Human Action Recognition Using Key Points Displacement

Kuan-Ting Lai[1,2], Chaur-Heh Hsieh[3], Mao-Fu Lai[4], and Ming-Syan Chen[1,2]

[1] Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2] National Taiwan University, Taipei, Taiwan, R.O.C.
[3] Ming-Chuan University, Taoyuan, Taiwan, R.O.C.
[4] Tungnan University, Taipei, Taiwan, R.O.C.
ktlai@arbor.ee.ntu.edu.tw
http://arbor.ee.ntu.edu.tw/

Abstract. Recognizing human actions is currently one of the most active research topics. Efros et al. first proposed using optical flow and normalized correlation to recognize distant actions. One weakness of the method is that optical flow is too noisy and cannot reveal the true motions; the other popular method is the space-time-interest-points proposed by Laptev et al., who extended the Harris corner detector to temporal domain. Inspired by the two methods, we proposed a new algorithm based on displacement of Lowe's scale-invariant key points to detect motions. The vectors of matched key points are calculated as weighted orientation histograms and then classified by SVM. Experimental results demonstrate that the proposed motion descriptor is effective on recognizing both general and sport actions.

Keywords: SIFT, Action Recognition, Optical Flow, Space-time-interest-points, SVM.

1 Introduction

Recognizing Human actions attracts lots of research interests in recent years. Many applications, such as video content analysis, video surveillance, remote home care, etc., are based on action recognition technology. Human actions are diversified and thus hard to be classified. I. Laptev et al. [1] proposed to use space-time-interest-points to recognize human actions. The authors extended the Harris corner detector to find the spatio-temporal corners. Based on the 3D Harris detector, C. Shuldt et al. [2] applied SVM to classify actions, and established a large action database including six actions and 2391 sequences to evaluate the recognition performance. Since then the KTH action database has become one of the most important action database and widely adopted as evaluation benchmark [2-5, 9, 14, 16, 19]. The space-time-interest-points method is elegant and effective. However, under certain circumstances, the spatio-temporal corners are rare and not enough for some recognition tasks. To alleviate this problem, P. Dollar et al. [3] chose to apply a quadrature pair of 1D Gabor filters in temporal dimension and developed the Cuboid detector.

A. Elmoataz et al. (Eds.): ICISP 2010, LNCS 6134, pp. 439–447, 2010.

The previous methods did not exploit the implicit semantic information between different actions. To improve the recognition rate, J. C. Niebles et al. [4] applied Probabilistic Latent Semantic Analysis (pLSA) model with "bag of video words" representation. They employed Dollar's method to extract action features. Wong et al. [5] further enhanced the recognition rate by introducing geometric information into the pLSA model. The reported recognition rate is impressive, but robust actor's centroid information is required and was extracted manually by the authors.

As mentioned above, most methods take strong temporal responses as action features, which are usually periodic motions. However, in real world, human actions are not so regular. In this paper, we propose a new temporal feature extraction method based on scale-invariant key points tracking, which is named as SIFT displacement. The idea was inspired by both Efros's optical flow and Laptev's space-time-interest-points methods. The goal is intuitive: extracting the human body parts' motions frame-by-frame. D. G. Lowe's [6] scale invariant feature transform (SIFT) is used to extract actors' features. The vectors of matched key points are surpassed as weighted orientation histograms and then classified by support vector machine (SVM). Experiments have been conducted on KTH as well as a new tennis action database. The results demonstrate that performances are promising on both general and sport actions.

2 Previous Works

During the past years, the sparse feature representation of local image has been extensively studied and widely adopted in many applications, e.g. object recognition, image stitching, and video tracking, to name a few [6-8]. Ivan Laptev et al. [1] first extended the Harris corner detector into temporal domain and proposed the space-time-interest-points method. Mathematical forms of the method are defined below. For a spatio-temporal image sequence *f*, its scale-space representation is defined as the convolution of *f* with *g*:

$$L(x, y, t; \sigma_l^2, \tau_l^2) = g(x, y, t; \sigma_l^2, \tau_l^2) * f(.) \tag{1}$$

where *g* is the spatio-temporal separable Gaussian kernel defined in (2):

$$g(x, y, t; \sigma_l^2, \tau_l^2) = \frac{\exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/2\tau_l^2)}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}} \tag{2}$$

The final form is a 3-by-3 matrix composed of first order spatial and temporal derivatives averaged with a Gaussian weighting function. As shown in (2), Laptev's method assumed that actions have fixed time spans. To overcome this problem, Dollar et al. [3] developed a separable linear filter in temporal domain using two quadrature pair of 1D Gabor filters.

The other popular method is optical flow based motion descriptor proposed by Efros et al. [10]. The authors treat these optical flow features not as precise pixel displacements at points, but simply as a spatial pattern of noisy measurement. The noisy feature vectors are carefully smoothed and aggregated to form a motion descriptor. This method is best suitable for small human figures around 30 pixels.

On the other hand, the human silhouette also provides important information for action recognition and attracted many research interests [20-22]. A well combination of shape and temporal information can generate a perfect recognition result on Weizmann dataset [21]. However, sometimes it is hard to extract robust shape information in scenes with complex backgrounds.

Recently, Wang et al. [19] has conducted a thorough performance evaluation of three space-time-interest-points detectors and six descriptors along with their combinations on three datasets. All experiments are reported for the same bag-of-features SVM recognition framework. One important discovery is that dense sampling consistently outperforms all tested interest point detectors in realistic video settings. However, dense sampling produces a very large number of features and hard to handle in real-time. With regard to performance of interest point detectors, Laptev's Harris 3D performs better on KTH, while the Cuboid detector gives better results for UCF and Hollywood2 datasets.

The method proposed in this paper aims to concentrate on human body motions between consecutive frames. The scale-invariant feature transform (SIFT) is applied to extract essential motions of actors. SIFT has been applied in action recognition task, e.g. 3-D SIFT [15, 16], but our method is different in spirit. The details of SIFT displacement are described in next section.

3 Proposed Method

The fundamental step of our method is to detect key points of actor's body. The first successful local interest points detector is developed by Harris [13] for efficient motion tracking. The Harris corner detector is very sensitive to changes in image scale. David G. Lowe [6] extended the local feature approach to achieve scale invariance. As a result, SIFT is selected to detect local features, and its scale-invariant ability provides more robust tracking.

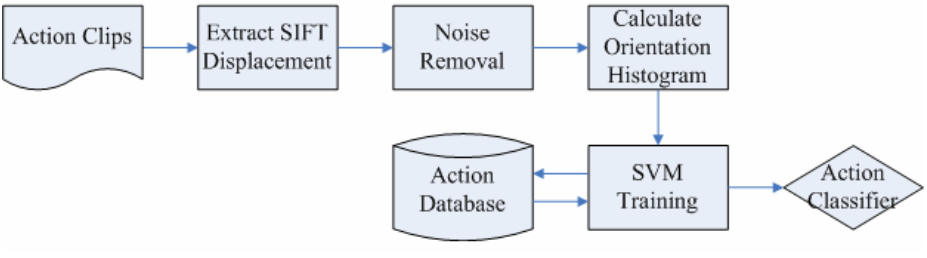

Fig. 1. Flowchart of the proposed action recognition method

Figure 1 shows the flowchart of the proposed algorithm. Initially, the action videos are segmented into action clips. The length of the clips depends on the action types. The next step is to extract SIFT feature points and match the key points in two consecutive frames. The movement between the paired keys is calculated as the SIFT displacement. Large erroneous matched keys need to be removed since the effects of false detections are magnified by large weightings. To increase accuracy, we remove

the outliers around two standard deviations larger than the mean vector length. The max key length is 10 pixels for running action, and 6 pixels for other actions in all experiments. After the noise are removed, the SIFT displacement vectors are quantized as several bins, weighted by vectors' lengths, and further summarized as histogram. Finally, SVM is used to learn and recognize the action features.

The SIFT method is widely used in recognizing rigid objects, but seldom applied to deformable human body. The reason is that human limb movements would change distinctive features. In our experiments, the key points of actor's body are indeed hard to be matched beyond a certain time period (>150 ms). Nevertheless, SIFT matching works well between consecutive frames in common broadcast videos (25 fps).

There are four major stages in calculating SIFT: scale-space extrema detection, key point localization, orientation assignment, and key point descriptor. The scale space function of an image is defined as $L(x,y,\sigma)$, which results from the convolution of a variable-scale Gaussian kernel $G(x,y,\sigma)$ with an input image $I(x,y)$.

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{3}$$

Lowe proposed to use a difference-of-Gaussian function to identify potential interest points that are invariant to scale and orientation changes. The candidate locations are filtered by testing their stability.

The magnitude and orientation of a key point are calculated in (4) and (5):

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \tag{4}$$

$$\theta(x, y) = \tan^{-1}((L(x, y+1) - L(x, y-1)) / (L(x+1, y) - L(x-1, y)) \tag{5}$$

where L is a Gaussian smoothed image with closest scale of the key point. An orientation histogram is formed from the gradient orientations of sample points within a region around the key point. Peak in the orientation histogram is assigned as the direction.

The nth matched key points at frame t and frame $t+1$ are defined as $k_n(x_t, y_t, t)$ and $k_n(x_{t+1}, y_{t+1}, t+1)$, respectively. The nth SIFT displacement's magnitude and angle are defined in Equation (6) and (7):

$$M_n = \sqrt{(x_{t+1}^n - x_t^n)^2 + (y_{t+1}^n - y_t^n)^2} \tag{6}$$

$$\theta_n = \tan^{-1}\left(\frac{y_{t+1}^n - y_t^n}{x_{t+1}^n - x_t^n}\right) \tag{7}$$

In evaluation of KTH action database, the whole video frame is divided into 3 x 3 regions, and the SIFT displacement vector orientation is quantized into 8 bins. The magnitude of a SIFT displacement is used as the weight while counting the orientation histogram. Several feature extraction results are shown in Fig. 2 and Fig. 3. The arrows represent the SIFT displacement vectors. As we can see, the motion descriptors correctly reveal the movements of different parts of the actor. A few erroneous key points are caused by the shadow of the actor.

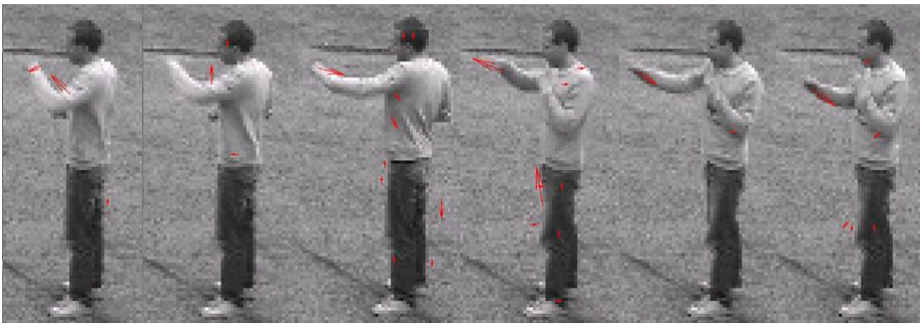

Fig. 2. Boxing action in KTH database; the arrows show the magnitude and orientation of motions

Fig. 3. Hand waving action in KTH database; most large SIFT displacement vectors are detected around the moving hands

4 Experimental Results

The proposed method was tested on both KTH as well as a new tennis action database established by the authors. We utilize the SIFT extraction source code developed by Rob Hess [17], and the SVM training tools provided by Chang and Lin [18]. The KTH dataset provides 25 people, six different actions (boxing, hand-clapping, hand-waving, jogging, running, walking), and four scenarios (D1: static background, D2: scale-variation, D3: different clothes, D4: lighting variation). The five-subset cross validation method is employed to evaluate the recognition rate. Two important parameters, orientation bin number and SIFT key points matching threshold, are scanned to test the performance. As shown in Fig. 4, the proposed method is quite stable and the recognition rates range between 75% and 79%. The best recognition rate is 78.67% (472/600) for all data and 81.78% (368/450) for all data except D2 dataset. The confusion matrices are shown in Table 1. The reason for the accuracy decrease caused by D2 dataset is that, despite the correct matching of key points, the lengths and orientations of SIFT displacements in D2 dataset are distorted by the camera zoom-in and zoom-out processes.

The SIFT key points matching threshold is the distance ratio from the closest neighbor to the distance of the second closest in the Best-Bin-First algorithm [6]. In short, there are more the matched keys as the ratio becomes higher, and also more noisy and erroneous pair of key points. There is a tradeoff between quantity and quality of SIFT displacement vectors.

In comparison with other methods, the accuracy rates are reported as 63% by Ke et al. [14], and 71.72% by KTH authors [10]. P. Dollar et al. [3] selected only D1 and D3 datasets and achieved 81.16% recognition rate. Since our contribution is mainly on developing a novel spatio-temporal feature extraction method, only the results of the methods without semantic models and bag-of-features are compared here.

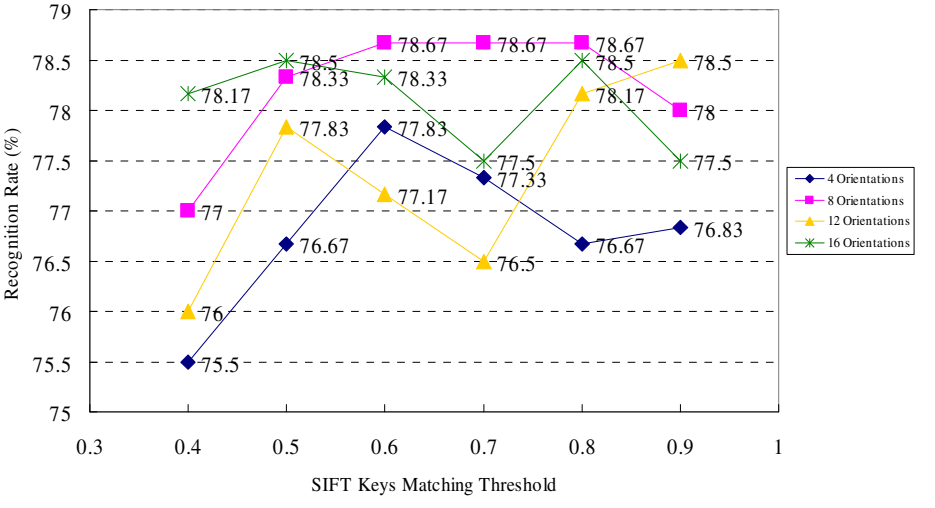

Fig. 4. The experimental results of KTH database with different SIFT key points matching thresholds and orientation bin numbers (4, 8, 12, 16)

Table 1. The left one is the confusion matrix of KTH database, and the right one is the confusion matrix without the scale variation dataset (D2)

KTH All	Boxing	Clapping	Waving	Jogging	Running	Walking
Boxing	0.82	0.09	0.08	0	0	0
Clapping	0.13	0.87	0.14	0	0	0
Waving	0.02	0.03	0.78	0	0	0
Jogging	0.01	0	0	0.68	0.13	0.16
Running	0.01	0.01	0	0.12	0.74	0.02
Walking	0.01	0	0	0.2	0.13	0.82

D1, D3, D4	Boxing	Clapping	Waving	Jogging	Running	Walking
Boxing	0.73	0.05	0.05	0	0	0
Clapping	0.24	0.93	0.16	0	0	0
Waving	0	0.01	0.79	0	0	0
Jogging	0	0	0	0.72	0.09	0.08
Running	0.01	0	0	0.15	0.89	0.08
Walking	0.01	0	0	0.13	0.01	0.84

In addition to KTH database, we evaluated our algorithm on a new tennis action database, which includes 60 actions extracted from 4 grand slams tournaments. The actions clips contain 11 “run left”, 12 “run right”, 16 “backhand stroke” (left-swing), 10 “forehand stroke” (right-swing) and 11 “serve” actions. To handle the large

deformations of the player body during fierce competitions, we select a short action time period (10 frames). The recognition rate of [Run Left, Run Right, Forehand, Backhand, Serve] are [0.82, 0.8, 0.81, 0.7, 0.82] individually.

Fig. 5. The player in the upper row performs the backhand-stroke action; while the player in lower row performs the serve action

5 Conclusions

In this paper, we presented a new motion descriptor based on SIFT displacement. The SIFT displacement vectors are further calculated as weighted orientation histogram and then classified by SVM. The proposed method has been tested on both general and sport actions. Experimental results are convincing and comparable to other state-of-the-art temporal feature extraction algorithms.

Several techniques can be exploited to improve recognition rate in the future. One possible method is key point tracking. The SIFT displacement method only considers the motions between two frames. However, the key points existing in several frames may contain more information than those only existing in two frames. In our preliminary experiment, a list of matched key points is kept in memory, and updated at every frame. The key points tracking results are shown in Fig. 6. The waving hands are successfully tracked while some unmoved keys are tracked in the body. In terms of running action, the best tracking points are at the actor's head. The other candidate method is to apply the bag-of-features framework [19]. More research efforts are required to find the best usage of the tracking information as well as exploring the full potential of SIFT displacement.

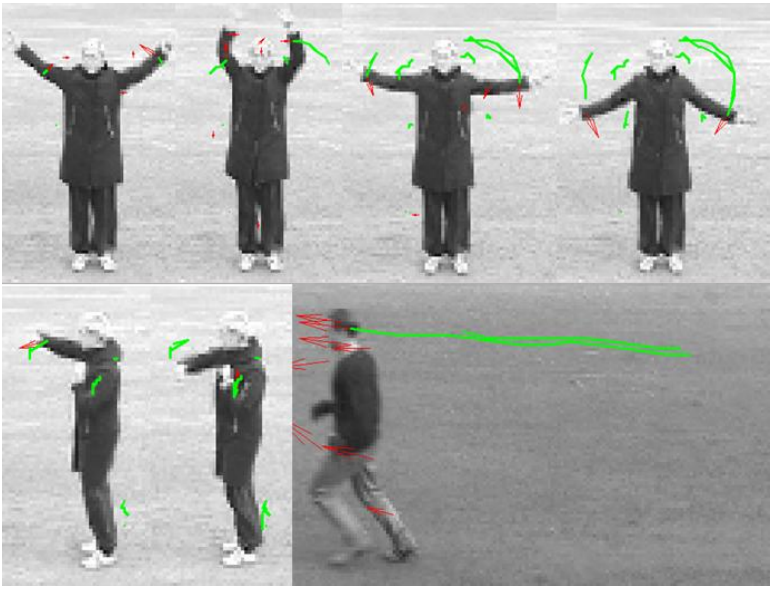

Fig. 6. Key points tracking results of hand-waving, boxing, and running. The green lines are the tracking trajectory, and the red arrows are SIFT displacements.

References

1. Laptev, I.: On space-time interest points. IJCV 64(2-3), 107–123 (2005)
2. Schuldt, C., Laptev, I., Caputo, B.: Recognizing Human Actions: A Local SVM Approach. In: Proc. ICPR (2004)
3. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: ICCV workshop: VS-PETS (2005)
4. Niebles, J.C., Wang, H.C., Li, F.F.: Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. In: Proc. BMCV (2006)
5. Wong, S.-F., Kim, T., Cipolla, R.: Learning Motion Categories using both Semantics and Structural Information. In: CVPR (2007)
6. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal on Computer Vision 60(2), 91–110 (2004)
7. Harris, C., Stephens, M.J.: A combined corner and edge detector. In: Alvey Vision Conference, pp. 147–152 (1988)
8. Schmid, C., Mohr, R., Bauckhage., C.: Evaluation of interest point detectors. IJCV 37(2), 151–172 (2000)
9. Liu, J., Shah, M.: Learning human actions via information maximization. In: CVPR (2008)
10. Efros, A.A., Berg, A., Mori, G., Malik, J.: Recognizing Action at a Distance. In: International Conference on Computer Vision (2003)
11. Hofmann, T.: Probabilistic Latent Semantic Indexing. In: Proceedings of the Twenty-Second Annual. International SIGIR Conference on Research and Development in Information Retrieval, SIGIR-99 (1999)
12. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003) doi:10.1162/jmlr.2003.3.4-5.993

13. Harris, C., Stephens, M.J.: A combined corner and edge detector. In: Alvey Vision Conference, pp. 147–152 (1988)
14. Ke, Y., Sukthankar, R., Hebert, M.: Efficient Visual Event Detection using Volumetric Features. In: International Conference on Computer Vision, pp. 166–173 (2005)
15. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional SIFT descriptor and its application to action recognition. In: Proceedings of the 15th International Conference on Multimedia, pp. 357–360 (2007)
16. Sun, X.H., Chen, M.Y.: Action Recognition via Local Descriptors and Holistic Features. In: CVPR (2009)
17. Hess, R.: SIFT Feature Detector, `http://web.engr.oregonstate.edu/~hess/`
18. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001)
19. Wang, H., Ullah, M.M., Klaser, A., Laptev, I.: Evaluation of Local Spatio-temporal Features for Action Recognition. In: CVPR (2009)
20. Bobick, A., Davis, J.: The Representation and Recognition of Action Using Temporal Templates. PAMI 23(3), 257–267 (2001)
21. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as Space-Time Shapes. IEEE PAMI 29(12), 2247–2253 (2007)
22. Yilmaz, A., Shah, M.: A Differential Geometric Approach to Representing the Human Actions. Comput. Vis. Image Und. 109(3), 335–351 (2008)