Pattern Recognition Letters 31 (2010) 639–647

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

# View-independent human action recognition with Volume Motion Template on single stereo camera ☆

Myung-Cheol Roh, Ho-Keun Shin, Seong-Whan Lee *

*Department of Computer Science and Engineering, Korea University, Anam-dong, Seongbuk-ku, Seoul 136-713, Republic of Korea*

ARTICLE INFO

*Article history:*
Received 5 November 2008
Received in revised form 5 October 2009
Available online 4 December 2009

Communicated by A. Shokoufandeh

*Keywords:*
View-independence
Human action recognition
Volume Motion Template
Motion History Image

ABSTRACT

Vision-based human action recognition provides an advanced interface, and research in the field of human action recognition has been actively carried out. However, an environment from dynamic viewpoint, where we can be in any position, any direction, etc., must be considered in our living 3D space. In order to overcome the viewpoint dependency, we propose a Volume Motion Template (VMT) and Projected Motion Template (PMT). The proposed VMT method is an extension of the Motion History Image (MHI) method to 3D space. The PMT is generated by projecting the VMT into a 2D plane that is orthogonal to an optimal virtual viewpoint where the optimal virtual viewpoint is a viewpoint from which an action can be described in greatest detail, in 2D space. From the proposed method, any actions taken from different viewpoints can be recognized independent of the viewpoints. The experimental results demonstrate the accuracies and effectiveness of the proposed VMT method for view-independent human action recognition.



## 1. Introduction

The time of robot assistance in human life is clearly visible in the near future, through the active research and development of a humanoid robot. An essential key which enables robots to assist humans is the interactions between human and robot. Research on human–robot interaction provides convenient and natural interfaces. Among many sensors, human action recognition based on a vision sensor provides an advanced interface and research on vision-based interaction has been actively studied.

We live in 3D space and human actions are performed in 3D space. This means, that for practical purposes, we must consider an environment from a dynamic viewpoint, where we can be in any position, any direction, etc. Many researchers have developed reconstruction methods for the 3D human body model, from input video sequences (Ben-Arie et al., 2002; Mori et al., 2004; Ramanan and Forsyth, 2003; Sminchisescu and Triggs, 2003; Yang et al., 2007). However, 3D-based methods have high computational cost and do not yet provide a precise enough solution to be used in practical environments.

The alternative method is to use the 2D template-based method, which is much simpler and requires lower computational costs than the model-based method. One of the representative methods is the MHI (Motion History Image) (Bobick and Davis, 2001). However, the template of human action changes dynamically depending on camera viewpoints and the speed of human action. Thus, for 2D-based human action recognition, the primary concern is the camera viewpoints problem. The view dependent problem is difficult to solve under a *single* camera environment, since the shape and the motion information, which depend on viewpoint, vary too greatly to represent actions efficiently.

Firstly, features for human action representation change drastically with changes in viewpoint. View-based methods suffer greatly from the changes of viewpoint. 2D-based methods have limitations in covering actions from different viewpoints. Also, the representation of an action has limitations, because the data obtained from a camera consists of the information about the projection from a 3D real world into a 2D image plane. For an example, two actions of 'Moving a hand forward' and 'Moving a hand backward', are barely distinguishable using MHIs, because the MHIs are generated on 2D images. However, MHI effectively describes orthogonal motions to the optical axis.

Secondly, variation of an action speed induces poor recognition performance. The maximum and minimum duration are manually determined in the MHI method and all MHIs are generated in a range between the maximum and minimum duration. The disadvantage of this method is that the greater the number of actions and the range, the greater the number of templates required. In addition, computational cost is increased and contrast is reduced. The solution to this problem involves using a different capture rate.

☆ A preliminary version of this paper has been presented at the 18th International Conference on Pattern Recognition, Hong Kong, August 2007.

* Corresponding author. Tel.: +82 2 3290 3197; fax: +82 2 926 2168.
*E-mail addresses:* mcroh@image.korea.ac.kr (M.-C. Roh), hkshin@image.korea.ac.kr (H.-K. Shin), swlee@image.korea.ac.kr (S.-W. Lee).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.11.017

The Timed Motion History Image (TMHI) is presented in (Bradski and Davis, 2002). The current time stamp is used as an intensity of motion template. However, the variation in the velocities of actions is not dealt with.

In order to overcome these problems, a Volume Motion Template (VMT), which is an extension of the MHI to 3D space, is proposed, based on a stereo camera. The VMT is generated by using motion history information in 3D space, using disparity maps, and the VMT is projected into a 2D plane that is orthogonal to an optimal virtual viewpoint. The optimal viewpoint is a viewpoint from which an action can be described in greatest detail in 2D space. The problem of variation of action speed is solved by the proposed temporal normalization method. Note that the VM (Volume Motion) in this paper does not mean an actual 3D volume but a *virtual* volume of motion that reconstructed from a pair of images by a stereo camera.

## 2. Previous work

The previous human action recognition research can be separated into two main categories: model-based and appearance-based methods. The model-based method analyzes components of the human body from the input images (Ben-Arie et al., 2002; Mori et al., 2004; Ramanan and Forsyth, 2003; Sminchisescu and Triggs, 2003). In this method, it is important to fit 2D or 3D human body models to an input video sequence and to accurately extract the coordinates of each human body articulation. Sminchisescu and Triggs proposed an approach to fit a 3D model to an image, with covariance scaled sampling from a single camera input (Sminchisescu and Triggs, 2003). Mori et al. proposed an approach to recover a human body configuration (Mori et al., 2004). Yang et al. proposed a gesture spotting and recognition method based on a stereo camera and they used a 3D model that is reconstructed from a pair of stereo image (Yang et al., 2007). The performance of these approaches depends on the accuracy of extracting the coordinates of human body articulations. Analyzing components of the human body is very complex, difficult, and computationally expensive.

The appearance-based method is simple and requires lower computational costs than the model-based method (Bobick and Davis, 2001; Bradski and Davis, 2002; Kirishima et al., 2005; Lee and Kim, 1999; Shan et al., 2004; Ye et al., 2004). Bobic and Davis proposed a simple method for temporal template matching, the MHI (Bobick and Davis, 2001). The MHI is a static image template where pixel intensity is a function of the recency of motion history at that location. Brighter values correspond to more recent motion in a sequence. They performed the MHI method in aerobic sequences. However, its performance will be degraded for non-trained viewpoints. In addition, the MHI cannot represent parallel motion in an optical axis, such as moving a hand forward and backward.

Prior research for solving the viewpoint problem was as follows: Seitz and Dyer proposed a view-invariant method for detecting cyclic motion (Seitz and Dyer, 1997). Affine invariant principles are used, and it is assumed that feature correspondences between successive frames are known. Bodor et al. introduced an image-based rendering method for view-independent human motion recognition (Bodor et al., 2003). Orthogonal views to the mean direction of motion with rendering are generated. This approach, however, assumes a multiple camera environment. Niu and Mottaleb proposed view-invariant human activity recognition based on shape and motion features with hidden Markov models (Niu and Mottaleb, 2004). This approach performs experiments on walking and running actions and requires a set of hidden Markov models for every viewpoint. Rao et al. proposed a method for view-invariant representative actions of a human hand (Rao et al., 2002). Meaningful action is characterized as dynamic instants and intervals. Parameswaran and Chellappa proposed a method for estimating view-independent human body pose with the epipolar geometry of the camera (Parameswaran and Cellappa, 2005). It is assumed that a torso twist is small and head-orientation is used for recovering epipolar geometry. The system requires the correspondence of several body landmarks for input, which is a strong restriction in the real world. Yilmaz and Shah proposed a novel action representation method that is view invariant (Yilmaz and Shah, 2005). Although their method is robust to viewpoint changes, it cannot distinguish actions that the motions is parallel to a camera, such as moving a hand backward and forward. Ahmad and Lee proposed a human action recognition method that is robust to view changes (Ahmad and Lee, 2008). They used multidimensional features of combined local–global optic flow and shape flow. However, it also has a problem with parallel motions to a camera. Weinland et al. proposed a free viewpoint action recognition method using motion history volumes (Weinland et al., 2006). In order for free-viewpoint representation of actions, they used multiple calibrated and background-subtracted video cameras. Holte and Moeslund proposed view invariant gesture recognition method using 3D motion primitives (Holte and Moeslund, 2008). They represent temporal motion using view invariant shape descriptor and classification is performed using the sequence of the motion primitives. However, in order to represent the accurate shape context good quality of depth map is required. Shin et al. showed the very first model of Motion History Image in 3D (Shin et al., 2005).

In this paper, the VMT is proposed, to cope with both the limited capability of 2D images representation and the duration problem, and the view dependent problem. The VMT is a 3D extension of the MHI and requires only one stereo camera.

## 3. Volume Motion Templates and Projected Motion Templates

The overview of the proposed action recognition method is represented in Fig. 1. This section describes the detailed algorithms to generate the VMT and PMT (Projected Motion Template), which is a projected template image from the VMT, by an optimal virtual viewpoint.

### 3.1. Generating volume motion template

The VMT is proposed to cope with problems of 2D motion analysis:limited capability of 2D motion representation, and the view dependent problem. The proposed VMT is defined by a 3D motion template, which includes 3D motion history information, using disparity maps of stereo input sequences. Algorithm 1 shows the algorithm for generating the VMT. However, the 3D motion template is *virtually* reconstructed, because it is impossible to get *real* 3D volume from single stereo camera.

**Algorithm 1.** An algorithm for generating the VMT

1: Extract silhouette images of foreground images and disparity maps, using the correspondence of stereo input images
2: Calculate volume object, $O_t$, in 3D space, using the silhouette image and the disparity map
3: Calculate the motion difference, $\sigma_t$, and the magnitude of motion, $\mu_t$, between two consecutive volume objects in motion sequence
4: Construct the VMT, $V_t$, by adding new motion, $\sigma_t$, and attenuating intensity of previous motion information with the magnitude of motion

We used background subtraction for extracting foreground silhouette images from the video sequence. The volume object is a 3D binary image reconstructed from a stereo image, by using its disparity map. The instant of the volume object at time $t$ is calculated by

$$O_t(x,y,z) = \begin{cases} 1 & \text{if } S_t(x,y) = 1 \text{ and } Z_t(x,y) = z \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $S_t$ and $Z_t$ are a binary silhouette image and its disparity map, respectively. In order to estimate the action's motion, the difference between two consecutive volume objects at time $t$, are calculated. The difference of two consecutive volume objects is calculated by

$$\sigma_t(x,y,z) = |O_t(x,y,z) - O_{t-1}(x,y,z)|. \quad (2)$$

The VMT is constructed from a set of the $\sigma_j$ where $1 < j < t$. The voxel's intensity at a point, $(x,y,z)$, in the VMT, at time $t$ is defined by

$$V_t(x,y,z) = \begin{cases} I_{\max} & \text{if } \sigma_t(x,y,z) = 1 \\ \max(0, V_{t-1}(x,y,z) - \eta\mu_t) & \text{otherwise} \end{cases} \quad (3)$$

$$\mu_t = \int\int\int \sigma_t(x,y,z)\,dxdydz, \quad (4)$$

where $I_{\max}$ is maximum intensity (e.g., 255 in an 8 bit grayscale), and $\eta$ and $\mu_t$ are an attenuating constant and magnitude of motion. Motion history information disappearance rate is reduced in proportion to $\mu_t$. The attenuating constant controls the amount of disappearance of the motion history information. In the experiments, we set the $\eta$ as follows:

$$\eta = \frac{1}{N}\sum_{i=1}^{N}\eta_i, \quad (5)$$

$$\eta_i = \frac{I_{\max} - 1}{\sum_{t=1}^{T_i}\mu_t}, \quad (6)$$

where $N$ and $T_i$ are the number of training data and the length of the $i$th training data. The magnitude of motion is a sum of all elements in $\sigma_t$ that is defined in Eq. (2).

### 3.2. Spatial and temporal normalization of the VMT

The silhouette of a human can be located at any position in an image plane. In addition, silhouettes in a video can have various sizes, depending on the distance between the human and camera and on personal body sizes. The spatial normalization is performed according to the height of a dominant region and to the center of the gravity of silhouette. Silhouette images are resized according to the height of the body and are shifted to a given location according to the center of gravity.

Each person performs actions at different speeds. This variation results in degraded recognition performance. The conventional MHI produces different shapes for the same actions at different speeds, because of the duration problem. The problem is that prior motion data disappears over time, before an action is finished. This problem can be solved by controlling motion history information dynamically, using the magnitude of motion. In order to prevent the loss of prior motion information, the magnitude of motion is considered to maintain the history information. The magnitude of motion, $\mu_t$, is an important clue for solving the duration problem. When movement does not occur, or the magnitude of the

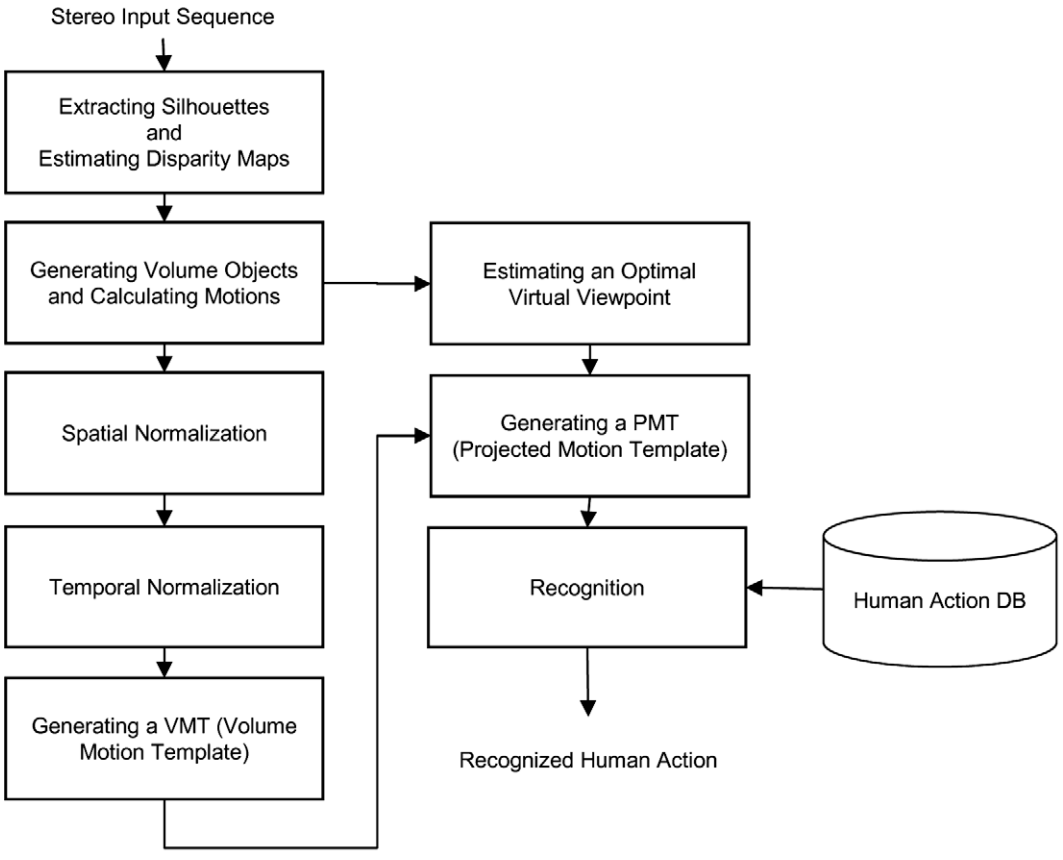


**Fig. 1.** An overview of the proposed VMT based action recognition method.

movement is small, such as in slow motion, history information is not updated. Therefore, motion history information can be maintained which is independent of the speed of movement.

### 3.3. Generating Projected Motion Template

As mentioned, the VMT refers to 3D space. It is not easy or efficient to match the generated VMT (3D image) with the reference VMT, whereas using the 2D template is much easier. Therefore, we proposed PMT (Projected Motion Template) for recognition. The PMT is a 2D image which is a projection image from 3D VMT, from an optimal virtual viewpoint. Algorithm 2 shows the algorithm for generating the PMT.

**Algorithm 2**. An algorithm for generating the PMT

1: Generate VMT and calculate motion orientation
2: Estimate an optimal virtual viewpoint by calculating an orthogonal motion to the motion orientation of the VMT
3: Generate a PMT (Projection Motion Template) image, $\rho$, from VMT from the optimal virtual viewpoint

Fig. 2 presents the overview of the means of generating VMT and PMT. Although the action involves moving a hand forward from a parallel viewpoint in an optical axis, the PMT can effectively describe motion information.

We define an optimal virtual viewpoint as the viewpoint from which an action can be described in greatest detail, in 2D space. The VMT is projected into 2D space by employing the optimal virtual viewpoint. For example, we can easily distinguish between the 'stretching arm forward' and 'stretching arm backward' actions by observing it from the side view (90° to the motion) rather than from the frontal view (0° to the motion). The virtual 90° is determined as an optimal virtual viewpoint.

The optimal virtual viewpoint for describing motion is an orthogonal orientation to the motion orientation. Therefore, determination of motion orientation is an important factor for estimating the optimal virtual viewpoint. Approaches for estimating the motion orientation do exist, such as optical flow and global motion orientation (Bobick and Davis, 2001). However, these demonstrate poor performance with respect to noise, since the disparity map obtained from the stereo camera is not reliable enough. This paper proposes a simple and robust method to estimate the motion orientation of action data with respect to noise, using time windows.

We can calculate a dominant motion orientation between two volume objects, at two different times, within a time window. Then, the optimal virtual viewpoint can be estimated from the orthogonal direction to the dominant motion. Fig. 3 represents the means to estimate dominant motion orientation between two input images at time $t$ and $t-w$. The dominant motion orientation is obtained by subtracting two volume objects at time $t$ and $t-w$. Firstly, the centers of gravity of new and old objects at time $t$ and $t-w$ are calculated. The new and old objects ($D_t^{new}$ and $D_{t-w}^{old}$, respectively) are defined as follows:

$$D_t^{new}(x,y,z) = \max(0, O_t(x,y,z) - O_{t-w}(x,y,z)) \tag{7}$$

and

$$D_{t-w}^{old}(x,y,z) = \max(0, O_{t-w}(x,y,z) - O_t(x,y,z)) \tag{8}$$

for all $(x,y,z)$. Let a moment vector of a volume object, $\mathscr{I}$, be

$$m(\mathscr{I}) = \frac{1}{\sum_x \sum_y \sum_z \mathscr{I}(x,y,z)} \sum_x \sum_y \sum_z \mathscr{I}(x,y,z)\overrightarrow{(x,y,z)}. \tag{9}$$

Secondly, the motion vector, $\vec{\delta}_t$, at time $t$ with time window $w$ is calculated by

$$\vec{\delta}_t = m(D_t^{new}) - m\left(D_{t-w}^{old}\right). \tag{10}$$

Thirdly, the rotation matrixes for the VMT are calculated by

$$R_x(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix}, \tag{11}$$

$$R_y(\beta) = \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix}, \tag{12}$$

$$R_z(\gamma) = \begin{pmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{13}$$

The angles are

$$\alpha = \cos^{-1}\left(-sgn(y)sgn(z)\frac{||\overrightarrow{\Delta_t}(y)||}{||\overrightarrow{\Delta_t}(y) - \overrightarrow{\Delta_t}(z)||}\right), \tag{14}$$

$$\beta = \frac{2}{\pi} - \cos^{-1}\left(-sgn(z)sgn(x)\frac{||\overrightarrow{\Delta_t}(z)||}{||\overrightarrow{\Delta_t}(z) - \overrightarrow{\Delta_t}(x)||}\right), \tag{15}$$

$$\gamma = \cos^{-1}\left(-sgn(x)sgn(y)\frac{||\overrightarrow{\Delta_t}(x)||}{||\overrightarrow{\Delta_t}(x) - \overrightarrow{\Delta_t}(y)||}\right), \tag{16}$$

where $sgn()$ is a sign indicate function and $\overrightarrow{\Delta_t}(k)$ is a projected vector of $\vec{\delta}_t$ into the $k$-axis.

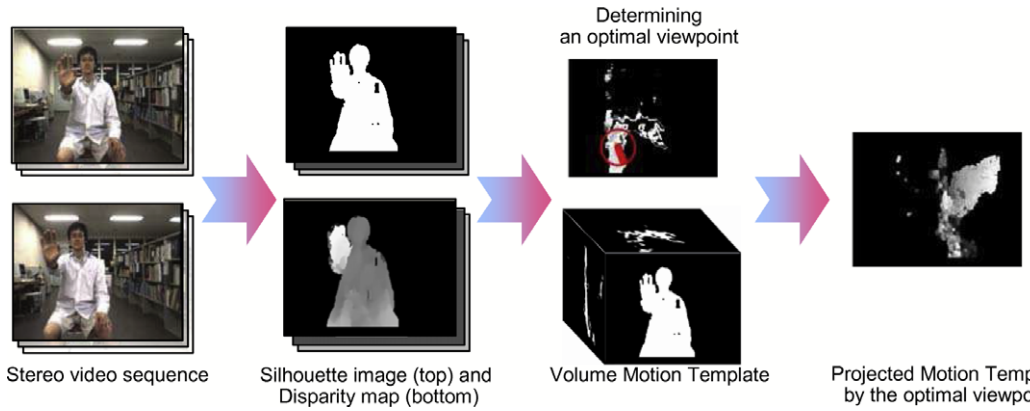


**Fig. 2.** An overview of the means of generating VMT and PMT. The bar in the red circle in third column image represents a motion orientation of the action.

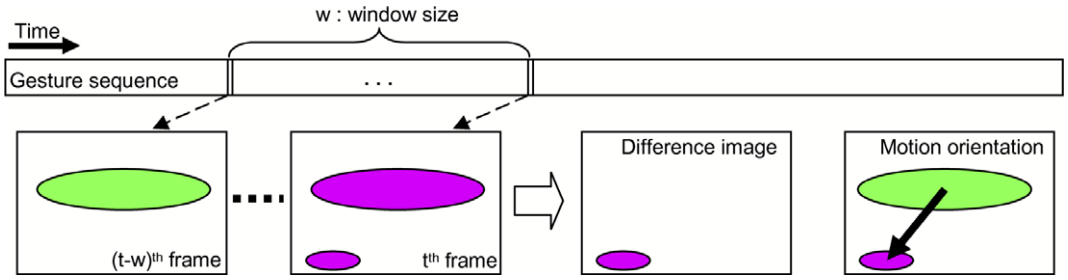


**Fig. 3.** Estimating motion orientation of movements, within a time window. The green and purple represent silhouette images at time $t - w$ and $t$, respectively, and the black arrow represents the estimated motion orientation.

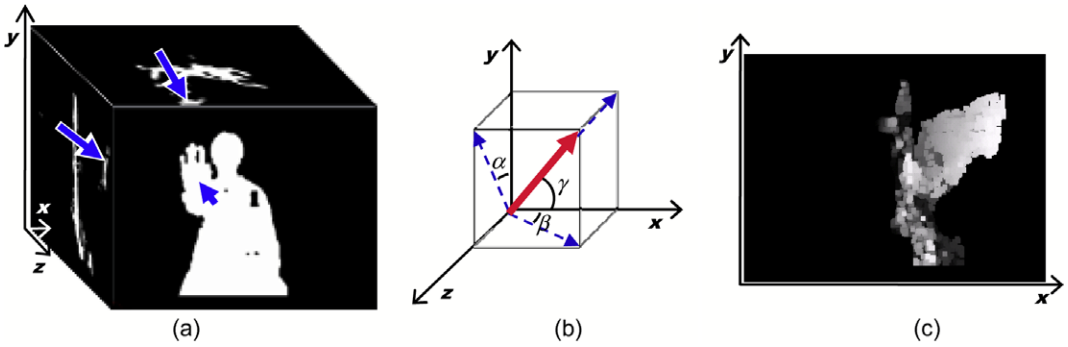


**Fig. 4.** Brief procedure for generating the PMT from the VMT. (a) VMT and motion vectors (blue arrows) in three planes ($xy$, $yz$, $zx$). (b) The angles for three planes and (c) PMTs which are projected into $xy$ plane, from the VMT, after rotating.

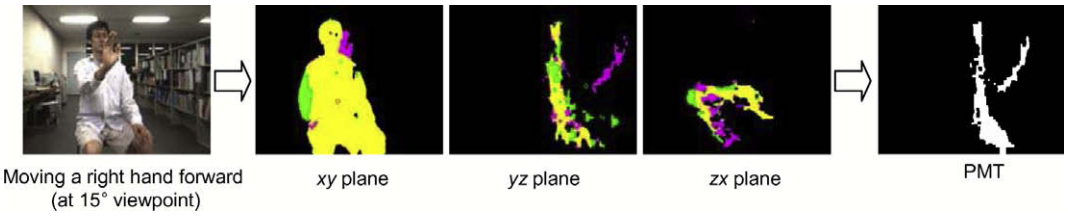


**Fig. 5.** Examples of projected images of a VMT into $xy$, $yz$, and $zx$ planes, and the PMT from an optimal virtual viewpoint, for an input action. The green and purple represent silhouette images at time $t - 1$ and $t$, and the yellow represents overlapped region, respectively.

Fourthly, the VMT is rotated by the matrix $(R_x(\alpha)R_y(\beta)R_z(\gamma))$ and the PMT is generated by projecting the rotated VMT into an $xy$-axis plane. Fig. 4 presents the brief procedure for generating the PMT from the VMT. Fig. 5 shows PMTs in $xy$, $yz$, and $zx$ planes, and the PMT from an optimal virtual viewpoint. Fig. 6 presents examples of actions (left-hand two images), PMTs (right-hand images), and its motion orientations (arrows on the right-hand images). The actions are; 'walking toward a stereo camera', 'sitting on a chair', 'bowing', and 'moving a right hand forward'.

## 4. Action recognition using the PMT

To recognize an action, we assume that a PMT of any action can be reconstructed from a linear combination of PMTs in a certain action class. This hypothesis is used in many fields and yields good results in recovering and reconstructing an image from noisy and distorted environments (Hwang and Lee, 2003). Because PMT includes noise which is caused by inaccurate depth information and silhouette images of input images, it is appropriate to use a recovery method from linear combination, rather than the simple template matching method for recognition. The training data set consists of $N$ PMTs for each action, as follows:

$$P_c = (p_1^c, p_2^c, \ldots, p_N^c), \quad \text{where } 1 \leqslant c \leqslant C, \tag{17}$$

where $N$ and $C$ are the number of training data sets (PMTs) for an action class, and the number of classes, respectively. $p_i^c$ represents a column vector of PMT of the $c$th action class.

From the assumption above, an input PMT, $\widetilde{P}$, can be reconstructed from linear combination, as follows:

$$\widetilde{P} \approx p_1^c\varepsilon_1 + p_2^c\varepsilon_2 + \cdots + p_N^c\varepsilon_N = P_c\varepsilon_c, \tag{18}$$

where $\varepsilon_c$ is a coefficient vector for the $c$th class.

The error function is defined by

$$E_c = \|\widetilde{P} - P_c\varepsilon_c\|. \tag{19}$$

An optimal coefficient $\varepsilon_c^*$ which minimize the error function can be estimated by

$$\varepsilon_c^* = (P_c^T P)^{-1} P^T \widetilde{P} = P_c^+, \widetilde{P}, \tag{20}$$

where $P^+$ is pseudo-inverse of $P_c$. Recognition is performed by finding a class $(P_c)$ which has highest correlation between the input PMT and the reconstructed PMT from the $P_c$.

In order to recognize gestures in a video sequence, where there are several gestures in a single video, a voting strategy is applied. We count the number of actions for each class within a time window. When one of the numbers has higher value than a given

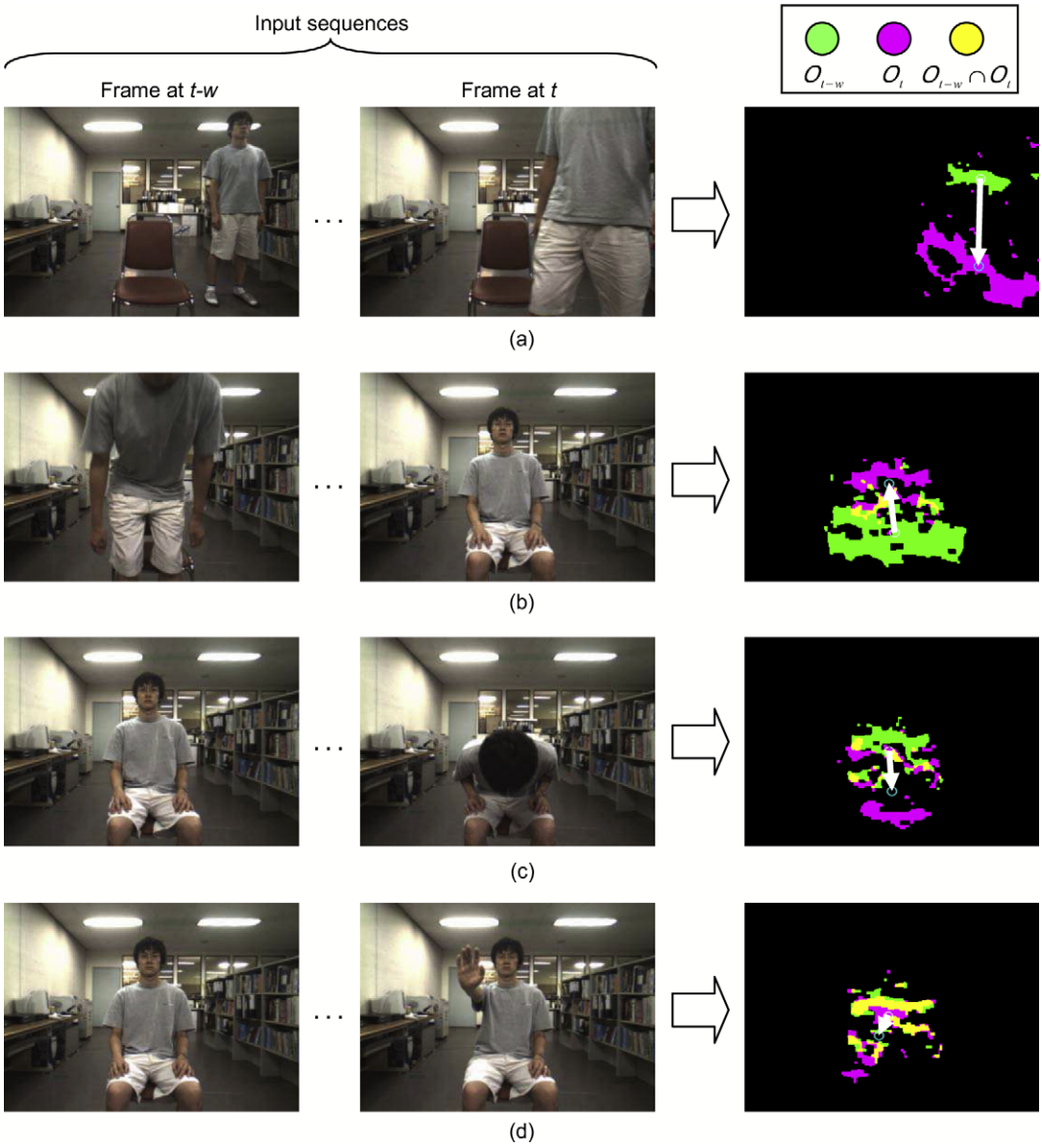


**Fig. 6.** Examples of estimation of motion orientation. The white arrows represent the dominant motion vectors for (a) walking toward camera, (b) sitting on a chair, (c) bowing, and (d) moving a right hand forward. The small white and blue circles represent the centers of gravity of $O_{t-w}$ and $O_t$.

threshold, the sequence within the time window is determined as the action corresponding to the number of the class.

## 5. Experiments and analysis

In order to make performance evaluations, we took videos from a stereo camera instead of using common action databases such as Weizmann action database (Gorelick et al., 2007), because they does not provide stereo videos. The training and testing data were captured by a stereo camera, Videre Design STH-MDCS2, with 4.8 mm focal length lenses. The resolution of the video sequence is of 320 × 240 24bit RGB color images, with a rate of 30 frames per second.

To measure the performance of the proposed VMT, two experiments were carried out.[1] The first experiment measures the extent of the VMT view change robustness. We measured the viewpoint change robustness, using various video data sets which are taken from different viewpoints. The second experiment measures consistency for various actions which are taken from the frontal viewpoint.

### 5.1. Viewpoint change robustness

Four actions are taken from seven different viewpoints (0°–90° for every 15° intervals). The actions are; 'moving hand forward', 'moving hand backward', 'bowing', 'raising a right hand'. Because the data taken at the −15° to −90° viewpoints are symmetrical to the data at 15°–90°, we anticipate that the result will be the same. Fig. 7 represents the PMTs and the MHIs of the four actions, from seven viewpoints. Fig. 7a–d shows the input videos (first rows), VMTs (second rows) and MHIs (third rows) for the actions; 'moving hand forward', 'moving hand backward', 'bowing', 'raising a right hand', respectively. The PMTs for an action from different viewpoints show stable images, whereas the MHIs change drastically as the viewpoint changes. Fig. 8 represents the correlation

[1] Example video clips are available at http://image.korea.ac.kr/3DGD/videoclips.html.

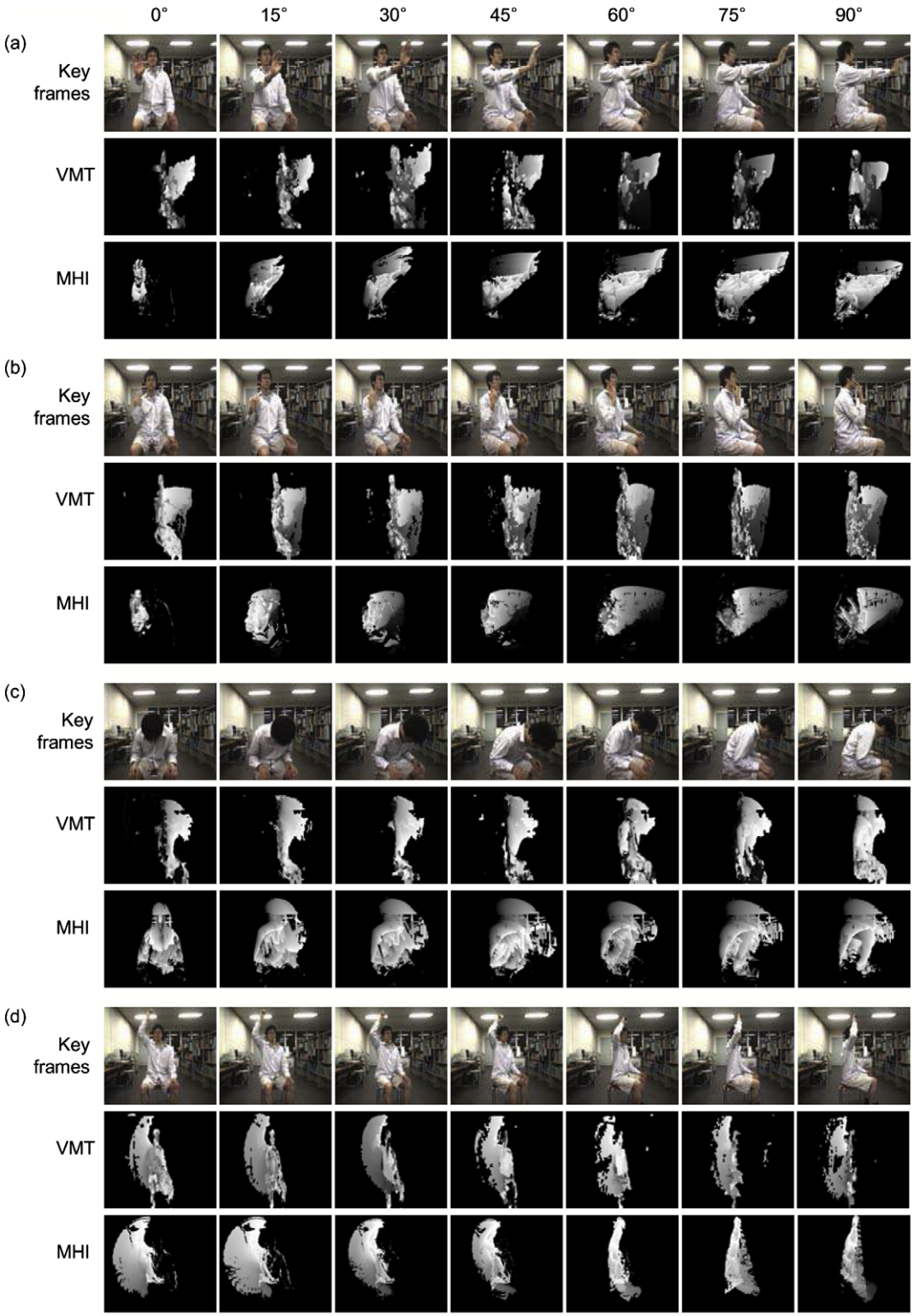


**Fig. 7.** Examples of input sequence, PMT and MHI for four actions taken from seven different viewpoints. The actions are; '(a) moving a right hand forward', '(b) moving a right hand backward', '(c) bowing', and '(d) raising a right hand'. The first rows represent representative frames of the video sequences. The second and third rows represent PMTs and MHIs.

coefficients between a template from the reference viewpoint (0°) and templates from the other viewpoints. The plots demonstrate that the proposed action representation is stable, despite dynamically changing viewpoints, whereas the similarity of MHIs changes drastically as the viewpoint changes.

The duration times of the actions are varying. The approximated duration time for each action is as follows: 'moving hand forward ($1.8 \pm 0.2$ s)', 'moving hand backward ($2.0 \pm 0.2$ s)', 'bowing ($6.1 \pm 0.3$ s)', and 'raising a right hand ($2.5 \pm 0.5$ s)'. The variation of the action duration of each action is not large. Thus, in the first

experiment, there is no significant difference between the results of the MHI method and the proposed VMT method using the proposed temporal normalization. However, there are differences in the second experiment.

### 5.2. Consistency for various actions

Ten actions are performed for this experiment; 'walking', 'swinging whole body', 'sitting down', 'bowing', 'swinging upper-body', 'moving hand forward', 'moving hand backward', 'raising an arm', 'swinging hand', 'standing'. 10 video sequences were taken for ten people, and the actions for each person were taken in a single video. Among 10 subjects' videos, 5 were used for training. Each subject was asked to follow a scenario from the list below, for which a video was taken. The approximated duration time is also described.

1. Approaching a chair located in front of a stereo camera, from a distance: $3.2 \pm 1.8$ s.
2. Swinging body several times, by the chair: $6.6 \pm 1.2$ s.
3. Sitting down on the chair: $3.4 \pm 0.3$ s.
4. Bowing: $5.8 \pm 0.5$ s.
5. Swinging upper body: $5.1 \pm 0.8$ s.
6. Moving a right hand forward: $1.9 \pm 0.2$ s.
7. Moving a right hand backward: $1.8 \pm 0.4$ s.
8. Raising a right arm by side: $2.4 \pm 1.1$ s.
9. Swinging a right hand: $3.6 \pm 1.4$ s.
10. Standing up from the chair: $6.0 \pm 1.2$ s.

The experiments were performed using continuous video data, therefore, deletion, insertion, and substitution errors should be considered, when evaluating performance. The performance was evaluated from two measurements: correctness and accuracy.

$$\textit{Correctness}(\%) = \frac{(N - D - S)}{N} \times 100,$$
$$\textit{Accuracy}(\%) = \frac{(N - D - S - I)}{N} \times 100, \quad (21)$$

where $N$ is the total number of actions in a sequence and $D$, $S$, and $I$ represent the number of deletion, substitution, and insertion errors (Lee and Kim, 1999).

**Table 1**
Action recognition results using MHI (Bobick and Davis, 2001).

| Actions | Total | Errors | | | Correct (%) | Accuracy (%) |
|---|---|---|---|---|---|---|
| | | D | S | I | | |
| Walking | 5 | 0 | 1 | 0 | 80 | 80 |
| Swinging body | 5 | 0 | 2 | 1 | 60 | 40 |
| Sitting down | 5 | 0 | 1 | 0 | 80 | 80 |
| Bowing down | 5 | 0 | 1 | 1 | 80 | 60 |
| Swinging upper-body | 5 | 1 | 1 | 1 | 60 | 40 |
| Moving a hand forward | 5 | 1 | 1 | 0 | 60 | 60 |
| Moving a hand backward | 5 | 1 | 2 | 0 | 40 | 40 |
| Raising an arm | 5 | 0 | 0 | 0 | 100 | 100 |
| Swinging a hand | 5 | 0 | 1 | 0 | 80 | 80 |
| Standing up | 5 | 0 | 1 | 0 | 80 | 80 |
| Total | 50 | 3 | 11 | 3 | 72 | 66 |

Tables 1 and 2 represent the results of the MHI-based method, and the proposed VMT-based method, respectively. Overall, the proposed method outperform the MHI-based method. In particular, substitution errors were greatly reduced when using the VMT-based recognition method. For all actions including parallel

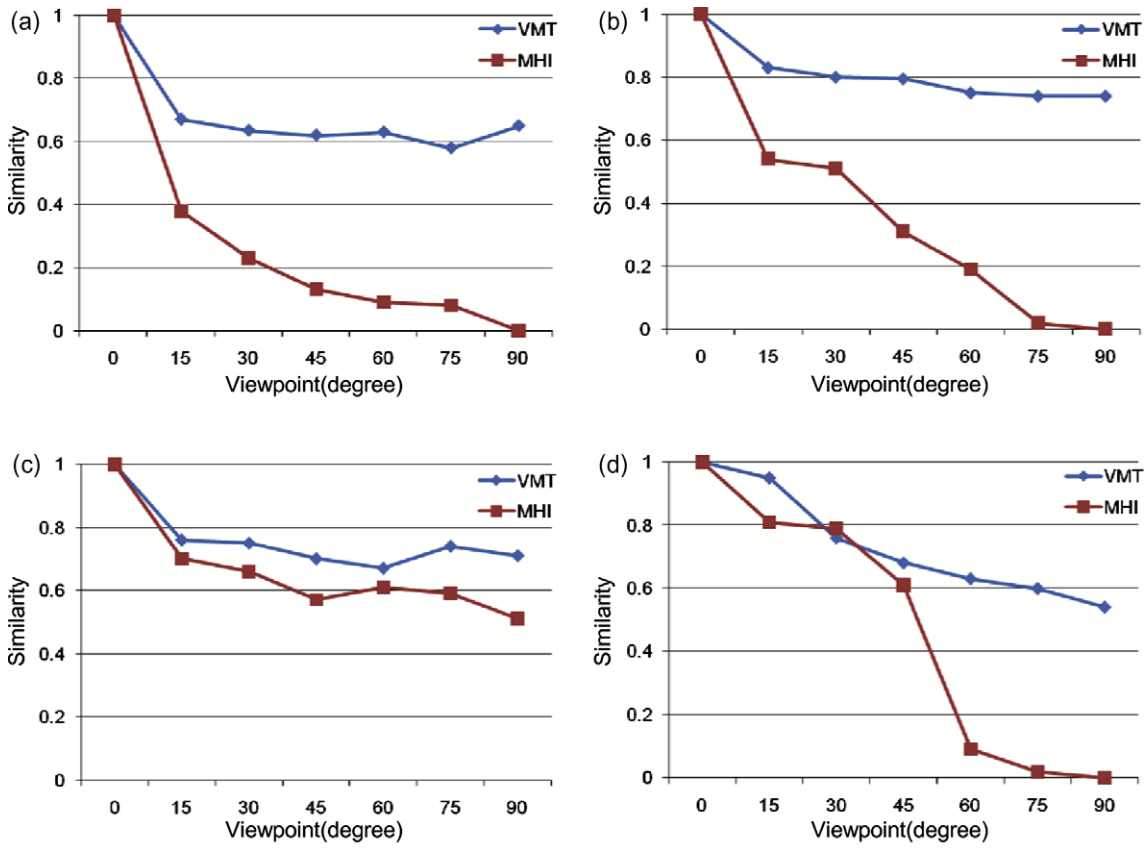


**Fig. 8.** Similarities between a template from the reference viewpoint (0°) and templates from the other viewpoints (15°–90°). The actions are '(a) moving a right hand forward', '(b) moving a right hand backward', '(c) bowing', and '(d) raising a right hand'.

**Table 2**
Action recognition results using the proposed VMT.

| Actions | Total | Errors | | | Correct (%) | Accuracy (%) |
|---|---|---|---|---|---|---|
| | | D | S | I | | |
| Walking | 5 | 0 | 0 | 0 | 100 | 100 |
| Swinging body | 5 | 0 | 0 | 0 | 100 | 100 |
| Sitting down | 5 | 0 | 1 | 0 | 80 | 80 |
| Bowing down | 5 | 0 | 0 | 1 | 100 | 80 |
| Swinging upper-body | 5 | 0 | 0 | 1 | 100 | 80 |
| Moving a hand forward | 5 | 0 | 1 | 0 | 80 | 80 |
| Moving a hand backward | 5 | 1 | 1 | 1 | 60 | 40 |
| Raising an arm | 5 | 0 | 0 | 0 | 100 | 100 |
| Swinging a hand | 5 | 1 | 0 | 0 | 80 | 80 |
| Standing up | 5 | 0 | 1 | 0 | 80 | 80 |
| Total | 50 | 2 | 4 | 3 | 88 | 82 |

movements to the camera, the VMT demonstrates much more stable results than the MHI. For the other actions, it also demonstrates better results, due to the temporal normalization in Section 3.1 and reconstruction method described in Section 4.

The variation of the action duration of each action is larger than that of the action in Section 5.1. Thus, for even non-parallel movements such as 'Swinging Body', there are differences between the results of the MHI and the proposed VMT methods. And the proposed VMT performed better than the MHI thanks to the proposed temporal normalization.

## 6. Conclusions and further research

The view-independent action recognition provides a natural environment for an advanced human–robot interface. However, the camera viewpoints problem is difficult to solve, since the presentation of an action changes dynamically, depending on camera viewpoints. In order to solve the problems, the VMT method is proposed. The proposed VMT method is an extension of the Motion History Image (MHI) method to 3D space and is generated by using motion history information in 3D space, using disparity maps of a stereo camera. We also proposed a Projected Motion Template (PMT) for viewpoint independent matching of two VMTs. The PMT is generated by projecting the VMT into a 2D plane that is orthogonal to an optimal virtual viewpoint. The optimal virtual viewpoint is a viewpoint from which an action can be described in greatest detail, in 2D space. Thus, the proposed method can recognize actions in a video taken from any viewpoint.

The proposed method was evaluated for various actions, in terms of viewpoint change robustness and consistency robustness, and compared with MHI. The experimental results demonstrated that the proposed method outperformed the MHI, and achieved consistent recognition results for various actions. Further research will include dealing with complex actions involving multiple motion orientations and achieving robust VMTs with respect to the noise of disparity maps.

## Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No. 2009-0060113).

## References

Ahmad, M., Lee, S.-W., 2008. Human action recognition using shape and CLG-motion flow from multi-view image sequences. Pattern Recognition 41 (7), 2237–2252.
Ben-Arie, J., Wang, Z., Pandit, P., Rajaram, S., 2002. Human activity recognition using multidimensional indexing. IEEE Trans. Pattern Anal. Machine Intell. 24 (8), 1091–1104.
Bobick, A.F., Davis, J.W., 2001. The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Machine Intell. 23 (7), 257–267.
Bodor, R., Jackson, B., Masoud, O., Papanikolopoulos, N., 2003. Image-based reconstruction for view-independent human motion recognition. In: Proc. IEEE Conf. Intelligent Robots and System, vol. 2. pp. 1548–1553.
Bradski, G.R., Davis, J.W., 2002. Motion segmentation and pose recognition with motion history gradients. Machine Vision Appl. 13 (3), 174–184.
Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R., 2007. Actions as space–time shapes. Trans. Pattern Anal. Machine Intell. 29 (12), 2247–2253. <www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html>.
Holte, M.B., Moeslund, T., 2008. View invariant gesture recognition using 3D motion primitives. In: Proc. IEEE Internat. Conf. Acoustics Speech and Signal Processing, pp. 797–800.
Hwang, B.-W., Lee, S.-W., 2003. Reconstruction of partially damaged faces based on a morphable face model. IEEE Trans. Pattern Anal. Machine Intell. 25 (3), 365–372.
Kirishima, T., Sato, K., Chihara, K., 2005. Real-time gesture recognition by learning and selective control of visual interest points. IEEE Trans. Pattern Anal. Machine Intell. 27 (3), 351–364.
Lee, H., Kim, J., 1999. An HMM-based threshold model approach for gesture recognition. IEEE Trans. Pattern Anal. Machine Intell. 21 (10), 961–973.
Mori, G., Ren, X., Efros, A., Malik, J., 2004. Recovering human body configurations: Combining segmentation and recognition. In: Proc. IEEE Internat. Conf. on Computer Vision and Pattern Recognition, pp. 326–333.
Niu, F., Mottaleb, M.A., 2004. View-invariant human activity recognition based on shape and motion features. In: Proc. IEEE Symposium Multimedia Software and Engineering, pp. 546–556.
Parameswaran, V., Cellappa, R., 2005. Human action–recognition using mutual invariants. Comput. Vision Image Understand. 98 (2), 295–325.
Ramanan, D., Forsyth, D.A., 2003. Finding and tracking people from the bottom up. In: Proc. IEEE Internat. Conf. Computer Vision Pattern Recognition, pp. 467–474.
Rao, C., Yilmaz, A., Shah, M., 2002. View-invariant representation and recognition of actions. Internat. J. Comput. Vision 50 (2), 203–226.
Seitz, S.M., Dyer, C., 1997. View-invariant analysis of cyclic motion. Internat. J. Comput. Vision 25 (3), 231–251.
Shan, C., Wei, Y., Qiu, X., Tan, T., 2004. Gesture recognition using temporal template based trajectories. In: Proc. IEEE Internat. Conf. on Pattern Recognition, pp. 954–957.
Shin, H.-K., Lee, S.-W., Lee, S.-W., 2005. Real-time gesture recognition using 3D motion history model. Adv. Intell. Comput. Lecture Notes Comput. Sci. 3644, 888–898.
Sminchisescu, C., Triggs, B., 2003. Estimating articulated human motion with covariance scaled sampling. Internat. J. Robot. Res. 22 (6), 371–391.
Weinland, D., Ronfard, R., Boyer, E., 2006. Free viewpoint action recognition using motion history volumes. Comput. Vision Image Understand. 104 (2), 249–257.
Yang, H.-D., Park, A.-Y., Lee, S.-W., 2007. Gesture spotting and recognition for human–robot interaction. IEEE Trans. Robot. 23 (2), 256–270.
Ye, G., Corso, J., Hager, G., 2004. Gesture recognition using 3D appearance and motion features. In: Proc. Internat. Conf. Computer Vision and Pattern Recognition, p. 160.
Yilmaz, A., Shah, M., 2005. Actions sketch: A novel action representation. In: Proc. Internat. Conf. Computer Vision and Pattern Recognition, pp. 984–989.