Pattern Recognition Letters 33 (2012) 438–445

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

# Human action segmentation and recognition via motion and shape analysis

Ling Shao [a,*], Ling Ji [b], Yan Liu [c], Jianguo Zhang [d]

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

A R T I C L E I N F O

*Article history:*
Available online 30 May 2011

*Keywords:*
Human action segmentation
Motion analysis
PCOG
Motion history image
Human action recognition

A B S T R A C T

In this paper, we present an automated video analysis system which addresses segmentation and detection of human actions in an indoor environment, such as a gym. The system aims at segmenting different movements from the input video and recognizing the action types simultaneously. Two action segmentation techniques, namely color intensity based and motion based, are proposed. Both methods can efficiently segment periodic human movements into temporal cycles. We also apply a novel approach for human action recognition by describing human actions using motion and shape features. The descriptor contains both the local shape and its spatial layout information, therefore is more effective for action modeling and is suitable for detecting and recognizing a variety of actions. Experimental results show that the proposed action segmentation and detection algorithms are highly effective.

## 1. Introduction

Recognizing human actions in video has many important computer vision applications, such as video surveillance, human computer interaction, video browsing, and analysis of sport events. The motivation of this work is: develop a low cost automated human action detection framework as the replacement of the expensive sensor network systems. Sensor network systems are widely used in sports and exercise apparatus to measure and track athletes' performance. Although the sensor network systems can provide precise and flexible measurements, the main drawback is their high cost. Our detection framework only employs a consumer camera as the input. Hence the cost is highly reduced.

In this paper, we propose a human action detection framework, which can detect different exercise types and count the exercise cycles in an indoor environment. The novelty of our approach is twofold. Firstly, unlike the existing action recognition framework, which only recognizes the single action from one video sequence, our approach can detect different action classes from a video sequence. To do this we have to partition the video sequence into cycles automatically and recognize the action types respectively. We propose a color based method and a motion based method for human action temporal segmentation under a stationary background condition. Secondly, we apply a shape-based feature descriptor: Pyramid Correlogram of Oriented Gradients (PCOG). The shape description calculated from the Motion Energy Images (MEI) and Motion History Images (MHI) (Bobick and Davis, 2001) gives a good presentation of motion and shape information respectively. By using the local and spatial layout properties, the PCOG descriptor captures the essential information of human actions and provides good discriminativity for classification.

## 2. Related work

The existing methods for vision-based human action recognition can be classified into three main categories: model based approaches, spatio-temporal template based approaches and "bag-of-words" based approaches.

### 2.1. Model based

Model based approaches depend on locating and tracking body limbs, which require a 3D or 2D view-based model of the body (Aggarwal and Cai, 1999; Gavrila, 1999). Gu et al. (2008) presented a model-based human action recognition approach, where the action is represented as a sequence of joints in a 4D spatial–temporal space, and modeled by two HMMs: a conventional HMM for the global movement feature and an exemplar-based HMM for the configuration feature. In (Liang et al., 2009), human behaviors are segmented into atomic actions, each of which indicates a basic and complete movement. The HMM model is made for learning and recognizing atomic human actions. Zhou et al. (2009) presented a tensor analysis based approach for tracking the object trajectory, which can be also applied in tracking body parts from action recognition.

[*] Corresponding author.
*E-mail address:* ling.shao@sheffield.ac.uk (L. Shao).

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.05.015

### 2.2. Spatio-temporal template based

In (Bobick and Davis, 2001), an input image sequence is used to construct MEI and MHI for determining where and when respectively motion occurs in the sequence. For recognition, a set of moment invariants is calculated for each MHI/MEI and a Mahalanobis distance (Mahalanobis, 1936) metric is applied between the sets in order to discriminate different activities. 3D extension of the temporal templates was proposed in (Weinland et al., 2006). They used multiple cameras to build motion history volumes and performed action classification using Fourier analysis in cylindrical coordinates. A spatio-temporal descriptor based on global optical flow measurements was also proposed in ( Efros et al., 2003).

### 2.3. Bag-of-words

In contrast, "bag-of-words" based approaches detect local salient descriptors as visual words, which are then used to recognize the actions. "Bag-of-words" has been used successfully for object categorization (Willamowski et al., 2004; Nister and Stewenius, 2161). Inspired by text categorization, it represents an object as a histogram of occurrence of local features. Recently, the "bag-of-words" representation has been applied in activity recognition (Schuldt et al., 2004; Doll ar et al., 2005). In (Thurau, 2007), Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2006) is used for human behavior recognition. The behavior can be classified by means of histogram comparison, interpreting behavior recognition as a problem of statistical sequence analysis. Laptev (2005) extended the Harris corner detector to 3D for detecting spatio-temporal features in video and applied on action recognition in (Laptev et al., 2008. Niebles et al. (2006) presented an action localization and recognition algorithm based on detected spatio-temporal features and a probabilistic Latent Semantic Analysis (pLSA) model. However, these approaches lack the relations between the features in the spatial and the temporal domains which are helpful for recognition.

There are many recent papers on extending "bag-of-words" to add the spatial relations in the context of object categorization. In particular, the pyramid match kernel (Grauman and Darrell, 2005; Lazebnik et al., 2006) utilizes the weighted multi-resolution histogram intersection as a kernel function for classification with sets of image features.

The model based methods require accurate body parts/limbs detection, which is often a difficult task, therefore are not feasible in many applications. The bag-of-words model is based on local features and is robust to background clutter, partial occlusion and viewpoint changes. However, it is essentially a sparse representation method and is less informative compared to the global representation. The spatial temporal template based approach is in between of the above two and is the one we adopt in this paper.

All the existing techniques of action recognition only identify whether a certain action is performed in a video sequence. They are not able to temporally segment a periodic action into cycles. In this paper, we attempt to do temporal segmentation of periodic actions and then to recognize them individually. The localization and recognition of actions will benefit the analysis of long videos where actions are only performed in particular parts of the video. For action representation, our feature descriptor (PCOG) is mainly inspired by two sources: (i) the image pyramid representation of Lazebnik et al. (2006), and (ii) The Color Correlogram of Huang (1998). The proposed descriptor tries to simultaneously model the spatial layout and temporal relations of the local motion features. The temporal information is encoded by integrating the motion histories (MHI) into an image. Correlogram of Oriented Gradients (COG) captures the local shape information and the spatial layout information is captured by using a hierarchical spatial pyramid in the representation. Initial results of the descriptor were presented in (Shao et al., 2010).

## 3. Methodology

In this section, the algorithmic details of the action segmentation and detection system are given. As for human body detection and MHI/MEI, we directly adopt the techniques proposed by Dalal and Triggs (2006) and Bobick and Davis (2001).

### 3.1. Human action detection framework

In this sub-section, we outline the framework for our human action detection system. A schematic diagram is drawn in Fig. 1. The detection process is illustrated in five steps and the correctly classified sequences against all sequences give the finial detection accuracy. The five steps are summarized as follows:

(1) Locate the human/exerciser from the input sequence (camera or AVI file) by matching extracted HOG descriptors with the prototypical action primitives (Dalal and Triggs, 2006). The output only contains the region of interest (ROI).
(2) Generate the MHI/MEI for each frame and calculate the motion gradients for the motion estimation. The obtained gradient vectors are used for periodical action partitioning.
(3) Once a complete exercise cycle is detected, two key frames and their corresponding MHI and MEI are selected to encode this movement.
(4) Use the PCOG descriptor to characterize both MHI and MEI. The concatenation of descriptors constitutes the representation of an action.
(5) Finally, the current action class is predicted by classifying the extracted feature descriptors (PCOG) using the offline trained classifier.

The classifier is trained offline by using the multi-class Support Vector Machine (SVM) (Cristianini and Taylor, 2000) with the RBF kernel. The training phase follows the same steps illustrated above: select the ROI, calculate the MHI, MEI and build the feature descriptor for training. To achieve the best training result, training sequences are partitioned into cycles manually with respect to the duration of each movement.

### 3.2. Temporal human action segmentation

To detect different actions from a video sequence, we need an automatic method to partition the input video sequence into cycles. A cycle is defined as a set of continuous frames, which forms a single and complete action. In this paper we propose two different methods for temporal human action partition.

#### 3.2.1. Color intensity based method

Color intensity information is used to partition the video sequence into cycles. To be specific, the video sequence is segmented by observing the intensity change within a specified region. To achieve this goal, we first convert the RGB image to grayscale, then choose the ROI manually. Intensity values of all pixels within this region are accumulated to present the total color intensity of the current frame. When the moving body part enters/leaves the ROI, the intensity value within the ROI will change immediately. To gain the maximum intensity variation, the ROI should be chosen properly. An example is shown in Fig. 2. The blue rectangular box indicates the manually selected ROI, and the periodic exercise (push up) passed through the pre-selected region during every

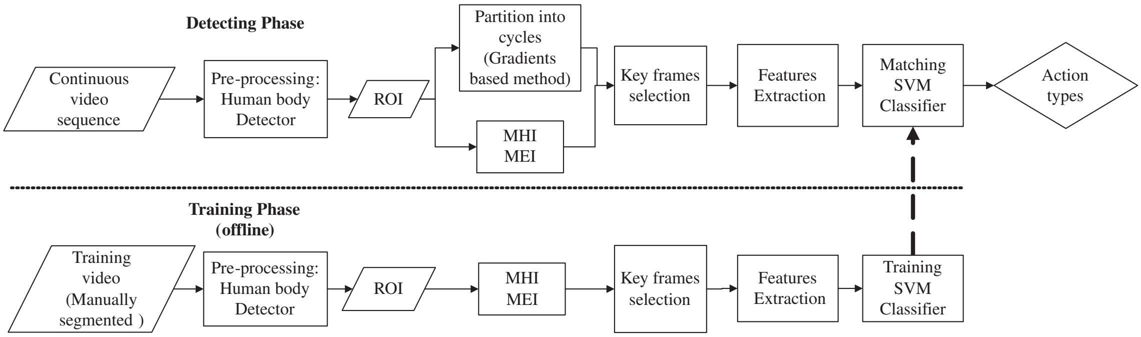


**Fig. 1.** Flowchart of the detection system.

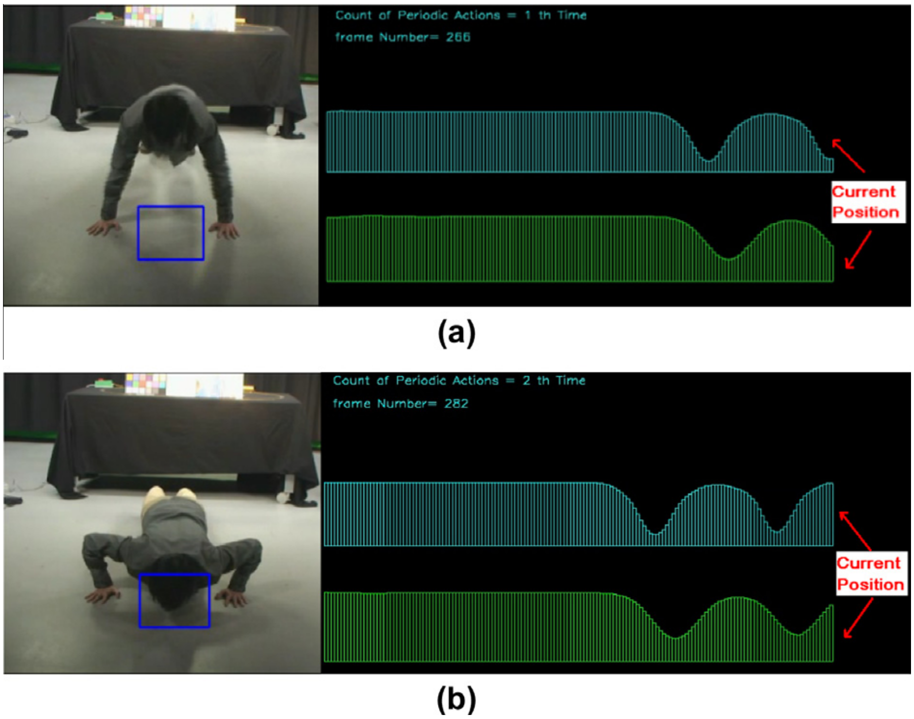

Count of Periodic Actions = 1 th Time
frame Number= 266
Current Position
(a)
Count of Periodic Actions = 2 th Time
frame Number= 282
Current Position
(b)

**Fig. 2.** (a) Before the moving body part (exerciser's head) enters the ROI, most pixels are gray, which corresponds to a high luminance value in the curve. (b) When the moving body part enters the ROI, the luminance value will drop due to the color differences between that body part (black hair) and the background.

exercise cycle. If the ROI is selected outside the range of the movement, there will be no observable luminance changes.

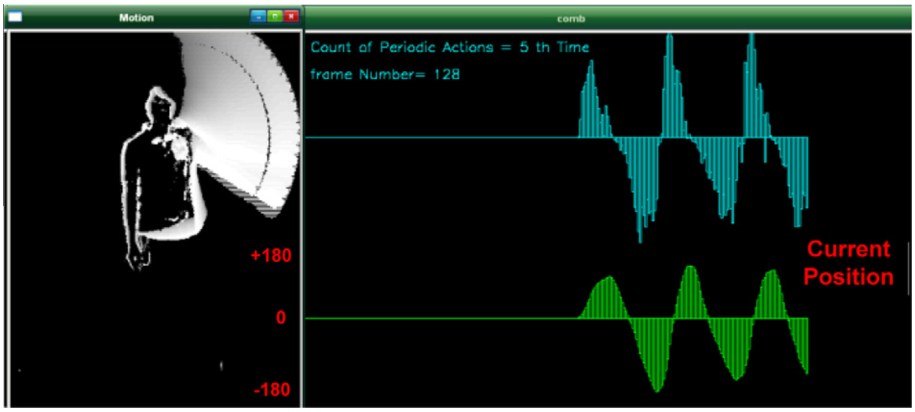
The blue[1] curve in Fig. 2 shows how the intensity value varies in time. Since most of the time the luminance value does not change linearly, we apply a "running average window" to convert the raw data to a smooth curve. Each time when the body part passes through the ROI, the luminance changes are recorded. In this way the input video sequence can be partitioned into cycles by searching the local maxima/minima. The main advantage of the color intensity based method is its low computational cost, which makes it suitable for different applications.

### 3.2.2. Motion based method

Most of the indoor exercises in gym, such as push up, squatting, hand waving, etc. consist of notable motions in either the vertical or the horizontal direction. Experimental results show that the periodic direction changes provide useful information for temporal human action segmentation. To obtain the direction information, we can use optical flow (Bradski and Davis, 2002) for motion estimation. In our approach, we use motion gradients, which is faster than optical flow. Orientations of the gradient vectors are then quantized into four directions: upwards, downwards, forwards and backwards. Gradient vectors are weighted by their orientations and magnitudes. For instance, a vector with magnitude 5 and direction upwards will be assigned a value +5. A vector with magnitude 1 and direction downwards will have a value $-1$. The global orientation in the vertical/horizontal direction is calculated by accumulating the gradient vectors in corresponding directions. Fig. 3 shows the sinusoidal motion patterns, which indicates how the global orientation varies in time. The overall motion patterns tend to be periodic when a periodic action is performed. The video sequence can be partitioned into cycles by locating the local maxima/minima on the sinusoidal curve.

The motion based method does not rely on the manually selected ROI and provides excellent segmentation results, if the background and illumination conditions are stable. Using motion

[1] For interpretation of color in Fig. 2, the reader is referred to the web version of this article.



Count of Periodic Actions = 5 th Time
frame Number= 128

**Fig. 3.** Positive parts on the curve indicate the upwards action negative parts on the curve refer to the downwards action.

gradient vectors instead of optical flow also highly reduces the computational cost.

### 3.3. Feature description

Our approach describes an action by key frames and their corresponding motion templates, which encode the temporal information by integrating the motion histories (MHI) into an image and represent this image by its local shape and the spatial layout of the shape. Key frames that contain the most complex motion are selected using the method in (Shao and Ji, 2009). The motion history image contains the trajectory information of the action being performed and recent motion is more emphasized than that happened in the past, which encapsulates the order and tendency of the motion. Further description of the shape and intensity of the MHI gives an overall representation of the action.

Traditional color histogram captures only the color distribution in an image and does not include any spatial correlation information. Images with very different appearances can have similar histograms. Therefore, adding spatial information into the histogram-based representation is important for refining color histogram based methods. Huang (1998) proposed color correlogram as a new color feature for image indexing/retrieval and it is proven to be an effective tool for the description of image content in the field of image retrieval (Huang, 1998; Hsu et al., 1995).

A color correlogram expresses how the spatial correlation of pairs of colors changes with distance $d$. For example, if we quantize an image $I$ into $m$ color bins, the color correlogram of $I$ is defined for $C_i, C_j \in [1, \ldots, m], d \in [1, \ldots, k]$ as:

$$\gamma^{(k)}_{C_i,C_j}(I) = \Pr(P_2 \in C_j | P_1 \in C_i, |P_1 - P_2| = k) \tag{1}$$

Given any pixel of color $C_i$ in the image $I$, $\gamma^{(k)}_{C_i,C_j}(I)$ gives the probability that a pixel at distance $d = k$ away from the given pixel is of color $C_j$. Fig. 4 illustrates an example of forming a correlogram descriptor at distance $d = 1$ schematically. Note that for an image with $m$ color bins, the size of the correlogram is $m^2 \times d$. While 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. A small $d$ might compromise the quality of the feature. We consider this tradeoff in the experiment part.

To add spatial information to our descriptor, we apply correlogram on the gradients, which captures the spatial co-occurrence of a pair of the gradient sets (COG). A COG descriptor is therefore a matrix instead of a histogram as in HOG. It incorporates the shape information (gradients) as well as the spatial correlations (relationship between gradients in neighboring locations) and is robust to small geometric deformation.

Since 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 COG are: (i) it includes the spatial correlation of orientations; (ii) it can be used to describe the global distribution of local spatial correlation of orientations if $D$ is chosen to be local.

Since 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 shape. To preserve the rough structure of shape the MHI is divided into sub regions. The idea of pyramid HOG is illustrated in Fig. 5. The PCOG descriptor consists of a correlogram of orientation 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, $l \in [1, \ldots, L]$, we divide the frame along $X$ and $Y$ dimensions into $2^{2\times l}$ sub-regions. Each cell can be described as the histogram of the weighted motion features in it. We can concatenate these histogram representations from all cells in all levels into a long histogram as the representation for the key frames.

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

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

.The final PCOG representation is the weighted concatenation of COGs at different scale levels.

## 4. Experiments and results

In this section, we first show the results of the temporal segmentation. A video sequence is automatically partitioned into separated cycles, which are compared with the ground truth. The segmented actions are then fed into the classifier for recognition. The performance of the PCOG descriptor is evaluated against two other frequently used methods.

### 4.1. Dataset and experimental setup

In this paper we build our own dataset for two reasons: (1) since our detection framework is designed to detect different exer-

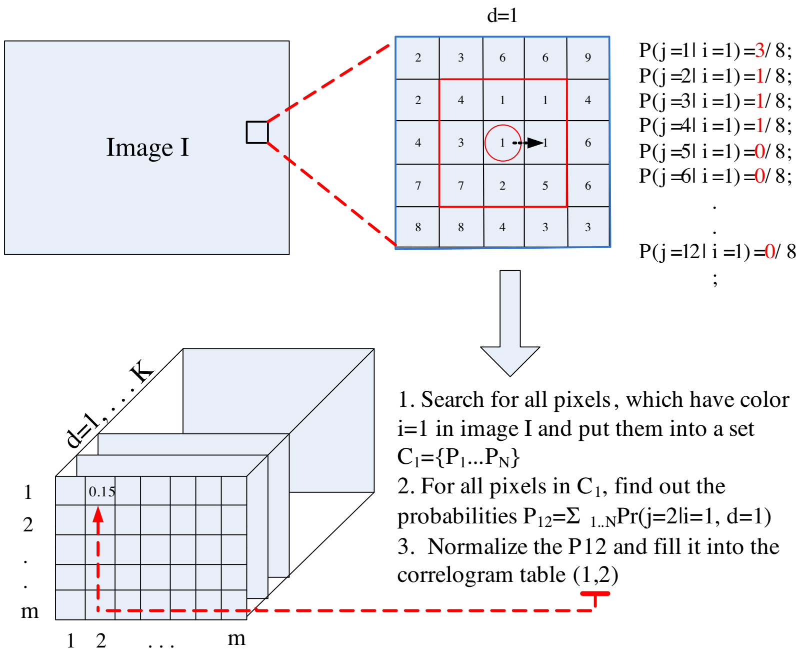

d=1
Image I
P(j =1| i =1) =3/ 8;
P(j =2| i =1) =1/ 8;
P(j =3| i =1) =1/ 8;
P(j =4| i =1) =1/ 8;
P(j =5| i =1) =0/ 8;
P(j =6| i =1) =0/ 8;
P(j =12| i =1) =0/ 8
d=1, . . . K
0.15
1. Search for all pixels, which have color i=1 in image I and put them into a set C1={P1...PN}
2. For all pixels in C1, find out the probabilities P12=Σ 1..N Pr(j=2|i=1, d=1)
3. Normalize the P12 and fill it into the correlogram table (1,2)
1
2
m
1 2 . . . m

**Fig. 4.** (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*. $C_j \in [1, \ldots, 12]$ 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* = 1 from a pixel of color $C_j = 2$ in the image.

cise types in an indoor environment such as sports center and gym and there is no such video dataset available; (2) the performance of our approach will be compared with Kellokumpu's approach (Kellokumpu et al., 2005), for which the dataset is not publicly available; therefore we try to build a similar dataset used in (Kellokumpu et al., 2005). We recorded video sequences of 8 indoor fitness exercises performed by 20 different subjects. This new dataset will be called Dataset A in the experiments.

Three views of the action (−30°, 0°, 30°) were recorded at two different scales (close to or far from the camera). We also select some indoor exercises including handwaving and boxing from the KTH (Schuldt et al., 2004) and the Weizmann (Blank et al., 2005) datasets as testing sequences. Exercises and their corresponding MHI are shown in Fig. 6, from left to right are: Handwaving, Right-handwaving, Handclapping, Reverse Cable Fly, Push Up, Jacking, Squatting, and Wood chopping.

The detection framework was implemented in C and Matlab environments. All video sequences were taken by a Logitech Web Camera with a resolution of 640 × 480 pixels at approximately 24 fps. Details of the detection system and the operational steps are given in Section 3.1.

### 4.2. Temporal human action partition

In this experiment we want to examine the accuracy of the proposed segmentation methods. A video sequence with one subject performing eight different exercises is first manually cropped into separate cycles, as shown in Fig. 7(a). The time stamps of all cycles in this sequence are recorded as the ground truth. After that the motion based method is applied to partition the same video sequence into cycles. Fig. 7(b) shows the MHI of the consecutive movements segmented from the same video sequence.

The experiment is repeated for all video sequences in the datasets and the total number of cycles and the start and end positions in each video sequence are recorded. From Table 1, we can see that our method can count the number of cycles precisely in all the three datasets used. The tolerance of the exact positions of each movement is also allowable.

### 4.3. Action detection in video

In this scenario action detection is defined as partitioning the different actions into cycles and recognizing the action classes simultaneously. The motion based method is adopted for temporal segmentation and the PCOG descriptor is used for recognition of actions. In this experiment we use the concatenation of MHI and MEI as motion templates, because they provide better discriminativity than individual ones. The MHI or MEI based features were also evaluated alone, but the number of false alarms was much higher than the combined features. This shows that both temporal (motion gradients) and shape information plays a significant role in detecting and recognizing human activities.

The action detection framework described in (Kellokumpu et al., 2005) is used for further comparison. In their work, the detection system is invariant to handedness of performing actions, for example, waving the right hand is considered the same as waving the left hand. In our approach these actions have different features. Thus instead of detecting 8 different movements, we can actually detect 11 different activities. The recognition and detection rates are defined as:

$$\text{recognitionrate} = \frac{\text{number of correct detection}}{\text{number of actions in video}} \tag{3}$$

$$\text{detectionrate} = \frac{\text{numbers of correct detection}}{\text{numbers of detections in videos}} \tag{4}$$

To show the better discriminativity of COG as a global representation, we compare the COG descriptor with two other frequently used global feature descriptors: Hu's moment invariants and the HOG descriptor at the finest level (whole image). Table 2 shows that COG has the highest recognition rate of 83%. Hu's moment invariants only produce a feature vector with a length of 7, which is usually too few for recognizing multiple classes. The increase in recognition rate from HOG to COG demonstrates that by taking into account the correlation between neighboring gradients the descriptor tends to be more informative.

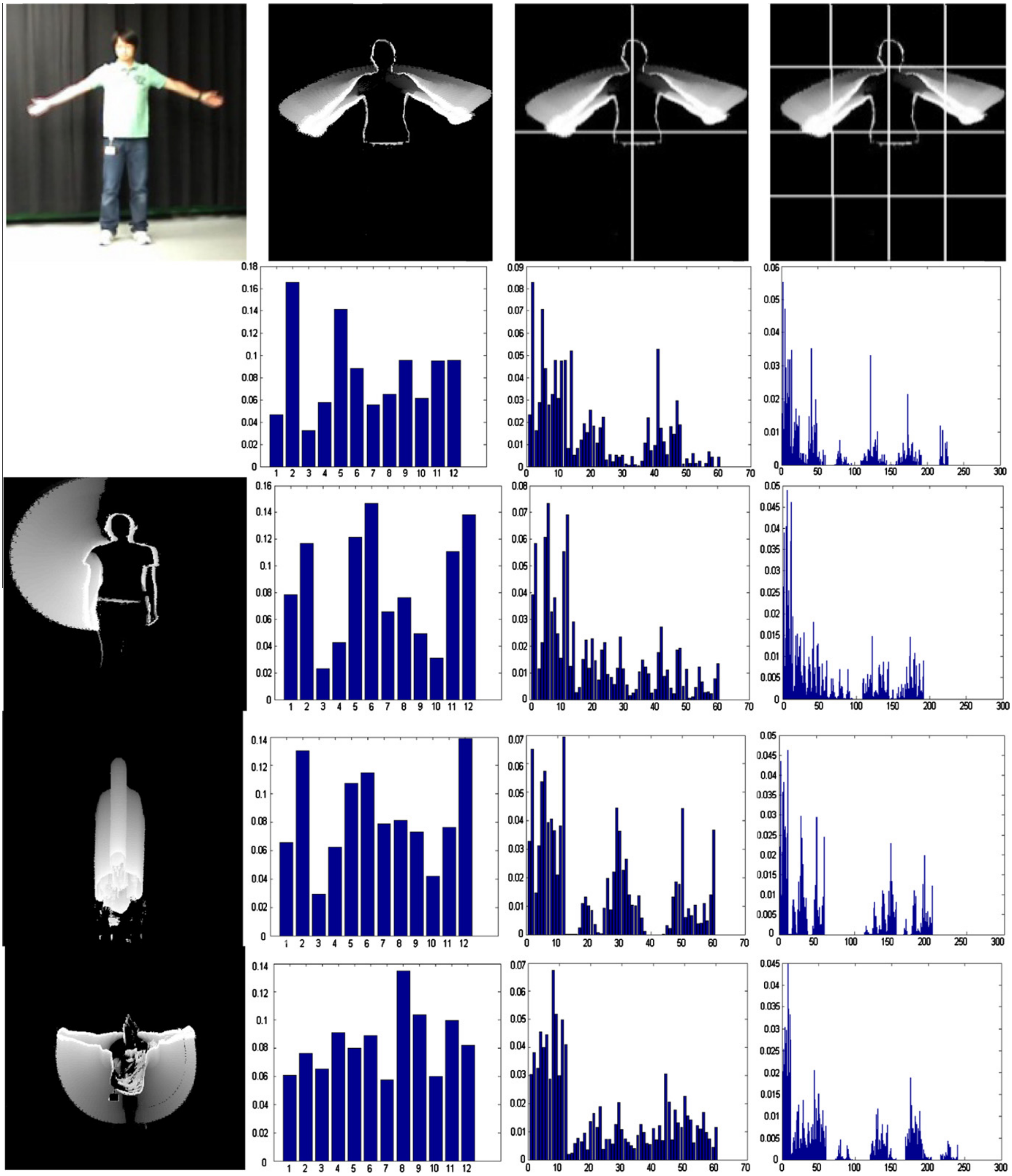


**Fig. 5.** Pyramid representation of MHI, at layer $L = 0$, l, 2. The first row illustrates the pyramid divisions at different levels on a certain action and their corresponding HOG descriptors. The remaining rows show HOG descriptors using different layers on different actions.

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 Table 2 we can observe that both methods achieve a higher recognition rate when the pyramid kernel is applied. At layer $L = 2$, PCOG outperforms PHOG by 1%. The gain of PCOG over PHOG decreases when more layers are used. This is because adding more layers also contributes to the spatial distribution information of the gradients, which makes the contribution from the correlogram less. Considering the computational overhead of calculating correlograms, PHOG can sometimes be a better choice when multiple layers are applied.

The total number of actions in test videos is 106, the number of detections is 112, and the number of correct detections is 104, giving a recognition rate of 94.6% and detection rate of 92.8%. Table 3

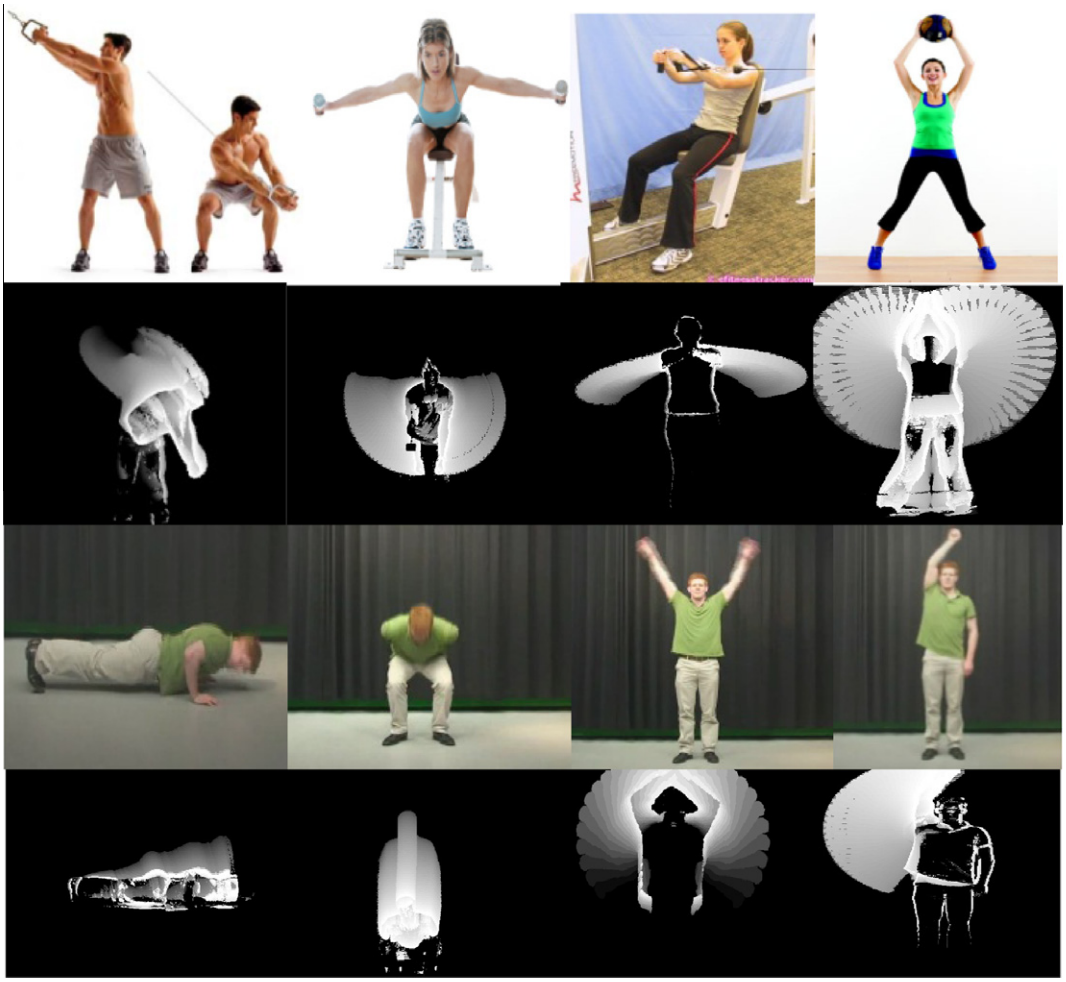


Fig. 6. Different indoor exercises and their corresponding MHI.

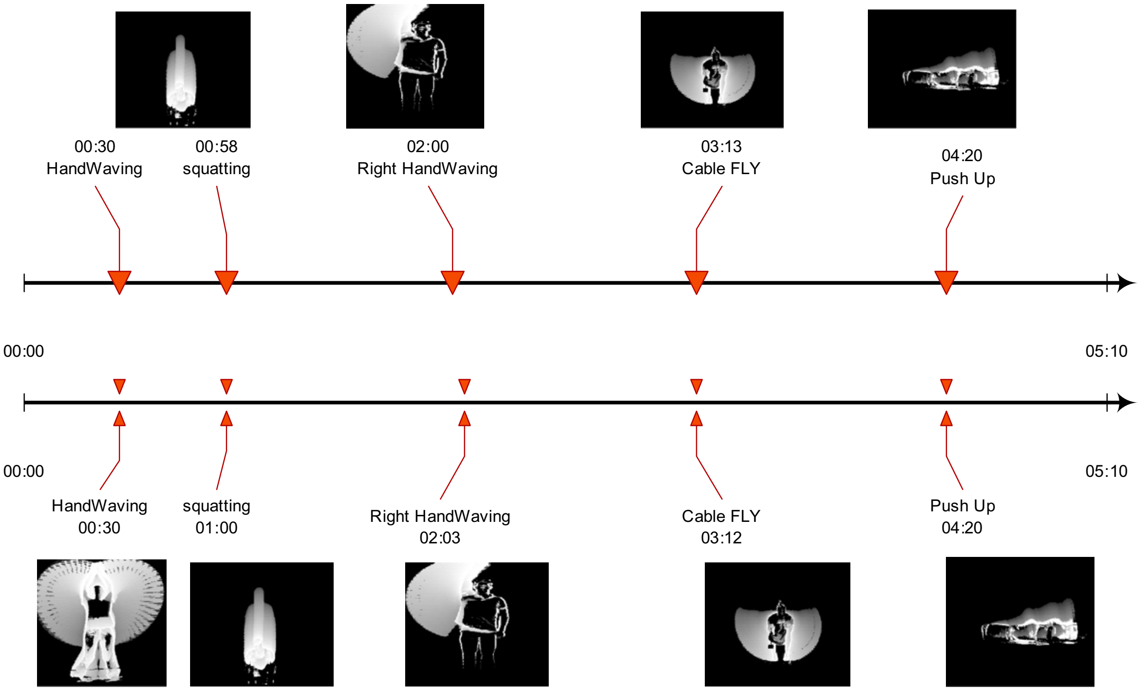


Fig. 7. MHI of the consecutive movements from the input video. (a) Top row shows the ground truth and (b) bottom row shows the result of our approach.

**Table 1**
Accuracy of action segmentation.

| | Dataset A (%) | Weizmann (Blank et al., 2005) (%) | KTH (%) |
|---|---|---|---|
| Total cycles | 100 | 100 | 100 |
| Tolerance | 1.5 | 1.3 | 2 |

**Table 2**
Recognition rates 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 |

**Table 3**
Detection rates in different layers.

| | Hu's moments (%) | PHOG (%) | PCOG (%) |
|---|---|---|---|
| $L = 0$ | 65.0 | 72.5 | 80.6 |
| $L = 1$ | – | 84.0 | 87.8 |
| $L = 2$ | – | 91.0 | 92.8 |

depicts the detection rates of the three methods. The number of detections exceeds the actual number of actions in test sequences because the test video sequences contain some irregular actions, for example, adjusting postures when they are changing the action types. Since the recognition and detection rates reported in (Kellokumpu et al., 2005) is 90% and 83%, respectively, from Table 3 we conclude that our approach performs better.

The system is first trained using a single viewpoint and one scale (distance to the camera). Experimental results show that the system tolerates some change in the viewing direction and different scaling. Different exercise types are still recognized even with the change of ±30° in the viewing angle. By adding more training samples with different viewpoints (−30°, 0°, 30°), the range can be extended to ±45°. The detection rate slowly starts to fall as the distance between exercisers and the camera increases due to insufficient motion information available.

## 5. Conclusion

In this paper we present a framework, which is able to detect different continuous human actions in real-time using a single stationary camera as input. The detection is based on describing the actions by key frames and the corresponding shape-based feature descriptor. The PCOG descriptor simultaneously integrates the spatio-temporal relations among local motion features with their shape information and embeds this rich information in the representation for the key frames.

The motion based temporal segmentation method achieved 100% detection rate if there is no intentional pause or any irregularities in the test videos. By using local properties (COG), our representation captures the essential information of human movements and allows variation in the performance of activities while still preserving discriminativity. The recognition performance is clearly improved by adding the spatial layout information to the descriptor. Experiments on action recognition and detection clearly show the effectiveness of the system.

## References

Aggarwal, J.K., Cai, Q., 1999. Human motion analysis: A review. Comput. Vision Image Understand. 73 (3), 428–440.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R., 2005. Actions as space-time shapes. In: IEEE Internat. Conf. on Computer Vision, vol. 2, pp. 1395–1402.

Bobick, A.F., Davis, J.W., 2001. The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Machine Intell. 23 (3), 257–267.

Bradski, G.R., Davis, J.W., 2002. Motion segmentation and pose recognition with motion history gradients. Machine Vision and Applications 13 (3), 174–184.

Cristianini, N., Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.

Dalal, N., Triggs, B., 2006. Histograms of oriented gradients for human detection. In: IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893.

Dollar, P., Rabaud, V., Cottrell, G., Belongie, S., 2005. Behavior recognition via sparse spatio-temporal features. In: VSPETS.

Efros, A.A., Berg, A.C., Mori, G., Malik, J., 2003. Recognizing action at a distance. In: The Ninth IEEE Internat. Conf. on Computer Vision. Nice, France.

Gavrila, D.M., 1999. The visual analysis of human movement: A survey. Comput. Vision Image Understand. 73 (1), 82–98.

Grauman, K., Darrell, T., 2005. The pyramid match kernel: Discriminative classification with sets of image features. In: ICCV, vol. 2, pp. 1458–1465.

Gu, J., Ding, X., Wang, S., Wu, Y., 2008. Full body tracking-based human action recognition. In: ICPR, pp. 1–4.

Hsu, W., Chua, T.S., Pung, H.K., 1995. An integrated color-spatial approach to content-based image retrieval. In: ACM Multimedia Conference, pp. 305–313.

Huang, J. 1998. Color-Spatial Image Indexing and Applications. Ph.D. Thesis, Cornell University.

Kellokumpu, V., Pietikäinen, M., Heikkilä, J., 2005. Human activity recognition using sequences of postures. In: IAPR Conf. on Machine Vision Applications, pp. 570–573.

Laptev, I., 2005. On space-time interest points. Int. J. Comput. Vision 64, 107–123.

Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B., 2008. Learning realistic human actions from movies. In: CVPR, Anchorage, USA.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR, pp. 2169–2178.

Liang, Y., Shih, S., Shih, C., Liao, H.-Y.M., Lin, C.-C., 2009. Learning atomic human actions using variable-length Markov models. IEEE Trans. Systems Man Cybernet. Part B Cybernet. 39 (1), 268–280.

Mahalanobis, P.C., 1936. On the generalized distance in statistics. The National Institute of Sciences of India 2 (1), 49–55.

Niebles, J.C., Wang, H., Fei-Fei, L., 2006. Unsupervised learning of human action categories using spatial–temporal words. In: BMVC.

Nister, D., Stewenius, H., 2006. Scalable recognition with a vocabulary tree. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2161–2168.

Schuldt, C., Laptev, I., Caputo, B., 2004. Recognizing human actions: A local svm approach. In: Internat. Conf. on Pattern Recognition, vol. 3, pp. 32–36.

Shao, L., Ji, L., 2009. Motion histogram analysis based key frame extraction for human action/activity representation. In: The 6th Canadian Conference on Computer and Robot Vision, Kelowna, Canada.

Shao,L., Ji, L., 2010. A descriptor combining MHI and PCOG for human motion classification. In: ACM Internat. Conf. on Image and Video Retrieval (CIVR), Xi'an, China.

Thurau, C., 2007. Behavior histograms for action recognition and human detection. In: HUMO, pp. 299–312.

Weinland, D., Ronfard, R., Boyer, E., 2006. Free viewpoint action recognition using motion history volumes. Comput. Vision Image Understand. 104 (2), 249–257.

Willamowski, J., Arregui, D., Csurka, G., Dance, C.R., Fan, L., 2004. Categorization nine visual classes using local appearance descriptors. In: IWLAVS.

Zhou, H., Tao, D., Yuan, Y., Li, X., 2009. Object trajectory clustering via tensor analysis. In: IEEE Internat. Conf. on Image Processing, Cairo, Egypt.