by Jhuang et al. [1], show excellent results on joint-annotated datasets, JHMDB. Three features in HLPF describe the positions of joints and the trajectories of their motions in the Cartesian coordinates as well as the polar coordinates. Four features describe the distance relations and the orientation relations between pairs of joints as well as their trajectories. In detail, the distance relations feature describes the distances between pairs of joints. The orientation relations feature describes the intersection angles between lines connecting different joints and the horizontal line. The other two features describe the angle relations among triplets of joints and their trajectories. Although the performance of HLPF is pleasurable on challenging datasets, the relative positions of joints in actual situations and the spatio-temporal information are not considered in the process of constructing features.

First, in action recognition, the viewpoint of a video and the distance between adjacent joints in a frame usually remain unchanged, when a person performs a kind of action. For example, when a person runs, the distance between the hip and knee almost remains the same, as shown in Figure 1 (a) and (b). In addition, when the same person performs different actions, the distance between the same adjacent joints usually remains the same. Consequently, the discriminative power of the distance relations feature is not strong enough. We can draw the same conclusion for the orientation relations feature. For the same action, different people may have different action amplitudes. For the same person performing the same action, the action amplitude may be different at different times. Thus for a particular action, the orientation of a pair of joints may vary in a wide range. Moreover, the orientation is seriously affected by the viewpoint. Therefore, the discriminative power of the orientation relations feature is not strong enough, either. The primary reason behind this conclusion is that the distance relations feature and the orientation relations feature treat each orientation equally; however, the discriminative power in varying orientations is different. In each frame, we use a minimum enclosing rectangle to locate the torso of a person performing an action, such as brushing hair, clapping and running. We can see that the long edge of the rectangle is often approximately perpendicular to the

978–1–4799–7082–7/15/$31.00

ground. In addition, for limbs, the direction of motion generally adopts a horizontal or vertical orientation in relation to the aforementioned long edge of rectangle. This means that the discriminative power of horizontal and vertical orientations is much stronger. So, it is advantageous to calculate horizontally and vertically relative positions between each pair of joints. Thus, we present a horizontally relative position feature and a vertically relative position feature.

Secondly, three kinds of relations are described in [1]: single joint position, relative position relation between each pair of joints, and relative position relation among each triplet of joints. For the relation between each pair of joints, the previous paragraph shows that the horizontally and vertically relative position features perform better than corresponding features in HLPF. From the perspective of inner product space, the horizontally (vertically) relative position feature is the inner product between the horizontal (vertical) unit vector and a vector from one joint to another. This inspired us to presume that the inner products between the vectors of joints in each joint triplet could also perform better. Thus, we propose a joint inner product feature which describes the position relations of triplets of joints in an original way. Our experiments verify that the joint inner product feature is more effective than the angle relations feature in HLPF.

Thirdly, some features in HLPF describe the spatial information of joints, e.g., the distance relations feature, while others describe the temporal information, e.g., the distance trajectory feature. But there is no feature that describes the spatial-temporal information. In order to describe this information, we propose a novel feature, namely, the joint energy change feature. We hold the view that for a particular action, energy from each joint will change due to the motion of other joints. For different actions, the energy changes of different joints vary. Thus, we calculate the inner products between the joint vectors and their trajectories to obtain this novel feature. This process is based, first, on the theorem that energy changes are equal to the net force acting upon them, and secondly, from observations regarding the direction and magnitude of the force between each pair of joints.

In summary, to address the problems of HLPF [1], we propose novel high-level pose features (NHLPF). Considering that the distance between pairs of adjacent joints usually remain unchanged when people perform different actions, we propose the horizontally and vertically relative position features. Using the method of inner product, the joint inner product feature is constructed. The joint energy change feature is designed to supply spatio-temporal information. We conduct experiments on the challenging dataset: JHMDB [1], for which manual annotation of human poses have been provided. NHLPF improves the state-of-the-art on this dataset.

2. RELATED WORK

Action recognition is a hot topic in computer vision and there is a mountain of scholarship about this topic. In this section, we introduce some recent works. Action recognition methods can be grouped into two categories: methods based on local space-time interest points and pose-based methods.

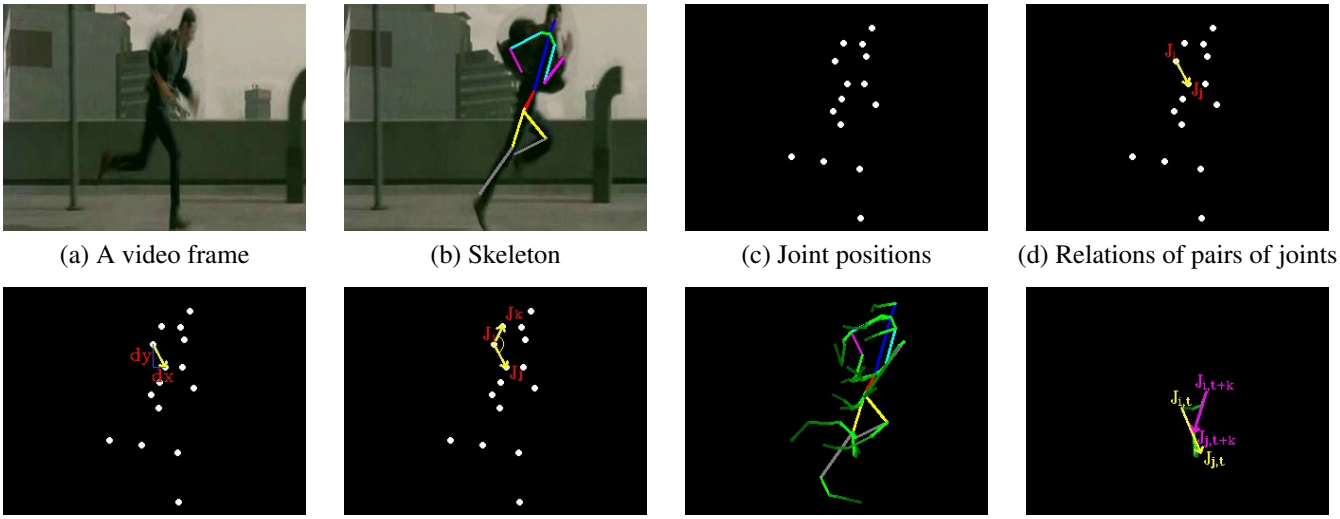
Methods based on local space-time interest points The framework of these methods can be divided into detection parts and description parts. Some classic detectors, such as Harris 3D detector [2], the Cuboid detector [3] and the Hessian detector [4] are spatio-temporal extensions of 2D detectors or saliency measures. After space-time interest points are detected, descriptors are computed at these points. HOG/HOF descriptors [5] characterize local motion and appearance. HOG3D [6], ESURF [7], 3D SIFT [8], and SOD [9] are 3D features computed in spatio-temporal space using temporal sequences of images. After descriptors are extracted, bag-of-words representation is used for classification. In particular, the improved dense trajectory features with Fisher vector representation have achieved an outstanding performance on a multitude of challenging datasets [10] [11].

Pose-based methods In these methods, pose estimation and pose-based description are two indispensable procedures for action recognition. Yao et al. demonstrate that when using pose estimation from multiple camera views, actions can be reliably recognized [12]. For relatively simple datasets composed of monocular videos, it is established that estimated poses are reliable to do the following action recognition [13]. Though providing reliable estimations for uncontrolled datasets is still complicated, Jhuang et al. propose a challenging dataset, the JHMDB dataset, in which all the joints are annotated [1]. Besides, Jhuang et al. present high-level pose features (HLPF) [1], which show excellent performance on the JHMDB dataset when joint positions are manually annotated in each frame. Cheron et al. design pose-based CNN (P-CNN) features [14]. When combined with the improved dense trajectory features encoded by Fisher vector, P-CNN features perform better than HLPF. However, due to high computational complexity, these combined features are not efficient.

In this work, we solve some problems in [1] which is closely related to our work, and use pose-based methods to construct features which show promising results in our experiments.

3. HIGH-LEVEL POSE FEATURES (HLPF)

In this section, we introduce high-level pose features (HLPF), which are used as a baseline in our algorithm. HLPF introduced in [1] describe simple spatial and temporal relations among the positions of the human joints. HLPF consist 9 features which can be grouped into spatial relations features and temporal relations features.



(a) A video frame (b) Skeleton (c) Joint positions (d) Relations of pairs of joints

(e) Relations of pairs of joints (f) Relations of triplets of joints (g) Trajectories of joints (h) Relations of hip and knee

Fig. 1: Overview of the construction of some features in HLPF and NHLPF. (a) A video frame from JHMDB dataset at frame t. (b) The image at frame $t+k$ with connections at adjacent joints in different colors. The frame is from the same video clip as the clip in (a). (c) Annotated joint positions at frame t. (d) Distance relations and orientation relations between pairs of joints. (e) Horizontal and vertical relative positions between pairs of joints. (f) Angle relations and joint inner product among triplets of joints. (g) Trajectories of joint positions from frame t to $t+k$. (h) The vector from right hip to right knee at frame t and frame $t+k$ as well as trajectories of right hip and right knee.

Spatial relations features are computed from the positions of three categories, namely, single joints, pairs of joints and triplets of joints. For each frame, $x-$ and $y-$ coordinates from all 15 joints are annotated, as shown in Figure 1 (c). For each single joint, the position is first normalized with respect to the human scale. Then, the relative position of each joint to the center of the human body is computed to form **normalized positions feature** that has 30 dimensions. For each pair of joints, **distance relations feature** is obtained by calculating the distance between this joint pair. **Orientation relations feature** is obtained by calculating the orientation of the vector connecting each pair of joints, as shown in Figure 1 (d). There are $C_{15}^2 = 105$ kinds of pairs, yielding 105 dimensions for both features. For each triplet of joints, **angle relations feature** is obtained by calculating the inner angles that span the vectors connecting this joint triplet, as shown in Figure 1 (f). Since there are $3 \times C_{15}^3 = 1365$ inner angles spanned by two vectors connecting all the triplets of joints, there are 1365 dimensions for this feature.

Temporal relations features are considered as the difference between spatial features along the trajectory at frame t and $t+k$, i.e. the feature of dimension f is a sequence: $(f_{(t+s)}-f_t, \cdots, f_{(t+ks)}-f_{(t+(k-1)s)})$, $k=(T-t)/s$, where T is the trajectory length, k is the sequence length, and s is the step size. $T=7$ and $k=2$ are used in [1]. For each single joint, they use the translation of normalized position along x and y coordinates $(x_{t_2}-x_{t_1}, y_{t_2}-y_{t_1})$, called **Cartesian trajectory feature**. Translation of orientation $arctan(\frac{y_{t_2}-y_{t_1}}{x_{t_2}-x_{t_1}})$ is computed as **radial trajectory feature**. For each pair or triplet of joints, the trajectories of each spatial feature are calculated to obtain the corresponding trajectory feature. Since the sequence length of each feature dimension is 2, the dimension of each temporal feature is double its corresponding spatial feature. This does not apply to the radial trajectory feature, which has 30 dimensions. Figure 1 (g) shows an instance of Cartesian trajectories with sequence length $k=5$. The pose is presented at the frame $t+k$ and the trajectories, shown in green, grow brighter with the passage of time.

For each dimension of these features, a codebook is generated using k-means with $k=20$ for quantization. After each video clip is described by a histogram, SVM is used to perform training. More details can be found in [1].

4. NOVEL HIGH-LEVEL POSE FEATURES (NHLPF)

In this section, we introduce our novel pose-based features in detail.

4.1. Horizontally and vertically relative position features

For pairs of joints, we design features from the perspective of the position relations of joints in spatial and temporal space.

For spatial features, the **horizontally and vertically relative position features** are obtained by calculating the difference of $x-$ and $y-$ coordinates from each pair of joints, as shown in Figure 1 (e). In detail, suppose that the positions of the pair of joints $\{J_i, J_j\}$ at frame t are $(x_{i,t}, y_{i,t})$ and $(x_{j,t}, y_{j,t})$, the horizontally relative position (HRP) from $J_{i,t}$

to $J_{j,t}$ is:

$$HRP(i,j,t) = x_{j,t} - x_{i,t}. \quad (1)$$

The vertically relative position (VRP) from $J_{i,t}$ to $J_{j,t}$ is:

$$VRP(i,j,t) = y_{j,t} - y_{i,t}. \quad (2)$$

There are $C_{15}^2 = 105$ kinds of pairs, yielding 105 dimensions for both features.

For temporal features, trajectory features are calculated according to the method mentioned in Section 3 with the same parameters. Since the length of a sequence is 2, the dimension of both features is 210.

4.2. Joint inner product feature

For triplets of joints, we also construct features from the perspective of the position relations of joints in spatial space and temporal space, respectively.

For the spatial feature of each triplet of joints, the **joint inner product feature** is obtained by computing the inner product of a pair of vectors , as shown in Figure 1 (f). At frame t, let $(x_{i,t}, y_{i,t})$, $(x_{j,t}, y_{j,t})$, $(x_{k,t}, y_{k,t})$ denote positions of joints $\{J_{i,t}$,$J_{j,t}$,$J_{k,t}\}$ in a triplet of joints, respectively. The joint inner product (JIP) of this triplet is calculated as:

$$\begin{aligned} JIP(i,j,k,t) &= \overrightarrow{J_{i,t}J_{j,t}} \cdot \overrightarrow{J_{i,t}J_{k,t}} \\ &= (x_{j,t} - x_{i,t})(x_{k,t} - x_{i,t}) + (y_{j,t} - y_{i,t})(y_{k,t} - y_{i,t}). \end{aligned} \quad (3)$$

$3 \times C_{15}^3 = 1365$ triplets of joints result in 1365 dimensions for this feature.

For the temporal feature, the trajectory feature is calculated as mentioned before. The dimension of the trajectory feature is 2730.

4.3. Joint energy change feature

To combine spatial information with temporal information, we propose a **joint energy change feature**. This feature is constructed after making some observations.

First, each joint influences other joints through the force. The magnitude is highly negatively correlated with the distance between each pair of joints. For each pair of adjacent joints, the direction of the force is from one joint to the other. For other pairs of joints, the force from one joint to the other is the vector sum of the force from one joint to an adjacent joint. Thus for each pair of nonadjacent joints, the direction of the force is also from one joint to the other.

Second, the magnitude of the force remains unchanged during a short period of time, i.e. a few frames. With these two observations, the work of the force can be computed.

As shown in Figure 1 (h), the force F from J_i to J_j at frame t is:

$$F(i,j,t) = \frac{C}{\| J_{j,t} - J_{i,t} \|^2}(J_{j,t} - J_{i,t}). \quad (4)$$

Without the loss of distinctiveness, here we set C to 1. As just mentioned, the magnitude of the force F remains unchanged during a short period time. So at frame t' ($t \leqslant t' \leqslant t+k$, k is a small integer, t' is an integer), the force F from J_i to J_j is:

$$F(i,j,t') = F(i,j,t). \quad (5)$$

The displacement S of the force F from J_i to J_j in k frames is:

$$S(i,j,t,k) = (J_{j,t+k} - J_{i,t+k}) - (J_{j,t} - J_{i,t}). \quad (6)$$

Accordingly, the work JEC of the force F from J_i to J_j in k frames is:

$$JEC(i,j,t,k) = F(i,j,t)S(i,j,t,k) \quad (7)$$

Thus, we get the joint energy change feature. There are $C_{15}^2 = 105$ kinds of pairs, yielding 105 dimensions for this kind of feature.

As mentioned above, we have described the position relations between each pair of joints and among each triplet of joints, but the position information of each single joint hasn't been described yet. To supply the position information of single joints, the normalized positions feature, the Cartesian trajectory feature, and the radial trajectory feature are used in our method. We combine the features computed from single joints (the normalized positions feature, the Cartesian trajectory feature, and the radial trajectory feature), the features obtained from the relations of joint pairs (the horizontally and vertically relative position features, their trajectory features, as well as the joint energy change feature), and the features calculated from the relations of joint triplets (the joint inner product feature and its trajectory feature) to construct the novel high-level pose features (NHLPF).

5. EXPERIMENTS

In this section we introduce the settings used in our experiments and show experimental results of our novel features.

5.1. Experimental settings

We conduct experiments on the JHMDB dataset [1] which contains 21 human actions involving *brush hair, catch, clap, climb stairs, golf, jump, kick ball, pick, pour, pull-up, push, run, shoot ball, shoot bow, shoot gun, sit, stand, swing baseball, throw, walk* and *wave*. Video clips are restricted to the duration of the action. There are $36-55$ clips per action class with each clip containing $15-40$ frames of size 320×240. Human poses are annotated in each frame. Consequently, there are 928 clips with 31838 frames annotated in total. 15 joints including *shoulder, elbow, wrist, hip, knee, ankle, neck, face* and *belly* are all annotated manually, no matter whether the joints are inside the frame.

We use l_2 normalization to normalize features. For quantization, a codebook is generated for each feature using k-means with $k = 20$. All the training samples are used as

Table 1: Performance of the horizontally and vertically relative position features and their trajectory features, as well as comparison of corresponding features in HLPF.

Method	Accuracy (%)
Horizontally Relative Position Feature (HRP)	43.1
Vertically Relative Position Feature (VRP)	51.5
HRP Trajectory Feature (HRPT)	31.1
VRP Trajectory Feature (VRPT)	41.1
HRP+VRP+HRPT+VRPT (H&V)	**72.5**
Distance Relations, Orientation Relations and their Trajectory Features (D&O)	69.9

inputs. Features are assigned to their closest codeword to generate histograms. To represent each video sample, we directly concatenate histograms of different feature into a long vector. Training and testing splits are generated randomly. The only constraint is that, for each action category, the ratio of the number of clips in the training set and the testing set is close to $7:3$. Ten splits are randomly generated and the performance reported here is the average of the ten splits. For classification, we use SVM with RBF kernel to train classifiers.

5.2. Performance of horizontally and vertically relative position features

To verify the the effectiveness of the horizontally and vertically relative position features, as well as their trajectory features in Section 4.1, we compare them with corresponding features in HLPF, which are computed from pairs of joints, namely the distance relations and orientation relations features, as well as their trajectory features. Table 1 shows the performance of the features mentioned above. We can see that the performance of the vertical features (VRP and VRPT) is much better than the horizontal features (HRP and HRPT). This is because actions are usually performed when the body is vertical to the ground. By combining our novel features (H&V), we achieve 2.6% improvement when comparing with corresponding features (D&O) in HLPF.

5.3. Performance of joint inner product feature

To verify the the effectiveness of the joint inner product feature and its trajectory feature in Section 4.2, we compare them with corresponding features in HLPF, which are computed from triplets of joints, namely the angle relations feature and its trajectory feature. Table 2 shows the performance of aforementioned features. The combination of our two features (P) shows better performance than corresponding features (A) in HLPF.

From Table 1 and Table 2, we can conclude that the temporal features (HRPT, VRPT and JIPT) are not as distinctive as the corresponding spatial features (HRP, VRP and JIP), but they are indispensable. These results also indicate that our novel features (H&V and P) improve upon the baseline performance for corresponding features (D&O and A).

Table 2: Performance of the joint inner product feature and its trajectory feature, as well as comparison of corresponding features in HLPF.

Method	Accuracy (%)
Joint Inner Product Feature (JIP)	53.8
JIP Trajectory (JIPT)	44.3
JIP+JIPT(P)	**64.9**
Angle Relations and its Trajectory Feature (A)	60.0

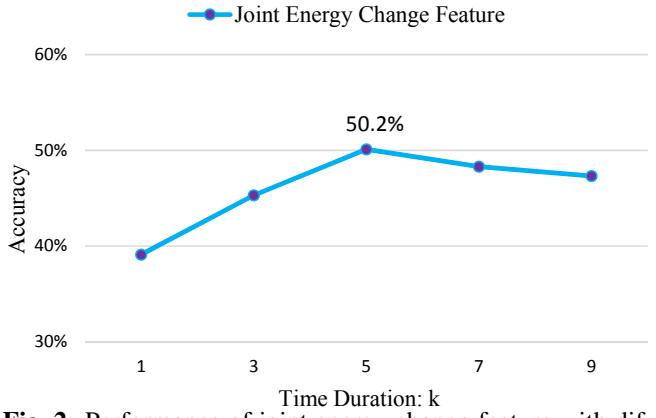


Fig. 2: Performance of joint energy change feature with different time duration k.

5.4. Performance of joint energy change feature

Figure 2 shows the performance of joint energy change feature with respect to the time duration k. It demonstrates that a suitable time duration ($k = 5$) results in the highest accuracy. Actually, for a small k, imperfect annotations may cause jittery trajectories, thus resulting in a lower performance. For a large k, the force during this period is changed, so that this feature no longer makes sense. Hereafter, the performance related to these two features is based on the setting that $k = 5$.

To show the complementarity of the joint energy change feature and HLPF, we compare them in Table 3. Results indicate that the JEC, which uses the spatio-temporal information, is an excellent supplement to the spatial features and the temporal features.

5.5. Comparison to the state-of-the-art

HLPF [1], dense trajectory features [10] encoded by Fisher vectors [11] (DT-FV) and posed-based CNN features [14] (P-CNN) are state-of-the-art methods for action recognition. We combine NHLPF and HLPF as improved HLPF (IHLPF). The comparisons of these methods are shown in Table 4. For P-CNN with DT-FV, we use the results reported in [14]. For HLPF, we use the publicly available code to compute fea-

Table 3: Comparison of HLPF and the combination of HLPF and JEC. JEC is the joint energy change feature.

Method	Accuracy(%)
HLPF [1]	76.0
HLPF + JEC	**78.2**

Table 4: State-of-the-art methods on the JHMDB dataset

Method	Accuracy(%)
HLPF [1]	76.0
P-CNN [14]	74.6
DT-FV [14]	65.9
NHLPF(ours)	**79.6**
P-CNN+DT-FV [14]	79.5
IHLPF(ours)	**80.4**

tures. It demonstrates that, for single methods, NHLPF improves upon the state-of-the-art methods by 3.6% on the JHMDB dataset. For combined methods, IHLPF improves unpon the state-of-the-art methods by 0.9%. Although the improvement is not obvious, the combined features we use are much simpler than both P-CNN and DT-FV. This is because P-CNN needs to train two CNNs, which require RGB image patches and flow patches around the joints as respective inputs. Besides, DT-FV has high computation complexity for computing descriptors around dense trajectories and for Fisher vector coding. Our method, however, only calls for the positions of joints to compute simple features. This manifests that our novel features are both efficient and effective.

A quantitative comparison per class is presented in Figure 3. It can be concluded that, NHLPF achieves large improvements over HLPF for actions that are difficult to distinguish, such as *run, walk* and *climb stairs*.

6. CONLUSION

In this paper, we propose a set of novel high-level pose features (NHLPF). The horizontally and vertically relative position features describe the relative position relations between pairs of joints, and solve the problem of distance invariance of adjacent joints. The joint inner product feature describes relative position relations among triplets of joints. The joint energy change feature combines spatial and temporal information to describe energy changes in joints due to the motion of other joints, and offers a new way to use spatio-temporal information. NHLPF improves the state-of-the-art on a benchmark dataset.

7. REFERENCES

[1] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," *ICCV*, pp. 3192–3199, 2013.

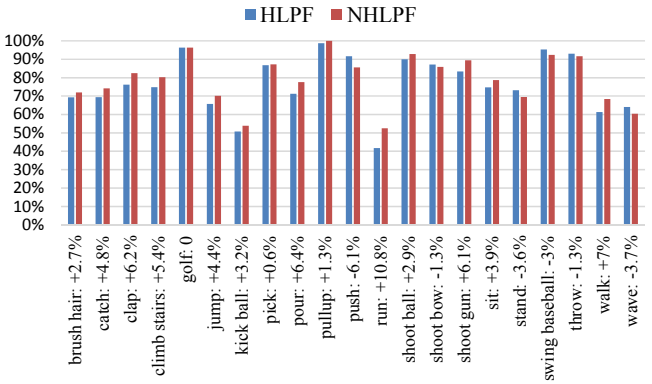


Fig. 3: Comparison of per class accuracy on JHMDB dataset between HLPF and NHLPF. Numbers correspond to the accuracy difference between NHLPF and HLPF (positive indicates that NHLPF performs better).

[2] I. Laptev and T. Lindeberg, "Space-time interest points," *ICCV*, pp. 107–123, 2003.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *VS-PETS*, pp. 65–72, 2005.

[4] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," *ECCV*, pp. 650–663, 2008.

[5] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," *CVPR*, pp. 1–8, 2008.

[6] A. Klser, M. Marszalek, and C. Schmid, "Learning realistic human actions from movies," *BMVC*, pp. 275:1–10, 2008.

[7] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," *ECCV*, pp. 650–663, 2008.

[8] R. Mattivi and L. Shao, "Robust spatio-temporal features for human action recognition," *MAPC*, pp. 351–367, 2011.

[9] H. Zhang, W. Zhou, C. Reardon, and L. E. Parker, "Simplex-based 3d spatio-temporal feature description for action recognition," *CVPR*, pp. 2067–2074, 2014.

[10] H. Wang and C. Schmid, "Action recognition with improved trajectories," *ICCV*, pp. 3551–3558, 2013.

[11] D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with fisher vectors on a compact feature set," *ICCV*, pp. 1817–1824, 2013.

[12] A. Yao, J. Gall, and L. Van Gool, "Coupled action recognition and pose estimation from multiple views," *IJCV*, pp. 16–37, 2012.

[13] V. K. Singh and R. Nevatia, "Action recognition in cluttered dynamic scenes using pose-specific part models," *ICCV*, pp. 113–120, 2011.

[14] G. Cheron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," *ICCV*, 2015.