# HUMAN ACTIVITY RECOGNITION BASED ON POSE POINTS SELECTION

*Ke Xu*[⋆] *Xinghao Jiang*[⋆†] *Tanfeng Sun*[⋆†]

[⋆]School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
[†]National Engineering Lab on Information Content Analysis Techniques, GT036001
Shanghai, China

## ABSTRACT

A novel method for human action recognition is proposed in this paper. Traditional spatial-temporal interest point detectors are easily affected by hair, face, shadow, clothes texture or the shake of camera. Inspired by the use of points distribution information, we propose a point selection method to select representative points (denoted by the "pose points"), which use HOG human detector and contour detector to select the points on human pose edges. The pose points carry both local gradient information and global pose information. 3D-SIFT scale selection method and novel descriptors called body scale and motion intensity feature are also studied. The descriptors calculate the width scale of different levels of human body and count motion intensity of activity in five directions. The descriptors combine spatial location with the moving intensity together and are used for further classification with SVMs. Experiments have been conducted on benchmark datasets and show better performance than previous methods, which achieved 99.1% on Weizmann dataset and 95.8% on KTH dataset.

***Index Terms***— Activity recognition, pose points, BOVW

## 1. INTRODUCTION

Human activities recognition is one of the most popular research topics in computer vision. It is very complex and difficult to recognize action automatically by computer due to the changes of viewpoints and the interference of background. In recent years, local based representations of human activity have been conducted on different datasets and made many outstanding achievements. In order to represent local features, many spatial-temporal detectors have been applied. Laptev proposed Harris3D detector[1], Mikolajczyk proposed Hessian[2] detector, and Bay proposed SURF[3] etc. Descriptors are studied to describe cube feature around the points. HOG/HOF[1] compute histograms of gradient and histograms of flow, 3D-SIFT[4] extended 2D-sift descriptor from static images to video sequences and SURF descriptor[3] provide comparable or even better results than SIFT while it can be calculated in a relatively efficient way.

Researchers have also proposed many well performed methods based on trajectories[5] or deep learning[6]. But these methods all face the same problem, i.e. the long training time due to the huge amount of tracking points and high computational cost. Traditional spatial-temporal interest points also have the following drawbacks. First, interest points detector are easily affected by background, texture and shadow, which makes many noisy descriptions when building the bag of visual words(BOVW). Second, interest points location information are missing when points are used as visual words, which actually contain important pose features.

In this paper, a novel spatial-temporal descriptor for human pose and motion representation is proposed. The descriptor contains three sub-descriptors, which are used to build BOVWs. The visual words in the BOVWs are further used to form feature histograms for support vector machine(SVM) training and classification. Inspired by[7] and [8], we construct a combined descriptor for each action including spatial-temporal multi-scale 3D-SIFT descriptor, body scale descriptor and motion intensity descriptor. Details of these descriptors will be explained in section 3 and section 4. Then with fusion training, statistic histograms are constructed for different activities. Finally, we train SVMs with the concatenated histograms for classification.

The rest of paper is organized as follows, Section 2 will elaborate the pose points extraction and selection methods. Section 3 will present an improved multi-scale 3D-SIFT descriptor applied on pose points and novel descriptors. The novel descriptors calculate relative width of body levels and human motion intensity. Section 4 explains our algorithm's framework. Section 5 will present the experimental setup and show the experimental results in comparison with some state-of-the-art approaches. Section 6 draws the conclusion.

## 2. POSE POINTS SELECTION

In recent years, spatial-temporal interest points are widely used in activity recognition. Traditional detectors tend to find

Xinghao Jiang is the corresponding author.

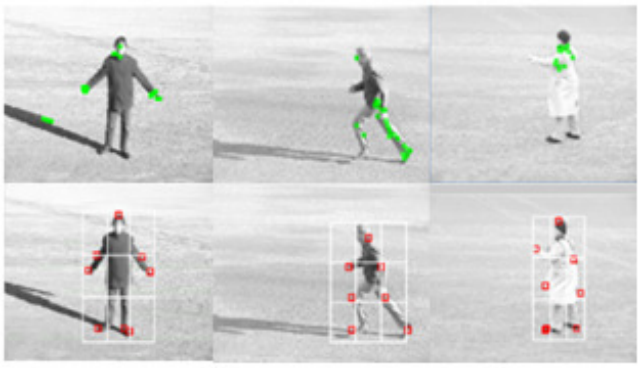


**Fig. 1**. The first row is traditional Harris3D interest points, which are marked with green color. The second row is the improved pose points, which are marked with red color.

points which have high response in both spatial region and temporal volumes. However, these detectors are easily affected by background, shadow, clothes texture, hair, face or the shake of camera (Figure 1).

In this paper, a novel points selection method to represent human pose and locations of moving parts is proposed. Inspired by the idea of [1] and [3], which usually extract points on corner or edges. In frame $t$, HOG human detector is first applied to locate body region. The body region height is $H_R$. Then Sobel contour detector is used in body region and the contour $C_t$ is stored. Hough transform is used to eliminate lines in the frame caused by shadow and background affection (Figure 2).

**Fig. 2**. Use Hough transform to eliminate shadow and background affection

Then human region is equally divided into $N$ levels in horizon. For each level $i$ ($i = 1$ to $N$), $l_i = H_R * i/N$ represents the level boundary value and is stored in vector $L$, $L = \{l_1, ..., l_N\}$ with $l_i < l_{i+1}$. Then as we proposed in (1) and (2), each point $P(x, y)$ in $C_t$ is tested and the points that have minimum $x$ and maximum $x$ are selected as pose points. Two pose points are extracted in each level and the total number for each frame is $2N$. Vector $EP$ store the pose points ordinates. Using pose points, selected interest points number is reduced down to twenty or less in each frame.

$$EP[2i-1] = min_x\{P(x,y)|P \in C_t, y \in (l_i, l_{i+1})\} \quad (1)$$

$$EP[2i] = max_x\{P(x,y)|P \in C_t, y \in (l_i, l_{i+1})\} \quad (2)$$

Pose points selection method is efficient as it selects points on human pose contour, these points can be used in tradition local descriptors and they also contain pose information in spatial domain.

## 3. FEATURE DESCRIPTIONS

In this section, an improved multi-scale 3D-SIFT descriptor is proposed to describe pose points feature. Body scale descriptor and motion intensity descriptor are proposed to describe pose feature and motion feature.

### 3.1. Multi-scale 3D-SIFT Descriptor

In this paper, an improved multi-scale 3D-SIFT descriptor is proposed to describe pose points in spatial and temporal domain. It calculates spatial gradient in eight directions and temporal gradient in four directions. Traditional 3D-SIFT[4] compute features in the given scale and the sub block sizes are always 2*2*2 or 4*4*4. For improvement, we use scales detected by SURF and change the sub block sizes used in 3D-SIFT automatically. For example, if an edge point locates in the points circle detected by SURF and the circle radius is $R/2$, then the histogram will be calculated with $\sigma_S = R$ and sub block width $R/2$. As in (3) and (4), $(x, y, t)$ represents the location of the pose point, and $(x^{'}, y^{'}, z^{'})$ represents the location of the pixel being added to the orientation histogram. $\omega$ is a normalize coefficient and $m_{3D}(x^{'}, y^{'}, z^{'})$ is the magnitude at $(x^{'}, y^{'}, z^{'})$. $\omega$ and $m_{3D}$ use the same definition in [4].

$$H(i_\theta, i_\phi) = H(i_\theta, i_\phi) + \frac{1}{\omega} m_{3D}(x^{'}, y^{'}, z^{'}) e^S \quad (3)$$

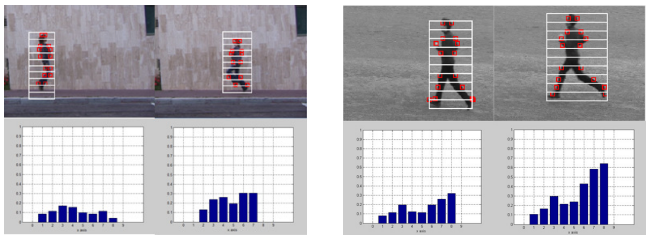
$$S = \frac{-(x - x^{'})^2 + (y - y^{'})^2 + (t - t^{'})^2}{2\sigma_S^2} \quad (4)$$

### 3.2. Body Scale Descriptor

After body levels $N$ and pose points $EP$ are detected, body scale features are extracted to represent the relative width of body parts. Body scale feature $H_{BS}$ extracts width from different levels of body and compute the width scale compared to human region width $H_W$. For each $i$ ($i = 1$ to $N$), the $i$ th value of $H_{BS}$ is defined as the distance between pose points in the $i$ th level.

$$H_{BS} = \{h_1, h_2 ... h_N\}, h_i = EP[2i] - EP[2i-1] \quad (5)$$

The advantages of this feature are the low computing cost and the effectiveness to similar actions like jump and skip or jogging and running, as shown in figure 3. Although jogging and running have similar width wavelets, their width scale have different values, running tend to have larger width and get higher scale between the maximum width and the minimum width in a gesture. The feature shows high discriminative power among actions.



(a) (b)

**Fig. 3**. (a)Body scale feature of jump(left) and skip(right). (b)Body scale feature of jogging(left) and running(right).

### 3.3. Motion Intensity Descriptor

After human moving region and the dense optical flow $(v_x, v_y)$ of moving parts at point $(x, y)$ is extracted, motion intensity vector is calculated as follows. First, divide human region into five directions with equal angle $D = \{d_1, d_2, d_3, d_4, d_5\}$. Then flow angle $\theta(x, y)$ and magnitude $M(x, y)$ are computed for each pixel in the region. Finally, if $\theta(x, y) \in d_i$ add up magnitude to direction $i$ to form a five dimension intensity vector $H_{MI} = \{h_1, h_2, h_3, h_4, h_5\}$ as shown in figure 4 and formula (6) to (8). This intensity description shows velocity difference among actions and can be used for later classification.

$$\theta(x, y) = tan^{-1}(v_y/v_x) \tag{6}$$

$$M(x, y) = \sqrt{v_x^2 + v_y^2} \tag{7}$$

$$h(d_i) += M(x, y),\ if\ \theta(x, y) \in d_i \tag{8}$$


**Fig. 4**. Process of calculating motion intensity feature.

## 4. ALGORITHM FRAMEWORK

In this section, our algorithm framework is built for dataset training and action classification, the procedure is described in the following steps, as shown in figure 5.

Step 1. Given an input video, HOG human detector is first applied on each frame to locate human region. Then SURF points are detected in the human region and sizes of human region, location of SURF points and points scales are stored. Next we detect Farneback optical flow and Sobel human contour in human region, with Hough transform to eliminate noise caused by background.

Step 2. Divide human region into levels and follow the pose points selection algorithm to detect pose points.

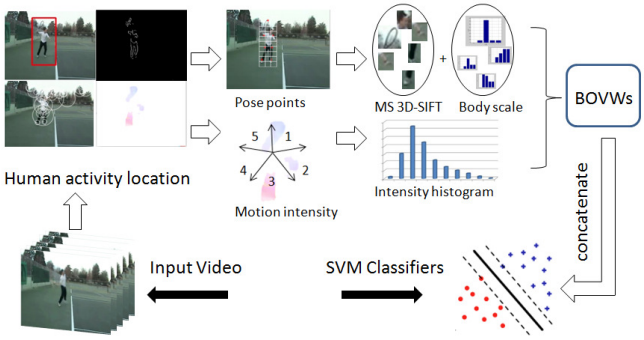



**Fig. 5**. Process of recognition, which contains pose points detection, points feature extraction, BOVWs building, statistical feature concatenation and classification .

Step 3. Multi-scale 3D-SIFT descriptors are used on pose points to construct a points bag of visual words(p-BOVWs) with K-means cluster method. Body scale feature is extracted to represent the relative width of body parts. Motion intensity is computed from optical flow, which represent the motion intensity of body parts. Different combination of words represent different actions.

Step 4. Building BOVWs(c-BOVWs) for concatenated body scale feature and Motion intensity vector with K-means cluster method. For each action volume, a statistical histogram of words is represented. The histograms are then used for classifier training. However, an activity volume may contains repeated sequences which is redundant and increase confusion rate for recognition. So we apply fusion training on body scale feature and motion intensity feature before K-means clustering, which means fusion the similar features in one activity volume to reduce computational cost and repetitiveness of visual words. After fusion, repeated gestures and motion patterns are removed and each activity volume contains only one entire action process.

Step 5. Combine p-BOVWs and c-BOVWs statistical histograms to form a vector and use SVMs for classification.

With the trained SVM classifier and the pre-set action label, the activity in an unlabeled video is able to be recognized as a specific action.

## 5. EXPERIMENT AND ANALYSIS

### 5.1. Datasets

In this paper, we employ Weizmann action dataset[9] and KTH dataset[10]. Weizmann dataset contains 92 videos of nine people performing following 10 actions: running, walking, skipping, jumping-jacks, jumping forward on two legs, jumping in place on two legs, jumping sideways, waving with two hands and waving with one hand. Each clip lasts about 2 seconds at 25Hz with frame size of 180*144 pixels. Sample

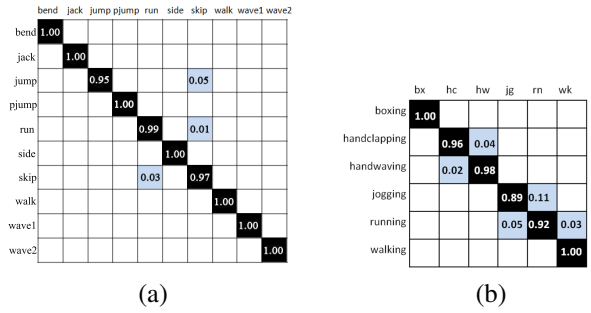
videos are also collected from the KTH human action dataset. It contain 6 types of human actions (walking, jogging, running, boxing, hand-waving and hand-clapping) performed by 25 actors with four different scenarios: S1 (outdoors), S2 (outdoors with scale variation), S3 (outdoors with different clothes), and S4 (indoors). There are totally 599 video clips with image frame size of 160*120 pixels.

### 5.2. Experiment Settings

In this section, we explain the experiment settings. In the experiment, Harris3D detector scales set to $\sigma = 2$ and $\tau = 2$. For both KTH dataset and Weizmann dataset, body scale feature is calculated in 10 levels and motion intensity is calculated in 5 directions. When building BOVWs, Multi-scale 3D-SIFT feature is calculated with K-means cluster center number set to 200, and concatenated vectors of body scale feature and motion intensity feature are calculated with K-means cluster center number set to 100. So the final training vector dimension is 300. For classification, we use a non-linear support vector machine(SVM) with a multi-channel $\chi^2$ kernel that robustly combines channels.

### 5.3. Classification Accuracy Experiment

In Weizmann dataset, following the Leave-one-out split method, each time we use 8 actors as training data and 1 actor as test data. Repeat the procedure 9 times, so each actor is used as both training data and test data. The experiment is repeated 20 times to compute an average confusion matrix, the average accuracy is 99.1%, which is shown in figure 6. There are still some confusions between jump, skip and run, due to the randomness of cluster center selection.




(a) (b)

**Fig. 6**. (a)Confusion matrix of Weizmann dataset, average the accuracy is 99.1%.(b)Confusion matrix of KTH dataset, average the accuracy is 95.83%.

In KTH dataset, using Leave-one-out split method, each time we use 24 actors as training data and 1 actor as test data. Repeat the procedure 25 times for each actors. The experiment is repeated 20 times to compute an average confusion matrix, the average accuracy is 95.83%, which is also shown in figure 6. Confusions mainly occurred between clapping and waving as well as jogging and running because some action remain insufficient frames after fusion and human movement viewpoint changes may affect body scale feature.

### 5.4. Comparison Experiment

The comparison experiment results is shown in in table 1, where [4] uses original 3D-SIFT feature, [7] and [11] uses local spatial-temporal feature, [8] use both spatial- temporal feature and global distribution of points. Our method uses multi-scale 3D-SIFT feature in points description, body scale feature and motion intensity in pose and motion description. It shows a better accuracy compared to previous spatial-temporal interest points methods, which achieves an accuracy of 95.8% on KTH dataset and 99.1% on Weizmann dataset.

**Table 1**. COMPARATIVE RESULT

| Method | KTH | Weizmann |
|---|---|---|
| **Our approach** | **95.8%** | **99.1%** |
| Bregonzio et al.[8] | 93.17% | 96.66% |
| Lu et al.[7] | 91.5% | 93.5% |
| Zhang et al.[11] | 91.33% | 92.89% |
| Scovanner et al. [4] | 82.6% | - |

## 6. CONCLUSION

In this paper, a point selection method is proposed, which use human detector and calculate pose contour to locate points that carry spatial pose information. This method reduces the interesting points number and shows effectiveness in experiment. Novel descriptors on spatial domain and temporal domain have also been proposed in our paper. An improved multi-scale 3D-SIFT descriptor is used to describe pose points feature, which use the scales detected by SURF detector. Body scale feature and motion intensity feature are also proposed in our paper. The descriptors present the body scale information of pose and the motion intensity information for action with low dimension and low computing cost. In the experiments, K-means clustering and SVMs classifier are used. The experiments are performed on Weizmann dataset and KTH human action dataset. The proposed approachs achieve an accuracy of 95.8% on KTH dataset and 99.1% on Weizmann dataset. The results show the effectiveness of our algorithm and our improvements compared to previous methods.

## 7. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (No. 61272439, 61272249), the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20120073110053).

## 8. REFERENCES


[1] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[2] Krystian Mikolajczyk and Cordelia Schmid, "Scale & affine invariant interest point detectors," *International journal of computer vision*, vol. 60, no. 1, pp. 63–86, 2004.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006*, pp. 404–417. Springer, 2006.

[4] Paul Scovanner, Saad Ali, and Mubarak Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 357–360.

[5] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3551–3558.

[6] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3d convolutional neural networks for human action recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 221–231, 2013.

[7] Mengmeng Lu and Liang Zhang, "Action recognition by fusing spatial-temporal appearance and the local distribution of interest points," in *2014 International Conference on Future Computer and Communication Engineering (ICFCCE 2014)*. Atlantis Press, 2014.

[8] Matteo Bregonzio, Shaogang Gong, and Tao Xiang, "Recognising action as clouds of space-time interest points," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1948–1955.

[9] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, December 2007.

[10] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. IEEE, 2004, vol. 3, pp. 32–36.

[11] Ziming Zhang, Yiqun Hu, Syin Chan, and Liang-Tien Chia, "Motion context: A new representation for human action recognition," in *Computer Vision–ECCV 2008*, pp. 817–829. Springer, 2008.