SOM-based Human Action Recognition Using Local Feature Descriptor CHOG3D

Ji, Yanli
Department of Advanced Information Technology, Graduate Student

Shimada, Atsushi
Department of Advanced Information Technology

Nagahara, Hajime
Department of Advanced Information Technology, Graduate Student | Department of Advanced Information Technology

Taniguchi, Rin-ichiro
Department of Advanced Information Technology

https://doi.org/10.15017/21946

出版情報：九州大学大学院システム情報科学紀要. 17 (1), pp.1-8, 2012-05-25. 九州大学大学院システム情報科学研究院
バージョン：
権利関係：

九州大学大学院
システム情報科学紀要
第 17 巻　第 1 号　平成 24 年 5 月

Research Reports on Information Science and
Electrical Engineering of Kyushu University
Vol.17, No.1, May 2012

SOM-based Human Action Recognition Using Local Feature Descriptor CHOG3D

Yanli JI*, Atsushi SHIMADA**, Hajime NAGAHARA** and Rin-ichiro TANIGUCHI**

(Received April 13, 2012)

Abstract: Human action recognition is applied in a wide field, such as video surveillance, intelligent interface, and intelligent robots. However, since various action classes, complex surrounding, interaction with objects, et al., it is still a complex problem to be solved. In this paper, we propose a method combining the Self-Organizing Map(SOM) and the classifier k-Nearest Neighbor algorithm (k-NN) to recognize human actions. We represent human actions in the form of local features using a compact descriptor, a histogram of oriented gradient in spatio-temporal 3D space(CHOG3D), which was proposed by us in the paper 1). Then we adopt SOM for feature training to extract key features of action information. With these key features, we adopt k-NN for action recognition. In our experiments, we test the optimal map size of SOM and the proper value k of k-NN for correct recognition. Our method is tested on KTH, Weizmann and UCF datasets, and results certify its efficiency.

Keywords: Human action recognition, CHOG3D, SOM, k-NN

1. Introduction

Human action recognition has a wide range of promising applications, e.g., video surveillance, intelligent interface, and intelligent robots. Many researchers have devoted great enthusiasm on it, and achievements are applied in our real life gradually. However, since there are various action classes, different action characteristics of human and complex surrounding, human action recognition becomes a complex problem. In recent years, a lot of researchers proposed algorithms to overcome these difficulties, and a serials of methods are produced. However, it is still the primary step of human action recognition.

Among the processing of action recognition, action representation is the primary and very important step. There are mainly two kinds of features for action representation, global features and local features. Since local spatio-temporal features are usually robust against illumination, cluster, and viewpoint changes, several local features were proposed to represent human actions. Laptev and Lindeberg et. al[2] proposed the Space-Time Interest Points (STIP) for a sparse representation of human action video data. In order to detect interest points in a space-time volume in their method, the author extended the detectors proposed in paper 3), and it was commonly used as "Harris3D." Following that, Schuldt[4] adopted STIP to represent human actions on a new database called "KTH actions dataset." Later, Dollar et al.[5] proposed a new spatio-temporal feature which calculated a vector of brightness gradients in a small 3-D volume called "cuboid" to describe the feature points. In addition, Scovanner et al.[6] designed the 3-D version of the SIFT descriptor, which was similar to the cuboid features[5]. Laptev[7] adopted histograms of oriented gradient(HoG) and histograms of optic flow(HoF) to describe human actions. Klaser[8] proposed a descriptor named as histograms of oriented 3D spatio-temporal gradients (HOG3D), which projected gradient into faces of an icosahedron. Willems[9] extracted scale-invariant spatio-temporal interest points which was called "Hessian detector." Then the author extented SURF[10] descriptor to describe the extracted interest points, named as "ESURF. "

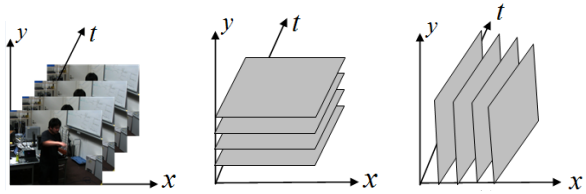
To test the performance of these descriptors, Wang[11] gave an evaluation on currently used detectors and descriptors through classifying human activities by SVM. All of the methods introduced above are included in the evaluation. It showed that the combination of detector Harris3D[2] and descriptor HoF, the combination of detector Harris3D and descriptor HoG/HoF, and the combination of detector Cuboids and descriptor HOG3D performed better than other algorithms. The average recognition rates of the three combinations on KTH dataset are 92.1%, 91.8% and 90.0%, respectively. Following that, Klaser updated the average recognition ratio of the combination of Harris3D and HOG3D in KTH dataset to be 92.4% in his thesis[12]. As the above results indicated, the descriptor HOG3D performed better among currently used local features for action recognition. However, descriptor HOG3D employed a complex procedure to calculate gradient, and it used a vector with more than 1000 elements to describe one feature

*Department of Advanced Information Technology, Graduate Student
**Department of Advanced Information Technology

point. It is a huge length for local feature description, and the length of descriptor increase the difficulty to distinguish two vectors correctly. Except that, the calculation cost of the descriptor HOG3D is high. To overcome these disadvantages, we proposed a descriptor which was a compact histogram of orientation gradient in spatio-temporal space(CHOG3D)[1] to represent human actions. In this paper, we employ the descriptor CHOG3D for action description.

For action recognition, various of statistical models, linear/nonlinear classification models, neuron network et al.. Among them, SOM has the ability of providing a low dimensional view on high dimensional data samples, and it has the advantage of handling complex situations. Therefore, it is suitable to solve the problem of human action recognition with large quantity of local features. In previous researches, Shimada et al.[13] proposed a Hierarchical Self-Organizing Map(HSOM) to recognize time-series gesture patterns which is represented by five 3D spatial coordinates of human body, head, two hands and two feet. In their method, SOM is used to train time-series patterns into three level features, posture, gesture elements and gestures, and they were arranged in SOM layer hierarchically. Using the sparse code in the bottom layer, they realized gesture recognition. They also researched into early recognition of gestures by SOM, and showed the good performance of SOM on dimension reduction and critical information extraction. In addition, Huang et al.[14] employed SOM for human action sequences recognition. By mapping action poses in a SOM, they represented a human action sequence by a trajectory of map units. Then a longest common subsequence algorithm was utilized to match action trajectories on the map robustly. Their approach was tested on Weizmann dataset and achieved promising results. These approaches showed the potential of SOM on action recognition. Therefore, we also employed SOM to train local features and obtain critical features for action recognition.

In this paper, we propose a SOM based system for local feature based human action recognition. In the proposed system, the compact descriptor CHOG3D is adopted for local feature calculation to represent human actions. Then the self-organizing map (SOM) is employed to train local features and extract key features of actions because of its advantage in mapping data into low dimension. After training, the key features are assigned action labels of the training data. For action recognition, we adopt k-NN to classify features of a testing action sequence into different action classes. According to calculating the statistics of feature classification, the action class of the testing sequence is determined. The method is tested on three datasets, KTH[4], Weizmann[15] and UCF sports[16], respectively. Furthermore, we test the optimal map size of SOM for in training and the proper value k for k-NN classification. With these optimal parameters, we perform action recognition on the three datasets, and results certify the efficiency of our method.



(a) xy channel (b) xt channel (c) yt channel

Fig. 1 Spatio-temporal channels.

The remaining parts of this paper are organized as follows. The descriptor CHOG3D is introduced in section 2. Then the detailed procedure of learning and recognizing by SOM and k-NN using local features are shown in section 3. The datasets adopted in this paper is introduced in section 4. In section 5, we shows the optimal parameter searching in our system and the experiment results obtained with these parameters.

2. Descriptor CHOG3D

To obtain spatio-temporal information of actions, we modified the FAST detector[17] to detect spatio-temporal interesting points in three dimensions. The detected points are described by the compact descriptor CHOG3D using the oriented gradient information in a spatio-temporal neighbor of the points.

2.1 Spatio-temporal FAST detection

Features from Accelerated Segment Test(FAST) was first reported by Edward R. and Tom D. in 2006, and it was modified to have more repeatability later. It is a faster and more stable feature detector comparing with other detectors. In this paper, FAST is extended to spatio-temporal space for action feature detection. As the **Fig. 1** shows, one video can be regarded as xyt spatio-temporal space, where x and y refer to spatial dimensions and t refers to temporal dimension. We detect spatio-temporal features on planes of xy, xt, yt channels(**Fig. 1(a)(b)(c)**), and integrate them to be feature points with x, y, t, three dimensions.

2.2 Descriptor calculation

The compact descriptor is calculated in a support region of a feature point with the size of $4 \times 4 \times 4$ pixels. The calculation procedure[18] is shown in **Fig. 2**.

In our method, gradient in x, y, t orientations of one

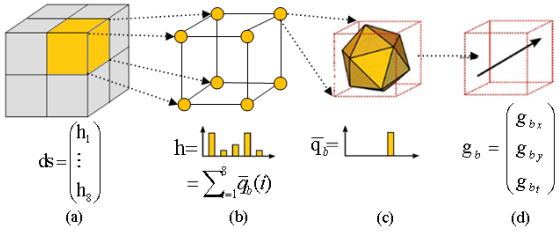


Fig. 2 Compact descriptor calculation. (a) the support region ($2\times2\times2$ cells) and descriptor of one interesting point; (b) histogram calculation of one cell ($2\times2\times2$ points); (c) orientation quantization; (d) three dimension gradient.

point b is adopted for descriptor calculation, recorded as $\mathbf{g_b} = (g_{bx}, g_{by}, g_{bt})$, as shown in **Fig. 2(d)**. In the step of orientation quantization, the three dimension gradient is projected to 20 surface normals of icosahedron. And the maximum element is kept in the quantized gradient vectors. It is recorded as $\bar{\mathbf{q}}_\mathbf{b}$, as shown in **Fig. 2(c)**.

For each cell, the quantized gradient vectors $\bar{\mathbf{q}}_\mathbf{b}$ of all elements were summed up to be one vector $\mathbf{h}$. Suppose $\bar{\mathbf{q}}_\mathbf{b}(i,j)$ refers to the jth bin in the quantized gradient vector of element i, and total s elements in the cell, the jth bin of $\mathbf{h}$ is calculated by formula (1). **Fig. 2(b)** shows the calculation of $\mathbf{h}$.

$$\mathbf{h}(j) = \sum_{i=1}^{s} \bar{\mathbf{q}}_\mathbf{b}(i,j). \tag{1}$$

After that, all the vectors in a support region of one feature point are finally concatenated to one feature vector, as shown in **Fig. 2(a)**. These local features represent human actions, and they are utilized for action recognition.

$$\mathbf{ds} = (\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_8) \tag{2}$$

3. Learning and Recognition by SOM

SOM[19] is an artificial neural network method proposed by professor T. Kohonen. It is trained using unsupervised learning to produce a low-dimensional, discretized representation of the training samples. In our method, batch learning is employed to train a SOM map, and k-NN is adopted for action classification.

3.1 SOM Training

Suppose that $\{\mathbf{m}_i\}$ are model vectors of neurons in a SOM map, where i refers to neuron number; $\{\mathbf{ds}_{lk}\}(l = 1,\cdots,L, k = 1,\cdots,C_l)$ are training local features, where L refers to action classes and C_l refers to the feature numbers of action class l.

At the first step, the SOM map with neuron vectors $\{\mathbf{m}_i\}$ is trained based on batch learning using all local features $\mathbf{ds}_{lk}$. Suppose that $\mathbf{R}_i$ is the neighborhood set of neurons which lie up to a certain radius from neuron i in the map, the procedure can be described as:

(1) Initialize the values of the $\{\mathbf{m}_i\}$ in some proper way, for instance, random initialization.

(2) Compare the $\mathbf{ds}_{lk}$, one by one, with $\mathbf{m}_i$ and list the features which are closest to $\mathbf{ds}_{lk}$ according to some distance, generally Euclidean distance.

(3) Let U_i denote the union of features which matched with model $\mathbf{m}_i$ and those matched with neurons in $\mathbf{R}_i$. Compute the means of the vectors $\mathbf{ds}_{lk}$ in U_i, and replace the old values of $\mathbf{m}_i$ by the respective means calculated by formula (3).

(4) Repeat these steps from the second step a few times until the trained map becomes steady.

$$\mathbf{m}_i^* = \frac{\sum_{\mathbf{ds}_{lk}\in U_i} \mathbf{ds}_{lk}}{n(U_i)}. \tag{3}$$

Where, $n(U_i)$ refers to the number of features $\mathbf{ds}_{lk}$ belonging to U_i.

After training, all the local features are mapped to neurons in the trained map, and action labels are assigned to neurons by the statistic of local features mapped to them. The features $\{\mathbf{ds}_{lk}\}$ are compared with neurons $\{\mathbf{m}_i\}$ to find the best matched neurons, and features are listed under the matched neuron $\mathbf{m}_i$. Following that, we count the feature number of each action category among the features matched with the neuron $\mathbf{m}_i$, and the action category which corresponds the maximum of statistics is given to $\mathbf{m}_i$ as its action label. The trained map is recorded as $\{\mathbf{m}_i, l_i, i = 1,\cdots,M; l_i \in \{1,\cdots,L\}\}$. Here, M is the neuron number in the trained map, and l_i is the action label of neuron $\mathbf{m}_i$, which is one element of action category set $\{1,\cdots,L\}$.

Theoretically speaking, there may exist neurons which are not matched by local features, so they are not given action label. However, since the quantity of training features $\{\mathbf{ds}_{lk}\}$ is much larger than the number of neurons $\{\mathbf{m}_i\}$ in our experiments, all neurons are given labels at last.

3.2 Recognition

Suppose that $\{\mathbf{ds}_n, n = 1,..,N\}$ are local features extracted from one testing video, and $\{\mathbf{m}_i, l_i, i = 1,\cdots,M; l_i \in \{1,\cdots,L\}\}$ are neurons in a trained map, k-NN is employed to classify testing local features into L action classes, and the the statistics of classification results determine the final action class of the testing video. Here, k is determined experimentally.

In the first step of recognition, we calculated the distances $\{D_{ni}\}$ (see Eq.(4)) between the testing local fea-

tures $\{\mathbf{ds}_n\}$ and the vectors of neurons $\{\mathbf{m}_i\}$. Then k-NN is employed to classify each local feature to some action class according to these distances. We count the numbers of local features classified to action class l, recording as $\{P_f(l), l, l = 1, \cdots, L\}$.

$$D_{ni} = \sqrt{(\mathbf{ds}_n - \mathbf{m}_i)^2}. \quad (4)$$

Considering the quantity of local features belonging to each action class differs much in the training processing, we normalize the statistics of recognition results $\{P_f(l)\}$ to obtain exact recognition result. Generally, the number of local features extracted from one video is different with others in all training videos. In our experiment, the total feature number of different action class varies a lot. Using these features for training, the numbers of neurons belonging to differ action classes also vary a lot in the trained map. If the neurons of one action category is much less, there will be less features being classified to this category. Thus videos of the category are easy to be recognized wrong. To obtain exact recognition, we analyze the feature quantity of each action class in the trained maps, recording the statistic of neuron numbers of action class l as $P_m(l)$. Here $\{P_m\}$ refers probability of action labels in the trained map, and l refers action category. Then we normalize the $\{P_f(l)\}$ by the formula (5). Following that, we find the $max(\hat{P}_f(l))$ corresponding action class l_0, and action class of the testing video is determined to be l_0.

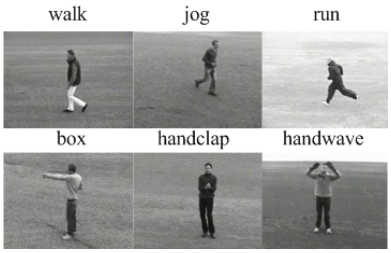
$$\hat{P}_f(l) = P_f(l)/P_m(l), l = 1, \cdots, L. \quad (5)$$

4. Dataset

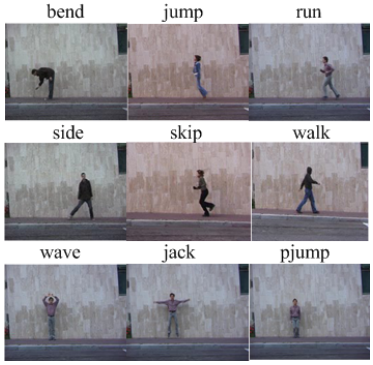
In this paper, three human action datasets are chosen to test our method. Among the three datasets, KTH and Weizmann dataset contain single human action in a relative clear background, and UCF dataset contains actions in various surrounding, interacting with objects. Some sample frames of these datasets are shown in **Fig. 3**.

4.1 KTH

The KTH dataset[4] contains six classes of human actions: walk, jog, run, box, hand wave, and hand clap. Each action is performed 4 or 6 times by 25 persons. The video sequences in the dataset are recorded in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, and indoors. Totally, the dataset consists of 600 video samples. Generally, more than 300 frames are contained per video. The background is homogeneous and static in most sequences. Comparing with other datasets, KTH contains more video samples. So in our experiment, the same



(a) KTH dataset



(b) Weizmann dataset

(c) UCF dataset

Fig. 3 Sample frames of three datasets.

with other researchers[7)11)12)], we separate video samples into training/validation set (8+8 persons) and test set (9 persons). Some sample figures of the dataset are shown in **Fig. 3(a)**.

4.2 Weizmann

Nine different action classes in the Weizmann dataset [15)] are used in our experiments: run, walk, skip, jump-jack(jack), jump, jump in place(pjump), gallop sideways(side), bend and handwave. In the dataset, each action category is performed once by nine persons. As a result, total 81 video samples are used in our experiments, and each video of them is composed of near 100 frames. As the previous researches did[7)11)12)], training and testing are also performed by leave-one-out on a per person basis, i.e., for each action category, training is done on the video samples performed by eight persons, and testing is done on the video sequences of

the remaining person[*1]. The average recognition rate of all tested video samples is calculated for comparison. Some sample frames of the Weizmann dataset are shown in **Fig. 3(b)**.

4.3 UCF

The UCF sport dataset[16] contains ten different types of human actions, swinging (on the pommel horse and on the floor), diving, kicking (a ball), weight-lifting, horse-riding, running, skateboarding, swinging (at the high bar), golf and walking. Each action is performed by several persons. The dataset consists of total 150 video samples which shows a large intraclass variability.

In our experiments, seven different action classes are chosen for action recognition, dive, ride horse, run, skateboarding, swing bench(on the pommel horse and on the floor), swing side (at the high bar) and walk front. For one action class, ten video samples are chosen for our experiment. We also adopt leave-one-out[7)11)12)] method to train and test these video sequences, and the average recognition ratio of all testing videos is calculated. Some frame samples are shown in **Fig. 3(c)**.

5. Experiments

In this section, we use the compact descriptor CHOG3D to describe human actions, and local features are calculated for all action classes. Then we train a SOM map using the calculated local features as it is introduced in section 3. The experiments are operated on KTH, Weizmann and UCF sport datasets.

5.1 Parameter Discussion

In KTH dataset, we test the optimal map size and the value k of k-NN for correct action recognition. We adopt the video samples in outdoor scenarios for the test, total 150 videos. We also separate the dataset samples into training/validation set (8+8 persons) and test set (9 persons) during the testing, thus in fact the final result of one parameter setting is obtained by testing 432 videos. Since the quantity of local features in each video is different, and one action is repeated several times in each video, we randomly choose 4000 local features in each video for this test experiment. Our experiments also indicate that 4000 local features are enough for action recognition.

To search for the optimal map size, we set the map size from 30×30 to 120×120 for SOM training, increasing 10 or 20 by one step. According to the method

[*1] Comparing with KTH, Weizmann and UCF contain less number of video samples. Therefore, here we adopt testing by leave-one-out as the previous works did.

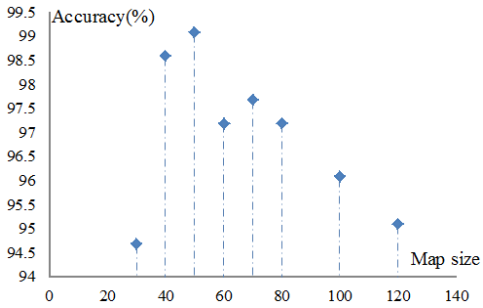


Fig. 4 Results of different map size in KTH.

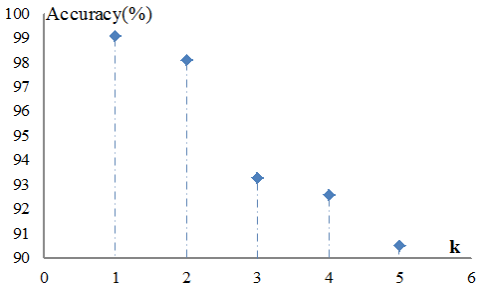


Fig. 5 Results of different k values in KTH.

introduced in section 3, we perform action recognition on the chosen video samples, and calculate the average recognition accuracy. Here the k is set to 1. **Figure 4** shows the recognition results of different map size. As the figure shows, the accuracy is highest when the training map size is 50×50. With map size increasing after 50, the recognition accuracy does not increase but decreases. It indicates the most proper map size for KTH dataset is 50×50. For Weizmann dataset, we also perform the testing in the same way, and the proper map size is determined to be 40×40. And the proper map size for UCF dataset is 40×40. These optimal parameters are utilized for map training in the recognition processing.

As the map size of 50×50 performs best in KTH dataset, we also set map size as it when we test the value k of k-NN. In this experiments, k is tested from 1 to 5 for feature classification, and the average recognition results are calculated in **Fig. 5**. The same as map size testing, each result is also obtained by testing 432 videos. As the figure shows, $k = 1$ is the most proper setting for k-NN classifier in our experiments.

5.2 Recognition Results

With the trained parameters, we perform recognition on the three datasets.

For the KTH dataset, the map size is set to 50×50, total 2,500 neurons. In the Weizmann dataset, the optimal map size is set to 40×40 in our experiments. And the size of SOM map is also set to 40×40 in the UCF

	box	wave	clap	run	walk	jog
box	0.806	0.157	0.037			
wave	0.005	0.958		0.037		
clap			1			
run				0.944		0.056
walk				0.005	0.972	0.023
jog				0.005	0.023	0.972

Fig. 6 Confusion matrix of KTH.

	bend	jack	jump	pjump	run	side	skip	walk	wave2
bend	1								
jack		1							
jump			0.45		0.11	0.11	0.33		
pjump				1					
run					0.78	0.11	0.11		
side			0.22			0.78			
skip			0.22			0.11	0.56	0.11	
walk						0.11		0.89	
wave2									1

Fig. 7 Confusion matrix of Weizmann.

	dive	ride	run	skate	swing bench	swing side	walk
dive	1						
ride		0.9	0.1	0.1			
run	0.2	0.2	0.5		0.1		
skate		0.1		0.8			0.1
swing bench			0.1		0.9		
swing side						1	
walk		0.1		0.1			0.8

Fig. 8 Confusion matrix of UCF.

sport dataset. The training and testing procedure is performed as it is introduced in section 3. The confusion matrix of recognition results in KTH, Weizmann and UCF datasets are shown in **Fig. 6**, **Fig. 7** and **Fig. 8**, respectively.

As the **Fig. 6** and **Fig. 7** show, most of the actions are recognized with a high recognition accuracy, while confusion recognition occurs among similar actions. In **Fig. 6**, the recognition accuracy of action "handclap" is 100%, and the action "walk" and action "jog" are 97.2%. Generally, action confusion occurs because these actions have similar gestures. For instance, 2.3% of the action "jog" is recognized as "walk," and 0.5% is recognized as "run" in **Fig. 6**. Similarly, 15.7% of action "box" is recognized as action "handwave," and 3.7% is recognized as action "handclap." The similar conclusion can be also obtained from the recognition result of Weizmann in **Fig. 7**. The 33% samples of "jump" are wrongly recognized as action "skip," and 22% "skip" are wrongly recognized as action "jump." The reason is that the two action classes are very similar. Except these actions, other actions can be distinguished correctly. As shown in **Fig. 7**, the recognition accuracies of actions "bend," "jack," "pjump" and "wave2" reach 100%.

However, the above conclusion is not adequate for UCF dataset. Some actions can not be distinguished correctly even they vary much, such as "run" with "dive" and "ride." This is because that these actions refer to interaction with some objects, which obstructs the correct recognition of actions. Nevertheless, most of actions are recognized with a high recognition accuracy by our method. These results certify that the proposed method is efficient in recognizing human actions.

Furthermore, we compare the recognition result of our proposed system with other related algorithms. The comparison of the average recognition accuracies of the datasets KTH, Weizmann and UCF is shown in **Table 1**. Here, we compare the best performance of frequently used local features with our algorithm. According to the evaluation in paper 11), the descriptor HoF, HoG/HoF and HOG3D performed better than other algorithms, and the average recognition accuracy of descriptor HoF and HoG/HoF were 92.1%, 91.8% respectively on KTH dataset. However, in paper 8), the average recognition accuracy of HoF and HOG3D were recorded as 86.7% and 91.4% on KTH dataset. In **Table 1**, we compare our result with the best performance of these algorithms. As the table shows, our system performs even better than these systems in KTH dataset. For Weizmann dataset, we compare the results of system adopting descriptor 3DSIFT[20] and the system utilizing descriptor HOG3D[12] with our proposed system, our result is similar with theirs. For UCF sport dataset, our result is 9.4% higher than the result in paper 11), and is 4.1% higher than HoF[11], 1.1% higher than HOG3D[12]. The comparison in the table shows that the result of our algorithm reaches the same level with others, even a little better.

In the proposed system, two main components contribute to the recognition results, descriptor CHOG3D for local feature calculation and SOM for action classification. Though CHOG3D is a compact descriptor, it has been certified in paper 1) that it is efficient to describe human actions. Furthermore, as a neuron network algorithm, SOM is able to solve complex, large quantity of features containing problems, which is suitable to map large quantity of local features to lower dimension for human activity recognition. The above

results also certify that SOM performs well on local feature used human action recognition. Therefore, our proposed method is able to solve the problem of action recognition.

5.3 Computation Cost Comparison

We compare the computation cost of our proposed system, CHOG3D + SOM, with the system HOG3D + SVM in paper 8). Since researchers are always interested in the computation cost of testing processing other than training, here we mainly compare the computation cost of testing processing in the two systems.

To compare the computation cost, we set the same parameters for the two system. Following is the definition of necessary parameters. First is the video number n_v, referring the videos to be tested. And local features number n_f, refers the features extracted from each video. The key words number is m, which refers the vocabulary size obtained by k-means in the system HOG3D + SVM and the neuron number in the trained SOM in system CHOG3D + SOM. Another parameter is n_s, which is the support vector number in SVM. The testing processing in our proposed method contains only k-NN classifier, the computation cost should be $O(n_f n_v m)$. However, in system HOG3D + SVM, the testing processing contains two steps. First, we need to integrate local features in a video to one histogram. Then we use SVM for classification. The computation cost of the two step processing is $O(n_f n_v m)+O({n_v}^2 n_s)$. It shows that the proposed system has lower computation cost.

We not only estimate the computation cost, but also test the time cost of the two systems on KTH dataset. With the same parameter setting, $n_v = 54, n_f = 2000, m = 2500$, we perform testing experiments on KTH dataset. The average time cost per video of our system is 0.27 seconds, while 1.5 seconds is cost to test one video by the method of system HOG3D + SVM. It can be seen that our system is much fast, and it certifies the above comparison result of computation cost.

Therefore, our proposed method obtain a good recognition result, and much faster.

6. Conclusion

In this paper, we proposed a SOM based system for local feature used human action recognition. In the proposed system, the compact descriptor CHOG3D was adopted to represent human actions. Then the SOM was employed to train local features to reduce feature dimension and extract key features of actions. The key features were assigned action labels, and k-NN was adopted to classify features in a testing action sequence to different action classes. According to the statistic of classification result, the action class of the the testing sequence is determined. Our method was tested on three datasets, KTH, Weizmann and UCF sports, respectively. Recognition results certified that the compact descriptor was efficient for action representation, and the proposed recognition system was able to solve the problem of single action recognition. Our method provides another choice for local feature used action recognition. In addition, we found that the results are not satisfactory when the actions interacting with objects. In future research, the objects will be considered to help improve recognition accuracy .

Table 1 Comparison of average recognition accuracy with other algorithms

Algorithm	KTH (%)	Weizmann(%)	UCF(%)
ESURF[9]	84.26	-	77.3[11]
3DSIFT[20]	-	82.6	-
HoF[11]	92.1	-	82.6
HoG/HoF[7]	91.8	-	81.6
HOG3D[12]	92.4	84.3	85.6
Ours	94.2	82.7	86.7

References

1) Y. Ji, A. Shimada, H. Nagahara, and R. Taniguchi, "A compact 3d descriptor in roi for human action recognition," in *IEEE TENCON 2010*, 11 2010.
2) I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. of IEEE International Conference on Computer Vision(ICCV'03)*, vol. 1, Nov. 2003, pp. 432–439.
3) C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. of Alvey Vision Conference*, 1988.
4) C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. of International Conference on Pattern Recognition(ICPR'04)*, vol. 3, Aug. 2004, pp. 32–36.
5) P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. of Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance(VS-PETS'05)*, Oct. 2005, pp. 65–72.
6) P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. of ACM International Conference on Multimedia*, vol. 20, Oct. 2007, pp. 357–360.
7) I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR'08)*, Jun. 2008.
8) A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d gradients," in *Proc. of British Machine Vision Conference(BMVC'08)*, Sep.

2008.

9) G. Willems, T. Tuytelaars, and L. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. of European Conference on Computer Vision(ECCV'08)*, vol. 5303, Oct. 2008, pp. 650–663.
10) H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded-up robust features," in *Proc. in European Conference on Computer Vision(ECCV'06)*, 2006, pp. 404–417.
11) H. Wang, M. U. Muhammad, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. of British machine vision conference*, 2009, pp. 127–137.
12) A. Klaser, "Learning human actions in video," Ph.D. dissertation, University of Grenoble, July 2010.
13) A. Shimada and R. Taniguchi, "Gesture recognition using sparse code of hierarchical SOM," in *Proc. of International Conference on Pattern Recognition(ICPR'09)*, Dec. 2008.
14) W. Huang and Q. Wu, "Human action recognition based on self organizing map," in *ICASSP*, March 2010, pp. 2130–2133.
15) M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. in IEEE International Conference on Computer Vision (ICCV'05)*, 2005, pp. 1395–1402.
16) M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach: a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. in CVPR(08)*, 2008, pp. 1–8.
17) E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, pp. 105–119, Jan. 2010.
18) Y. Ji, A. Shimada, and R. Taniguchi, "Human action recognition by som considering the probability of spatio-temporal features," in *Proc. of the 17th international conference on Neural information processing*, Nov. 2010, pp. 391–398.
19) T. Kohonen, *Self-Organizing Maps*. Berlin, Germany: Springer-Verlag, 1995.
20) P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. in ACM International Conference on Multimedia*, 2007, pp. 357–360.