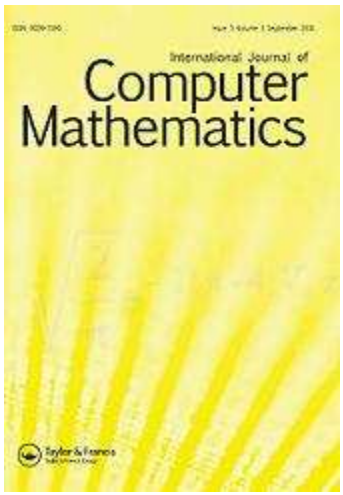
Human Action Representation Using Pyramid Correlogram of Oriented Gradients on Motion History Images

Ling Shao, Xiantong Zhen, Yan Liu, Ling Ji



Human Action Representation Using Pyramid Correlogram of Oriented Gradients on Motion History Images

Journal:	*International Journal of Computer Mathematics*
Manuscript ID:	GCOM-2010-0983-A.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	02-Mar-2011
Complete List of Authors:	Shao, Ling; The University of Sheffield, Electronic and Electrical Engineering Zhen, Xiantong; The University of Sheffield Liu, Yan; Hong Kong Polytechnic University Ji, Ling; Philips
Keywords:	Action recognition, Feature representation, Correlogram, HOG, Motion history image
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
IJCM-LS.rar	

International Journal of Computer Mathematics
Vol. 00, No. 00, 2010, 1–13

RESEARCH ARTICLE

Human Action Representation Using Pyramid Correlogram of Oriented Gradients on Motion History Images

Ling Shao[a,*], Xiantong Zhen[a], Yan Liu[b], Ling Ji[c]

[a]*Department of Electronic & Electrical Engineering,The University of Sheffield Sheffield,UK;* [b]Department of Computing, Hong Kong Polytechnic University, Hong Kong; [c]*Philips Healthcare, Philips Electronics, The Netherlands*

(*Received 00 Month 200x; in final form 00 Month 200x*)

The representation of human actions in video sequences is one of the key steps in action classification and recognition, performances of which are greatly dependent on the distinctiveness and robustness of the descriptors used for representation. In this paper, a novel descriptor, named Pyramid Correlogram of Oriented Gradients (PCOG), is presented for feature representation. PCOG, combined with the motion history images, captures both shape and spatial layout of the motion and therefore gives more effective and powerful representation for human actions and can be used for detection and recognition of a variety of actions. Experiments on challenging action datasets show that PCOG performs significantly better than Histogram of Oriented Gradients (HOG) both as a global descriptor and as a local descriptor.

Keywords: human action recognition;feature descriptor;pyramid correlogram of oriented gradients;motion history image;

AMS Subject Classification: F1.1; F4.3 **(... for example; authors are requested to provide some AMS Subject Classification codes and/or some CR Category numbers, and/or some MCS codes, and/or some Computing Classification System codes)**

1. INTRODUCTION

Automatic recognition and categorization of actions in video sequences is a very active research topic in computer vision and machine learning, and can be applied on many areas, including content-based video indexing, detecting activities and behaviors in surveillance videos, organizing digital video library according to specified actions, human-computer interfaces and robotics. The challenge is how to obtain robust action recognition and classification under variable illumination, background changes, camera motion and zooming, viewpoint changes, and partial occlusion. Moreover, the intra-class variation is often very large and ambiguity exists between actions such as running and jogging [1]. Feature representation as a fundamental part of action recognition will greatly influence the performance of the recognition system. Human actions from video data inherently contain both spatial and temporal information, which requires that descriptors of actions in video sequences accurately capture and robustly encode this kind of information.

*Corresponding author. Email: ling.shao@sheffield.ac.uk

ISSN: xxxx-xxxx print/ISSN xxxx-xxxx online

DOI: 0020716YYxxxxxxxx
http://www.informaworld.com

Many feature description methods have been proposed recently, however the current state-of-the-art in the representation methods are still far from tackling many of the above problems.

In this paper, we present a novel feature descriptor named Pyramid Correlogram of Oriented Gradients (PCOG). The shape descriptor is based on and extends the Histogram of Oriented Gradients (HOG) and is inspired mainly by two sources: (i) the image pyramid representation of Lazebnik et al. [2], and (ii) The Color Correlogram of Huang et al. [3]. The descriptor simultaneously models the spatial layout and temporal relations of motion features. The temporal information is encoded by integrating the motion histories (MHI) into an image. Correlogram of Oriented Gradients (COG) captures the shape and local relationship information, and through applying a hierarchical spatial pyramid in the representation, the spatial layout information is included in the descriptor. Therefore, the PCOG descriptor is more discriminative than HOG and provides good means for human action classification.

In the rest of the paper, we first briefly review the related work in human action classification and recognition in Section 2. The key methods used in our action recognition algorithm are described in Section 3. Then Section 4 details the experiments and results. Finally, we draw the conclusion in Section 5.

2. RELATED WORK

Different approaches in the field of human motion recognition have been proposed and developed for both global representation and local representation. A comprehensive overview can be found in a recent survey on vision-based human action recognition [4]. The global representation treats the vision observation as a whole. A person is detected as the region of interest and localized in the video using background subtraction or other localization methods. The global representation attempts to capture the whole human body characteristics including contours and poses, but does not consider the human body being composed of body parts. Silhouettes, edges, shapes or optical flow are often used for the global representation. Bobick and Davis [5] extract silhouettes from a single view and aggregate differences between subsequent frames of an action sequence. Then motion energy image (MEI) which indicates where motion occurs and motion history image (MHI) which records how the motion is moving are constructed. However, this method depends on the background subtraction. Weinland et al. [6] extended this method using motion history volumes by means of five independent views of the same actor. To achieve translation and scale invariance, Wang et al.[7] perform R-transform on extracted silhouettes to represent low-level features. The advantage of the R-transform lies in its low computational complexity and geometric invariance. Star skeleton as feature for action recognition is used by Chen et al. [8]. The feature is defined as a five-dimensional vector in star fashion because the head and four limbs are usually local extremes of the human shape. An action is composed of a series of star skeletons over time. Silhouettes and contours are combined as descriptors by Wang and Suter [9]. Average motion energy and mean motion shape are derived to characterize actions. Motion and trajectories are also commonly used features for human action recognition. Motion trajectories obtained from tracking the body parts are used by Forsyth et al. [10] to perform recognition. Motion information such as optical flow has also been used in global representations, when background subtraction cannot be performed reliably. The global representation is powerful due to the capture of the holistic and distinctive information on the performed action; however, it may greatly depend on the recording conditions such as position

of the pattern in the frame, spatial resolution and relative motion with respect to the camera. Moreover, global image representation can be influenced by motions of multiple objects and variation in the background and suffer from occlusions.

Local representation methods have been successfully used for many recognition tasks including object and scene recognition as well as human motion recognition [1, 11-14]. Local representation methods describe the observation as a collection of local descriptors or patches. Space-Time Interest Points (STIPs) are firstly extracted from the video using certain detection methods. Then a robust description of the area around each STIP is applied and a model based on the independent features (Bag of Words) or a model that can also contain structural information is employed to represent an action. Those methods do not require tracking and stabilization and are often more resistant to cluttering and occlusions as only few parts may be occluded. Usually they perform well even with a non-uniform background. The Gradient descriptor was introduced in the field of human action recognition by Dollar et al. [15] and is obtained through computing the brightness gradients of the cuboids along x, y and t dimensions. Histogram of Oriented Gradients (HOG) extended from its 2D version through encoding the spatial distribution of local gradients is computed by dividing an image into small spatial regions called cells. Histogram of gradient directions is accumulated over the pixels of each cell. The representation is formed by combining histograms of all the cells in the image. Histogram of Optic Flow (HOF) with the same idea as HOG is used by Laptev et al. [14] and they prove that the combination of HOG and HOF, named HOG-HOF, to perform better than each separate method. The final descriptor is simply a concatenation of HOG and HOF. 3D-SIFT as an extension of the 2D-SIFT has been developed by Scovanner er al. [1]. The extended SURF was proposed by Willens et al. [11]. Compared with 3D SIFT, 3D SURF is computationally much faster. Wang et al. [16] have evaluated the local descriptors and they demonstrated that descriptors combining image gradient and flow information have the best performance. Due to its advantages in representation of human actions, local representation draws a lot of attention. Despite recent developments, it is still an open and active aspect of research in human action recognition.

3. METHODS

3.1 *Motion Templates*

Motion templates was proposed by Bobik et al. [5] including motion energy images (MEI) and motion history images (MHI), and are used to represent the motions of an object in a video sequence. Assume $I(x, y, t)$ is an image sequence and let $D(x, y, t)$ be a binary image sequence indicating regions of motion, which can be obtained from image differencing. The binary MEI $E_\tau(x, y, t)$ (τ is the duration) is defined as:

$$E_\tau(x, y, t) = \bigcup_{i=1}^{\tau-1} D(x, y, t) \tag{1}$$

Motion history images (MHI) $H_\tau(x, y, t)$, are used to represent how the motion image is moving, and are obtained using a simple replacement and decay operator:

$$H_\tau(x, y, t) = \{_{max(0, H_\tau(x,y,t-1)-1)}^{\tau \quad if D(x,y,t)=1} \quad _{otherwise} \tag{2}$$

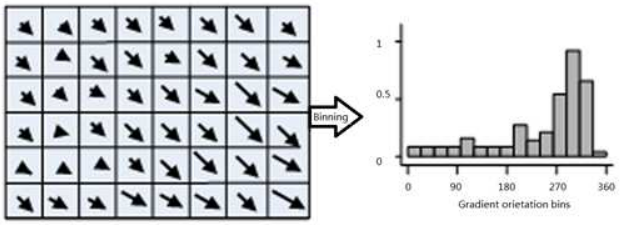


Figure 1. Histogram of motion gradients.

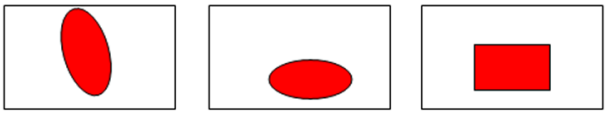


Figure 2. Three images with the same histogram have different appearances.

Once cuboids are extracted from the video sequences, MHI can be computed with the above equation through projecting all frames onto one image, namely the MHI image.

3.2 *Motion History Gradients*

Having the MHI image of a video sequence through projecting all the frames onto one, we can derive an indication of the action by measuring the gradients of the MHI image. Fig. 1 shows an example of the gradients of the MHI and the histogram of motion gradients. Note that these gradient vectors point orthogonally to the moving object boundaries at each "time step"in the MHI. Gradients of the MHI can be calculated efficiently by convolving with separable Sobel filters in the X and Y directions yielding the spatial derivatives: $F_x(x, y)$ and $F_y(x, y)$, The gradient orientation at each pixel is given by:

$$\phi(x, y) = \arctan \frac{F_x(x, y)}{F_y(x, y)} \tag{3}$$

3.3 *Pyramid Correlogram of Oriented Gradients*

The proposed algorithm describes human actions by motion templates, namely MHI images, which encode the temporal information by integrating the motion histories into an image and represent this image by its local shape as well as the spatial layout of the shape. Traditional color histograms capture only the color distribution in an image without any spatial correlation information; however, images with similar histograms can have very different appearances. Fig. 2 shows an example that three images have the same histogram but different appearances. Therefore, it is necessary and important to add spatial information into the histogram-based representation for the redefinition of color histogram based methods. Color correlogram was proposed by Huang as a new color feature and proven effective for the description of image content in the field of image retrieval [3].

A color correlogram expresses how the spatial correlation of pairs of colors changes with distance d. Assume we have an image I , which is quantized into

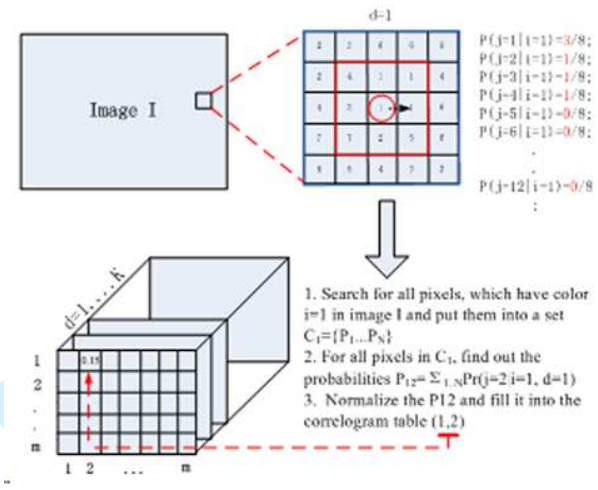


Figure 3. (a) Image I is quantized into m=12 color bins. Given a pixel P has the color $C_i = 1$, what is the probabilities that a pixel P', which has color C_j at distance d=1 away from P. in this case. (b) Element (1, 2) in PCOG d=1, shows the probability of finding a pixel of color $C_i = 1$ at distance d=1from a pixel of color $C_j = 2$ in the image.

m color bins, the color correlogram of I is defined for $C_i, C_j \in [1 \dots m]$,$d \in [1 \dots k]$ as:

$$\gamma^{(k)}_{C_i,C_j}(I) = Pr(P_j \in C_i | P_i \in C_i, |P_i - P_j| = k) \tag{4}$$

Given any pixel of color C_i in the image I, $\sigma^{(k)}_{C_i,C_j}(I)$ denotes the probability that a pixel at distance d=k away from the given pixel belonging to color C_j and Figure. 3 illustrates the formation of a correlogram descriptor at distance d=1 schematically. The size of the correlogram is noticeable as it is $d^2 \times k$ for an image with m color bins. When choosing d to define the correlogram, we need to address the following issues. A large d would result in expensive computation and large storage requirements, while a small d might compromise the quality of the feature. So a tradeoff between them should be considered.

We apply the idea of correlogram on the gradients to add spatial information to our descriptor, which captures the spatial co-occurrence of a pair of the gradient sets. Therefore it incorporates the shape information (gradients) as well as the spatial correlations (positions of gradients) and is robust to small geometric deformation. As local correlations between orientations are more significant than global correlations in an image, a small value of d is sufficient to capture the spatial correlation. The properties of Correlogram of Oriented Gradients (COG) are: (i) it includes the spatial correlation of orientations; (ii) it can be used to describe the global distribution of local spatial correlations of orientations if d is chosen to be local. The COG descriptor of an image only contains information about the local spatial structures and does not give any information about the overall structure of the shape. To preserve the rough structure of the global shape, the MHI image is divided into sub regions. The idea is illustrated in Fig. 4. The descriptor consists of correlograms of oriented gradients over each image sub-region at each resolution level. This results in a higher-dimensional representation that preserves more information. For each level l, $l \in [1 \dots L]$, we divide the frame along X and Y dimensions into $2^{2\times l}$ sub-regions. Each cell can be described as a correlogram of the motion features in it. We can concatenate these correlogram representations from

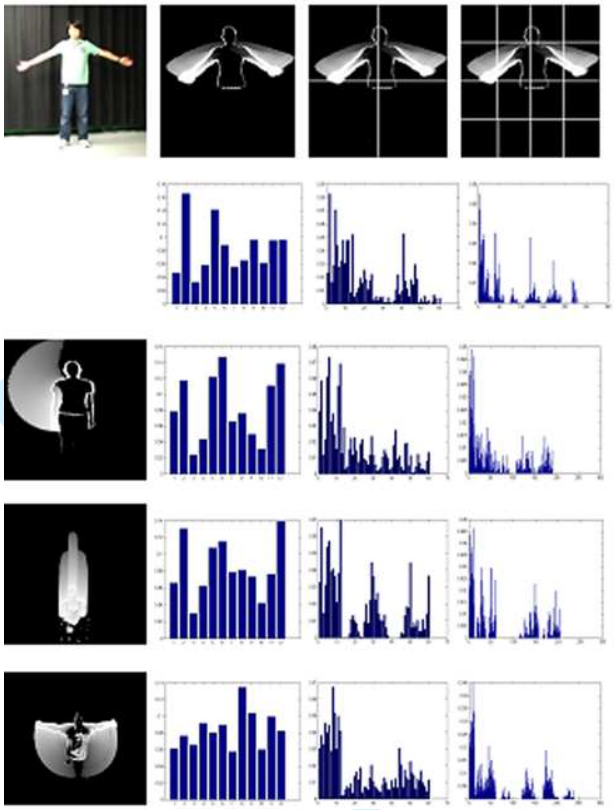


Figure 4. Pyramid representation of MHI, at layer L=0, l, 2.

all cells in all levels into a long vector as the representation for the MHI image. Hence, the proposed descriptor is named Pyramid Correlogram of Oriented Gradients (PCOG). Histogram of oriented gradients (HOG) was used as a descriptor of human actions by Laptev et al. [14]. Similarly, we apply the pyramid strategy on HOG and obtain pyramid histogram of oriented gradients (PHOG).

Different weights are needed to be assigned to each of various levels of the pyramid as different information is captured at each of them. At a finer resolution (l=0), the captured correspondence between two sets is more accurate. Therefore, the similarity information gained at a coarser level is penalized and give more weights to the similarity measured at a finer resolution. The weight we assign at level l is:

$$W(l) = \frac{1}{2^{l-1}}, l \in [1 \dots L] \tag{5}$$

3.4 *Feature Detection*

The proposed descriptor can be also used for the local representation, thus local features should be extracted from the video sequences first. The feature detection method used for the local representation is the periodic feature detector proposed by Dollar et al. [15]. The detector is based on a set of separable linear filters which treat the spatial and temporal dimensions in different ways. The response function is given by:

$$R = (I * g * h_{even})^2 + (I * g * h_{odd})^2 \tag{6}$$

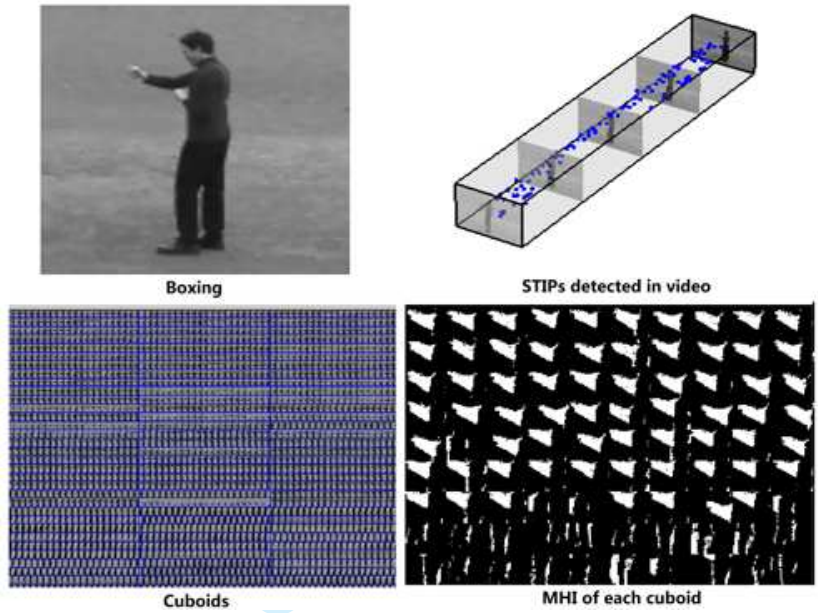


Figure 5. Spatial-temporal interest points and their corresponding MHIs.

where $*$ denotes the convolution operation,$g(x, y, \sigma)$ is a 2D Gaussian kernel, applied only along the spatial dimensions, and h_{even} and h_{odd} are a quadrature of 1D Gabor filter applied only temporally. They are defined as:

$$h_{even}(t; \tau, \omega) = -\cos(2\pi t\omega e^{-t^{2/\tau^2}}) \tag{7}$$

and

$$h_{odd}(t; \tau, \omega) = -\sin(2\pi t\omega e^{-t^{2/\tau^2}}) \tag{8}$$

The parameters σ and tau correspond roughly to the spatial and temporal scales of the detector, and they are set manually by user. The authors suggest keeping $\omega = 4/\tau$. The response function gives the strongest responses where there are periodic motions, however the detector also responds strongly to a range of other motions, e.g. local regions exhibiting complex motion patterns such as space-time corners. This method can detect a high number of space-time interest points, was proven to be faster, simpler, more precise and gives better performance, even though only one scale is used [17]. So it is adopted for STIP detection in this work. Fig. 5 gives an example of feature detection from a video sequence and the MHI images of the cuboids.

3.5 *Bag of Words*

In the local representation, the bag of words technique is used for modeling the human actions. Space-Time Interest Points (STIPs), which are the locations where the interesting motion is happening, are detected from the video sequences by a feature detection method. Small video patches, namely cuboids, are extracted from around each detected interest point and contain the local information. Each cuboid is described by a feature description method. The result is that the video sequences

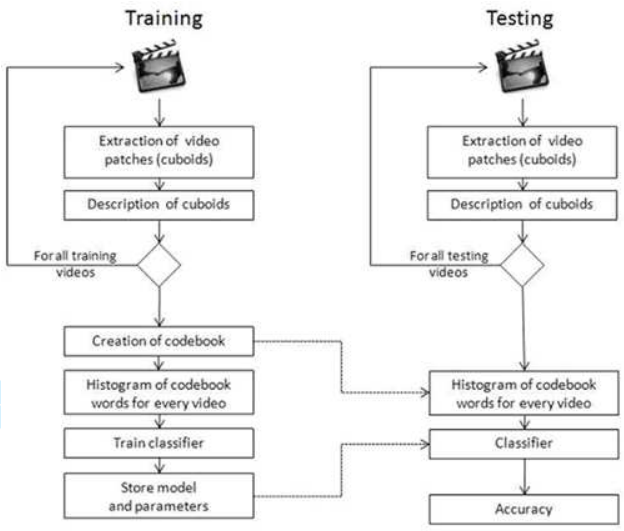


Figure 6. Methodology for Human Action Recognition based on the local representation.

are discarded and represented as series of cuboid descriptors. Linear discriminant analysis (LDA) is performed to reduce the dimension of the feature vectors. A visual vocabulary (codebook) is built by applying clustering on descriptors from all the training samples using the k-means algorithm. The center of each cluster is defined as the 'visual word', the length of which depends on the length of the feature vectors. A histogram of occurrence of the visual words in the entire video is then computed by assigning each feature descriptor to the closest (using Euclidean distance) vocabulary word. Thus each action video is represented as the spatial-temporal words from the codebook in the form of a histogram, which is eventually used for classification. The entire methodology is shown in Figure 6.

3.6 *Classification Method*

Support Vector Machine (SVM) is a popular technique for classification [18]. The goal of SVM is to produce a model which predicts the class labels based on the given feature values in the testing set. In the binary case according to theory, SVM finds a linear separating hyperplane with the maximal margin in the high dimensional space. The margin is the distance between two hyperplanes defined by the so-called support vectors. When the margin reaches its maximal value as hyperplanes adjust, the distance between the middle hyperplane and its nearest point is maximized. The nonlinear Radial Basis Function (RBF) kernel is chosen in our approach since it has the following advantages: (1) The RBF kernel can handle the case when the relation between class labels/action types and attributes/features is nonlinear; (2) The number of parameters, which affect the complexity of model selection, is less than that of linear kernel functions; (3) Moreover, the RBF kernel has less numerical difficulties.

4. EXPERIMENTS AND RESULTS

We firstly validate the choices of the three major parameters of the PCOG descriptor for the global representation of human actions: the number of the layers L, the number of bins K, and the searching radius d. This is implemented on our own dataset which includes video sequences of 8 indoor fitness exercises performed by

Figure 7. Samples of human actions from our dataset.

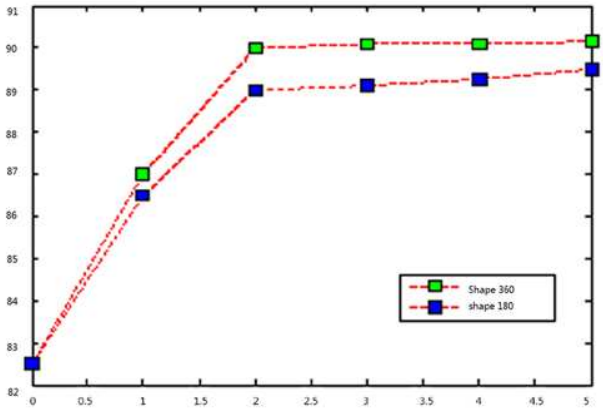


Figure 8. Layer L=0, . . . ,5

20 different subjects and some sequences selected from the KTH dataset [11-12, 14, 19-22] including handwaving, boxing, etc. Fig. 7 shows snapshots of our dataset.

Then the PCOG descriptor used for the local representation of human actions is evaluated on the KTH dataset which consists of six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping. Each action is performed by 25 different persons in different scenarios of outdoor and indoor environments. The dataset has in total 600 action sequences. We divide them into two parts: 16 persons for training and 9 persons for testing, as it has been done in [19]. In our tests, we use the SVM implementation in the public available machine learning library libSVM [23].

4.1 *Parameter optimization*

In the global representation experiment, two PCOG descriptors are compared: one with orientations in the range [0,180] (where the contrast sign of the gradient is ignored) and the other with the range [0, 360]. We refer to these as *Shape*180 and *Shape*360 respectively. Number of layers: We consider the number of layers firstly. A large number of layers will lead to a higher computational cost. From Fig. 8 we can see that the performance is optimal when L is 2. When L is 4, the recognition rate is slightly improved, while the computation time is tripled.

Number of bins K: We change the value of K in the range [8. . . 40] for *Shape*180 and [20. . . 80] for *Shape*360. Note that the range for *Shape*360 is doubled, so as to preserve the original orientation resolution. The performance is optimal with K = 20 orientation bins for *Shape*180, and K = 40 for *Shape*360 as shown in Fig. 9, from which we can see the performance is not very sensitive to the number of bins used. Searching radius d: As discussed in Subsection 3.3, the local correlations between orientations of gradients are more significant than global correlations of gradients in an image, a small value of d is sufficient to capture the spatial correlation. We test the different values d with the range [3, 10]. Fig. 10 shows that when d=6 the

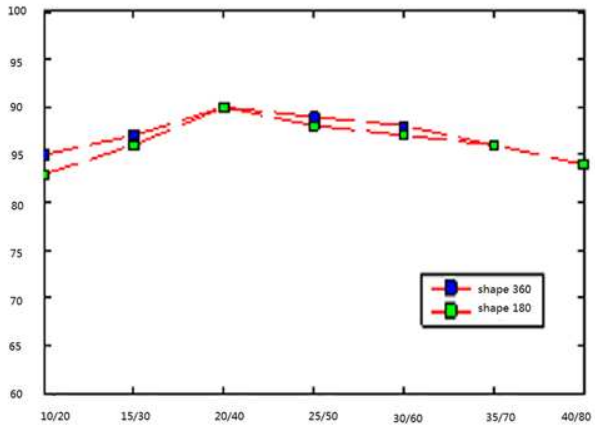


Figure 9. Orientation bins K=10,,40.

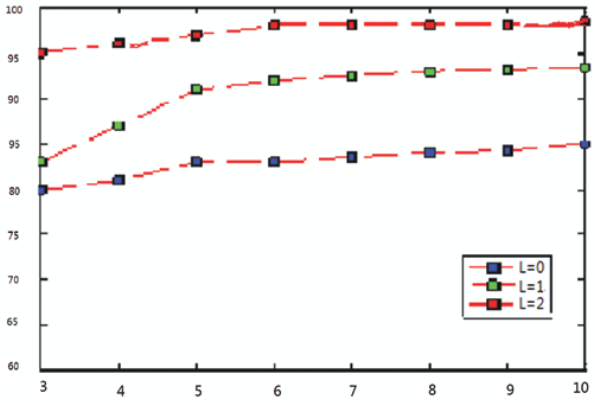


Figure 10. Performance test for searching radius d.

performance is optimal. A large d would result in expensive computation and large storage requirements.

4.2 *Human action recognition*

The descriptor used for global and local representations in human action recognition is evaluated and the accuracy of the recognition is defined as:

$$Accuracy = \frac{number of corrected recognition}{number of actions} \tag{9}$$

4.2.1 Global representation

To show the better discriminativity of the COG descriptor for global representation, we compare it with two other frequently used global feature descriptors: Hu's moment invariants and HOG at the finest level (whole image). The confusion matrices are shown in Fig. 11 and from Table 1 we can see that COG has the highest recognition rate of 83%. To show the improvement caused by adding the spatial layout information to the descriptor, we apply the pyramid kernel to HOG and COG descriptors. From the confusion matrices we can observe that both methods

Table 1. Recognition rate in different layers

	Hu's Moments	PHOG	PCOG
L=0	70.0%	75.0%	83.0%
L=1	–	90.0%	92.0%
L=2	–	97.0%	98.0%

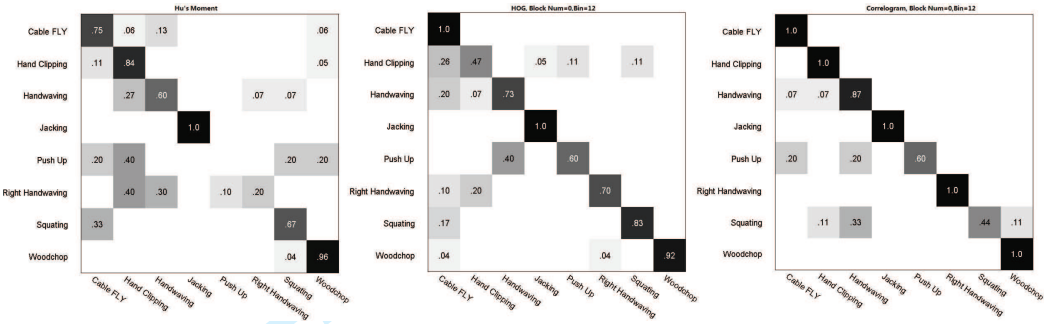


Figure 11. Confusion matrices of Hu's Moment, HOG and COG (whole image).

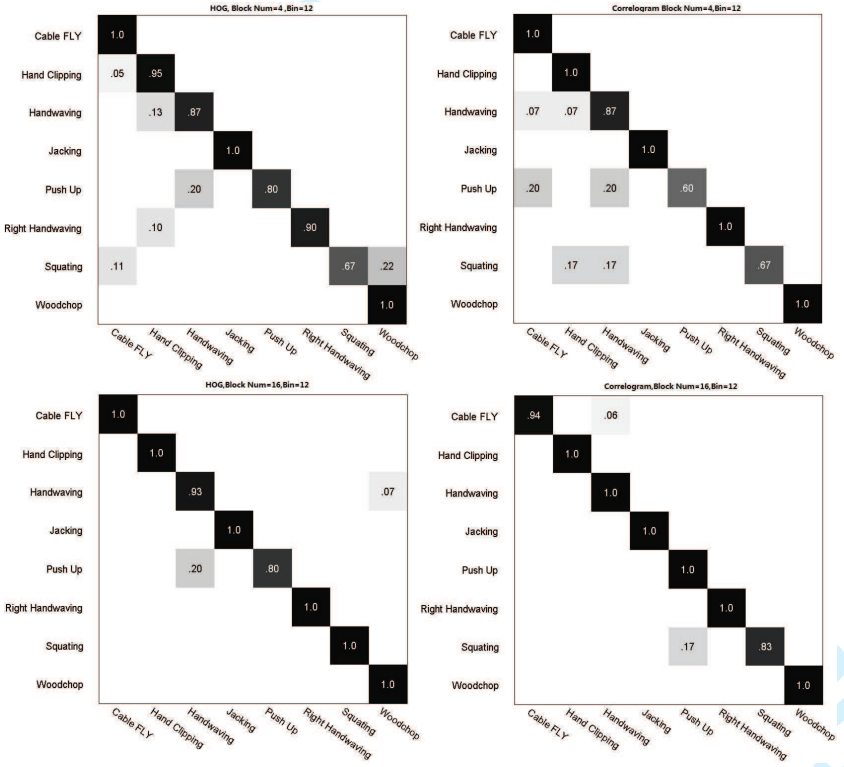


Figure 12. First row: PHOG and PCOG at L=1, second row: PHOG and PCOG at L=2.

achieve a higher recognition rate. At layer L=2, PCOG outperforms PHOG by 1%. The confusion matrices are given in Fig. 12.

4.2.2 Local representation

We also compare the performances of PHOG and PCOG for local representation. During feature detection, we set sigma=2.8 and tau=1.6, which give better results [17]. The PCOG and PHOG descriptors with orientations in the range [0-360], searching radius in the range [1-4] and 2 layers are used. Descriptors with there different orientation bins are compared in the experiments. The performances of

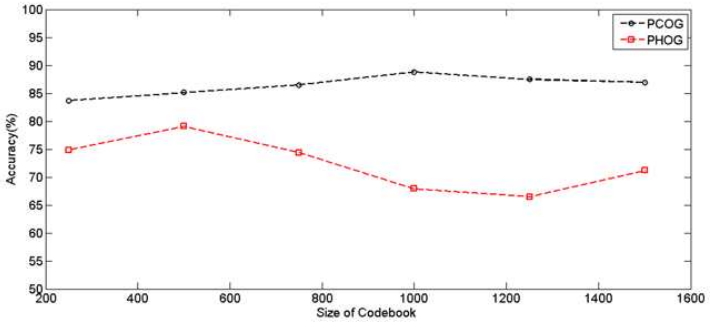


Figure 13. PHOG and PCOG descriptors with orientation bins K =12

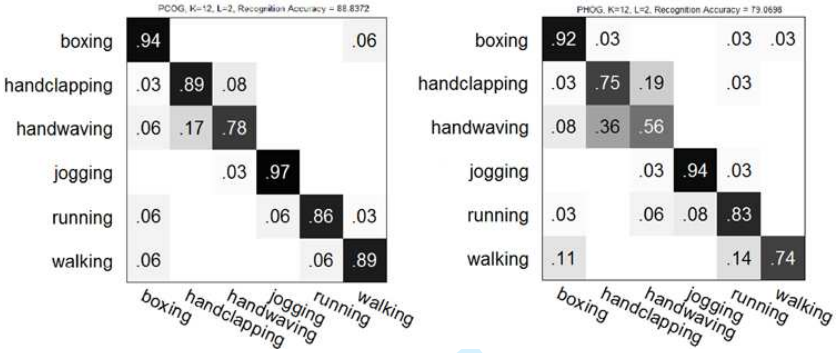


Figure 14. Confusion matrixes of PHOG (right) and PCOG (left)

PHOG and PCOG as local descriptors for action recognition are shown in Fig. 13 and the confusion matrixes are depicted in Fig. 14.

5. CONCLUSION

In this paper, we present a novel descriptor for the representation of human actions. The proposed PCOG descriptor as an effective extension of the HOG descriptor simultaneously integrates the spatial-temporal motion information in video sequences. The motion information of a video sequence is encoded by projecting multiple frames into a single image using the motion history image. Through adding the spatial layout information to the descriptor, the recognition performance is greatly improved. We have demonstrated that PCOG as a global descriptor and a local descriptor has a good performance in action recognition and outperforms the PHOG descriptor, which proves that correlating of orientated gradient is an effective means to capture the spatial layout information and can improve the discriminability and distinctiveness of the descriptor.

References

[1] P. Scovanner, S. Ali and M. Shah, A 3-dimensional sift descriptor and its application to action recognition. in Proceedings of the 15th international conference on Multimedia. 2007. New York,NY, USA.
[2] S. Lazebnik, C. Schmid and J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. in mputer Vision and Pattern Recognition, 2006. CVPR 2008. IEEE Conference on 2006. 2006.
[3] Jing Huang, Color-Spatial Image Indexing and Applications. 1998, Cornell University.

[4] Ronald Poppe, A survey on vision-based human action recognition. Image and Vision Computing, 2010. 28(6): p. 976-990.
[5] A. F. Bobick and J. W. Davis, The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001. 23(3): p. 257-267.
[6] D. Weinland, R. Ronfard and E. Boy, Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding, 2006. 104(2-3): p. 249-257.
[7] Ying Wang, Kaiqi Huang and Tieniu Tan, Human activity recognition based on R transform. in Proceedings of the Workshop on Visual Surveillance (VS) at the Conference on Computer Vision and Pattern Recognition. 2007. Minneapolis, MN, USA.
[8] Hsuan-Sheng Chen, Hua-Tsung Chen, Yi-Wen Chen and Suh-Yin Lee, Human action recognition using star skeleton. in Proceedings of the International Workshop on Video Surveillance and Sensor Networks (VSSN'06). 2006. Santa Barbara, CA.
[9] Liang Wang and David Suter, Informative shape representations for human action recognition. in Proceedings of the International Conference on Pattern Recognition (ICPR'06). 2006. Kowloon Tong, Hong Kong.
[10] D. Ramanan and D. A. Forsyth., Automatic annotation of everyday movements. NIPS, 2003.
[11] G. Willems, T. Tuytelaars and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector, European Conference on Computer Vision,ECCV. 2008.
[12] A. Kläser, M. Marszalek and C. Schmid, A spatio-temporal descriptor based on 3Dgradients. in Proceedings of the British Machine Vision Conference (BMVC '08). 2008.
[13] I. Laptev and T. Lindeberg, Local descriptors for spatio-temporal recognition. in First International Workshop on Spatial Coherence for Visual Motion Analysis, LNCS. 2004.
[14] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, Learning realistic human actions from movies. in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on 2008.
[15] P. Dollar, V. Rabaud, G. Cottrell and S. J. Belongie, Behavior Recognition via Sparse Spatio-Temporal Features. in Proc. of ICCV Int. work-shop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VSPETS). 2005.
[16] Heng Wang, Muhammad Muneeb Ullah, Alexander Kläser, Ivan Laptev and Cordelia Schmid, evaluation of local spatio-temporal features for action recognition. in Proceedings of the British Machine Vision Conference (BMVC '09). 2009.
[17] Ling Shao and Riccardo Mattivi, Feature detector and descriptor evaluation in human action recognition. in Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR). 2010. Xi'an, China.
[18] N. Cristianini and J. Taylor, eds. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. 2000, Cambridge UP.
[19] C. Schuldt, I. Laptev and B. Caputo, Recognizing human actions: A local svm approach. in Proceedings of the Pattern Recognition, 17th International Conference on,(ICPR'04). 2004. Washington, DC, USA.
[21] C. Harris and M.J. Stephens. A biologically inspired system for action recognition, Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on 2007.
[22] S.F. Wong and R. Cipolla. Extracting spatio-temporal interest points using global information, The IEEE 11th International Conference on Computer Vision. 2007. Rio de Janiero, Brazil.
[23] C.C. Chang and C.J. Lin, LIBSVM: a library for support vector machines. 2001; Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.].