Model-based Human Action Recognition

Nattapon Noorit, Nikom Suvonvorn, and Montri Karnchanadecha
Department of Computer Engineering
Faculty of Engineering, Prince of Songkla University
Hat Yai, Thailand
pinggtar@hotmail.com, kom@coe.psu.ac.th, montri@coe.psu.ac.th

***Abstract*—The identification of human basic actions plays an important role for recognizing human activities in complex scene. In this paper we propose an approach for automatic human action recognition. The parametric model of human is extracted from image sequences using motion/texture based human detection and tracking. Action features from its model are carefully defined into the action interaction representation and used for the recognizing process. Performance of proposed method is tested experimentally using datasets under indoor environments.**

Keywords - human action recognition; activity recognition, human modeling; video surveillance

I. INTRODUCTION

Recognizing human activities from video sequence is one of the most challenging problems of surveillance application. In this paper, we propose the method for recognizing the basic actions of human activities, which is necessary for the event of interest detection, for example abnormal or rare events. Here, “action” is referred to simple motion patterns normally executed by a single person, such as walking, standing, laying, bending and sitting, and “activity” refers to the complex sequence of actions performed by several humans.

A large number of publications work for action recognition [1]. Efros et al.[2] attempt to recognize a set of simple actions (walking, running with direction and location) using a set of features that are based on blurred optic flow. Robertson and Reid [3] propose an approach where complex actions can be dynamically composed out of the set of simple actions by building a hierarchical system that is based on reasoning with belief networks and HMMs. Yilmaz and Shah [4] extract information such as speed, direction and shape by analyzing the differential geometric properties of space-time relation. [5][6] also analyze the space-time volume. Zelnik-Manor et al. [8] define dynamic actions as long-term temporal objects at multiple temporal scales features. Bradski et al.[7] develop MHI for motion segmentation that allow determination of the optical flow.

The rest of the paper is organized as follows. Section 2 describes our parametric human model. Section 3 depicts how to represent the actions for the recognition process. The experimentation and conclusion are discussed in the section 4 and 5 respectively.

II. HUMAIN MODEL EXTRACTION

In this section, we describe how to determine the parametric model of human from the images sequence. The overall process can be divided into two major steps: motion detection and tracking and human model construction with parameter estimation.

A. *Motion-texture based Human Detection and Tracking*

Moving object is separated from background by using background subtraction technique which small noises are removed by morphological opening and closing filters. Generally, an object might be detected in several fragmented image regions. In that case, a region-fusion operation is needed. Two regions are considered to be the same object if they are overlapped or their distance less than a specific threshold. With these constraints, the method is again very sensible to light condition, such as shadow, contrast changing and sudden changes of brightness. Intuitively, introducing some special characteristics of object, for instance texture properties, will probably improve the better results. Therefore, in the fusion process the color probability density of object’s texture is additionally applied for computing the similarity between regions using Mean-shift algorithm [9]. This mixture of motion and texture of object for detection and tracking can reduce significantly noises and increases consequently the effectiveness of our tracking algorithm. However, there are always additive noises superposed with detected objects that will be eliminated later by human model constraints.

The extracted foreground that supposed to be a human is then segmented into three regions representing the three important parts of human structure, for instance, head, body, and legs. Noticed that both of hands are denied and only one component is used for representing the both legs. We assume that with only three components of human model the five basic actions could be indentified correctly. Standing action is used for initiating the human model construction which the relations between three parts are specified explicitly; body region is bigger than others and located at the middle, the upper and lower connected regions are assumed to be the head and legs parts respectively. In this step, we have to deal with errors emitted from motion detection and tracking process. The human model constraints are used for noise suppression.

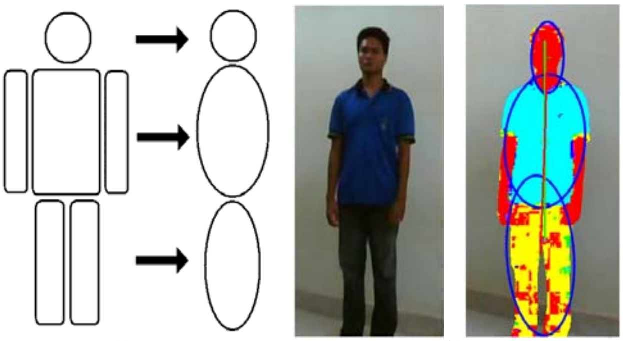


Figure 1. (a) Simplification of human structure (b) Example of reconstructed human model from image secquences.

B. *Parametric Model Definition*

Our human model for action recognition is presented in the figure 2. The parameters of model are considered into two groups: internal and external parameters. The internal parameters $[\vec{v}_h, \vec{v}_l]$ represent the characteristic of human structure, which describes the distances $[\vec{\mathrm{v}}_\mathrm{h}^{s}, \vec{\mathrm{v}}_\mathrm{l}^{s}]$ and directions $[\vec{\mathrm{v}}_\mathrm{h}^{\theta}, \vec{\mathrm{v}}_\mathrm{l}^{\theta}]$ from body component to the head and legs components respectively. The external parameters $[\vec{v}_{m1}, \vec{v}_{m2}, \vec{v}_{m3}]$ correspond to the properties of human movement from frame to another. More precisely, it characterizes the velocity $[\vec{\mathrm{v}}_{\mathrm{m}1}^{s}, \vec{\mathrm{v}}_{\mathrm{m}2}^{s}, \vec{\mathrm{v}}_{\mathrm{m}3}^{s}]$ and direction $[\vec{\mathrm{v}}_{\mathrm{m}1}^{\alpha}, \vec{\mathrm{v}}_{\mathrm{m}2}^{\alpha}, \vec{\mathrm{v}}_{\mathrm{m}3}^{\alpha}]$ of the head, body, and legs components respectively.

Note that we define our parameters relatively up to the centroid of human components in order to reduce noises produced during the low-level process, for example motion detection, segmentation, tracking and etc.

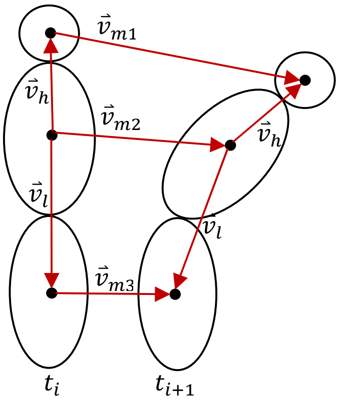


Figure 2. Humain model with parameters.

III. ACTION REPRESENTATION

In this section, we explain how to define the basic actions. Experimentally, we found that almost activities can be decomposed into the five basic actions: standing, bending, sitting, walking and laying. So, an activity can then be described by a sequence of these actions with some transition parameters mostly depending on motion vectors. Figure 3 shows the actions represented by our simplified human model.

To recognize the actions, we establish the features of each action from the parameters of human model by the following $[\frac{\vec{v}_h^{\theta}}{\vec{v}_\mathrm{l}^{\theta}}, \frac{\vec{v}_h^{s}}{\vec{v}_\mathrm{l}^{s}}, \frac{\partial \vec{v}_{m1}^{s}}{\partial t}, \frac{\partial \vec{v}_{m2}^{s}}{\partial t}, \frac{\partial \vec{v}_{m3}^{s}}{\partial t}, \frac{\partial^2 \vec{v}_{m1}^{\alpha}}{\partial t^2}, \frac{\partial^2 \vec{v}_{m2}^{\alpha}}{\partial t^2}, \frac{\partial^2 \vec{v}_{m3}^{\alpha}}{\partial t^2}]$:

- $\frac{\vec{v}_h^{\theta}}{\vec{v}_\mathrm{l}^{\theta}}$ the angle ratio between head and legs with respect to the body.
- $\frac{\vec{v}_h^{s}}{\vec{v}_\mathrm{l}^{s}}$ the distance ratio between head and legs with respect to the body.
- $\frac{\partial \vec{v}_{m1}^{s}}{\partial t}, \frac{\partial \vec{v}_{m2}^{s}}{\partial t}, \frac{\partial \vec{v}_{m3}^{s}}{\partial t}$ the velocity of each human components.
- $\frac{\partial^2 \vec{v}_{m1}^{\alpha}}{\partial t^2}, \frac{\partial^2 \vec{v}_{m2}^{\alpha}}{\partial t^2}, \frac{\partial^2 \vec{v}_{m3}^{\alpha}}{\partial t^2}$ the angular acceleration of each human components.

Our assumption, by combining the above features we could identify the different actions correctly. The figure 3 simply shows the complete basic actions with its features. We classify actions into two types: static and dynamic actions.

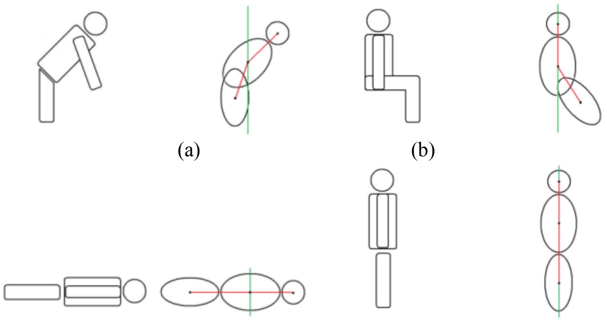


Figure 3. Basic actions (a) Bending (b) Sitting (c) Laying (d) Standing or Walking.

A. *Static actions*

The actions are considered static if only if there are at least one component which the velocity is null. By definition, the static actions are comprised of standing, bending and sitting. Consequently, Its features are then combined by only from the internal parameters $[\vec{v}_h, \vec{v}_l]$, that can be used later for action identification or action discrimination.

1) *Standing features*

In general, the standing action will provide $\vec{v}_h$ and $\vec{v}_l$ that tend to be paralleled with vertical axis, independently from the camera view of point. And certainly, the velocities of very components are near to zero. During the transition state, for example from standing to sitting or to bending, the angles of $\vec{v}_h$ and $\vec{v}_l$ increase while inversely decreasing for the opposite transition.

$$\vec{v}_h^\theta \cong 0, \vec{v}_l^\theta \cong 180, \frac{\partial \vec{v}_{m1}^s}{\partial t} \cong \frac{\partial \vec{v}_{m2}^s}{\partial t} \cong \frac{\partial \vec{v}_{m3}^s}{\partial t} \cong 0 \quad (1)$$

$$\frac{\partial \vec{v}_l^\theta}{\partial t} > 0 \quad (2), \qquad \frac{\partial \vec{v}_h^\theta}{\partial t} >0 \quad (3)$$

2) *Bending features*

For the bending case, the head component is blending down while the body and legs components are still fixed. Considering the parameters model, the vector $\vec{v}_l$ is likely to be positioned vertically. During the transition state from bending to standing the angle of $\vec{v}_h$ will decrease.

$$\vec{v}_l^\theta \cong 180, \frac{\partial \vec{v}_{m2}^s}{\partial t} \cong \frac{\partial \vec{v}_{m3}^s}{\partial t} \cong 0 \quad (4)$$

$$\frac{\partial \vec{v}_h^\theta}{\partial t} < 0 \quad (5)$$

3) *Sitting features*

In the sitting situation, the characteristic of features is opposite to the bending action. The vector $\vec{v}_h$ tends to be paralleled with y-axis and the legs component does not move. During the transition state from sitting to standing the angle of $\vec{v}_l$ will decrease.

$$\vec{v}_h^\theta \cong 0, \frac{\partial \vec{v}_{m3}^s}{\partial t} \cong 0 \quad (6)$$

$$\frac{\partial \vec{v}_l^\theta}{\partial t} < 0 \quad (7)$$

B. *Dynamic actions*

The actions are considered dynamic if only if all components of human model move. We found that only the internal parameters of human model cannot discriminate or identify the action itself. It needs to consider additionally the external parameters as transition factors.

1) *Walking features*

In case of walking action, every part of human move generally and approximately in the same direction and speed. Therefore, the walking action can then be identified by the standing action features with motion transitions of all components. So, their velocities are superior to zero.

$$\frac{\partial \vec{v}_{m1}^s}{\partial t} \cong \frac{\partial \vec{v}_{m2}^s}{\partial t} \cong \frac{\partial \vec{v}_{m3}^s}{\partial t} > 0 \quad (8)$$

2) *Laying features*

For the laying action, every parts of human body move with different directions and speeds. In general, head component is faster than body and legs respectively. We found that with some specific camera view point the $\vec{v}_h$ and $\vec{v}_l$ are lay on the same line, which is recognizing as standing action. Thus, we cannot use only internal parameters for identifying laying action. We define then the laying action with the standing action and motion transition features. Generally, the acceleration of head is more than body and legs components. And certainly the acceleration of body components is bigger than the legs one.

$$\frac{\partial^2 \vec{v}_{m1}^\theta}{\partial t^2} > \frac{\partial^2 \vec{v}_{m2}^\theta}{\partial t^2} \geq \frac{\partial^2 \vec{v}_{m3}^\theta}{\partial t^2} >0 \quad (9)$$

The global relation between actions can be represented as state diagram, shown in figure 4.

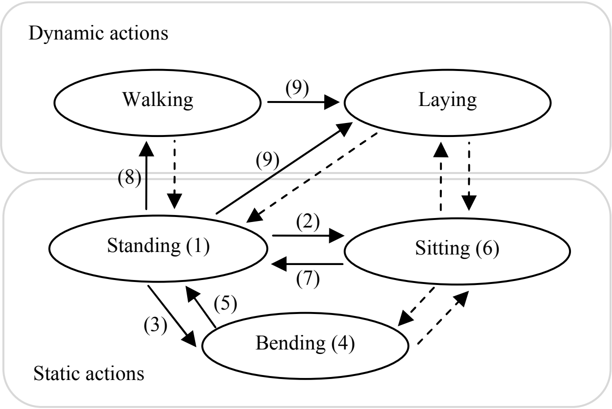


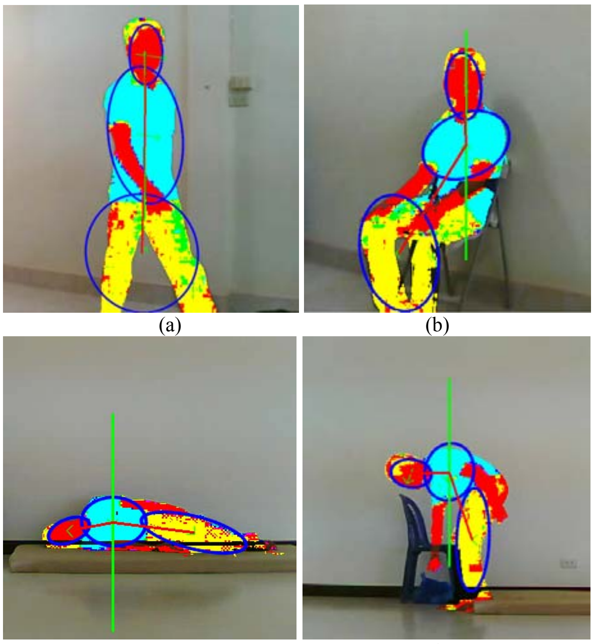
Figure 4. Action representation: dark arraws with number refers to features defining above, dash arrows are the N/A action transition.

IV. EXPERIMENTATION AND DISCUSSION

The experimentation is performed using the datasets established under indoor environments. Five peoples with different clothes act randomly and continuously with respect to the testing actions, consisting of 77 standing, 28 sitting, 50 bending, 44 walking, and 11 laying actions.

Figure 5 shows the results of the parametric human model obtained by the extraction process: motion-texture human detection and tracking, segmentation, and modeling. The image regions in red, blue, and yellow colors represent the head (hands), body, and legs of human structure. We can notice that the red color describes not only the head but also hands; due to both of these regions have the same skin color model. However, using the constraints on human model the simplification of human body is done quite well in the most cases. The blue ellipses on the figure depict the principal components of model and the red line linked between its centroids representing the internal parameters $[\vec{v}_h, \vec{v}_l]$. Consequently, we can determine the external parameters $[\vec{v}_{m1}, \vec{v}_{m2}, \vec{v}_{m3}]$ by computing the movement of each component using results obtaining from the previous frames during the tracking process. The green arrows at centroid of each component show the velocity and direction of movement. In this way, the features for identifying the corresponding actions can be estimated.

Figure 6 shows the features of actions extracted from human model during a testing scenario. A man acts the actions: bending, sitting, laying, standing, and walking respectively. Using the static and dynamic action features defined in section 3, we can recognize correctly the corresponding actions, shown in the rectangle box. However, we can also observe the errors obtained in a short period of time during the transition state from laying to standing action, which our method recognizes as N/A actions. This is quite normal because these transitions are not modeled explicitly in our action representation, figure 4.



(a) (b) (c) (d)

Figure 5. Examples of modeled action : (a) Standing (b) Sitting (c) Bending (d) Laying

The table 1 shows the experimentation results of our method. Number of frames representing each action is compared with frames in total of the identified action. From the highest to lowest recognition rate the actions are ordered by following: laying, standing, bending, sitting and walking respectively. The lowest rate can correctly identify actions up to 80%. In the worst case for walking action, we found that errors are produced when both of legs are separated, which leads to incorrect computation of centroid of legs component. This may be corrected by improving the human detection and tracking method. In the best case for laying action, our proposed method can completely identify all testing datasets. In average, up to 93% of basic actions are recognized appropriately. Note that N/A actions mean that the frames are not recognized by any modeled actions, but it can be identified as the non-definition transition actions showing as dash-lines in the figure 4.

However, using our method the errors of recognition can happen probably in the case where the camera is perfectly perpendicular to the actions. This make the internal parameters of human model became insignificant for the action discrimination. In the future work, to increase the recognition rate, it is necessary to correct this special case. More complex parameters model may be established or more cameras are used for recognition process, such as multi-camera based action recognition. Accordingly, this may also lead to the solution of the N/A action recognition problem too.

TABLE I. EXPERIMENTATION RESULTS

Actions	Number of Frames	Detected actions	Recognition rate
Standing	2886	2869	99.41 %
Sitting	1220	1089	89.26 %
Bending	2250	2123	94.35 %
Walking	2243	1809	80.65 %
Laying	1334	1334	100 %
N/A *	999	999	100 %
Total	10932	9224	93.95%

*N/A represents action transitions that are not modeled by our action representation (dash-lines in figure 4)

V. CONCLUSION

In this paper, we proposed an approach for identifying the human basic actions required for human activities recognition, such as laying, standing, bending, sitting and walking. The parametric model of human is presented for defining the action features. The action representation based on their features is then proposed as the identification process. The evaluation of our method is done over a large datasets with encourage results; up to 93% of actions are correctly recognized.

ACKNOWLEDGMENT

Granted by ENG/PSU
Partially funded by NECTEC: Project id P-00-10638

REFERENCES

[1] Krüger, Volker; Kragic, Danica; Ude, Aleš; Geib, and Christopher, The meaning of action: a review on action recognition and mapping, Advanced Robotics, Volume 21, Number 13, 2007 , pp. 1473-1501(29)

[2] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing Action at a Distance. In Internatinal Conference on Computer Vision, volume II, pages 726–733, Nice, France, Oct 13-16, 2003

[3] N. Robertson and I. Reid. Behaviour Understanding in Video: A Combined Method. In Internatinal Conference on Computer Vision, pages 808–815, Beijing, China, Oct 15-21, 2005

[4] A. Yilmaz and M. Shah. Actions Sketch: A Novel Action Representation. In Computer Vision and Pattern Recognition, volume I, pages 984–989, San Diego, California, USA, June 20-25, 2005

[5] B. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. In Internatinal Conference on Computer Vision, pages 1395–1402, Beijing, China, Oct 15-21, 2005

[6] M. Bregonzio, S. Gong and T. Xiang. Recognising Action as Clouds of Space-Time Interest Points. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, June 2009

[7] Lihi Zelnik-Manor and Michal Irani, Statistical Analysis of Dynamic Actions, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 28 , Issue 9 (September 2006), Pages: 1530 – 1535

[8] G. Bradski and J. Davis. Motion Segmentation and Pose Recognition with Motion History Gradients. Machine Vision and Applications, 13(3):174–184, 2002

[9] Y. Cheng.Mean shift, mode seeking, and clustering,IEEE Trans. on Pattern Analysis and Machine Intelligence, 17(8):790-799, 1998.

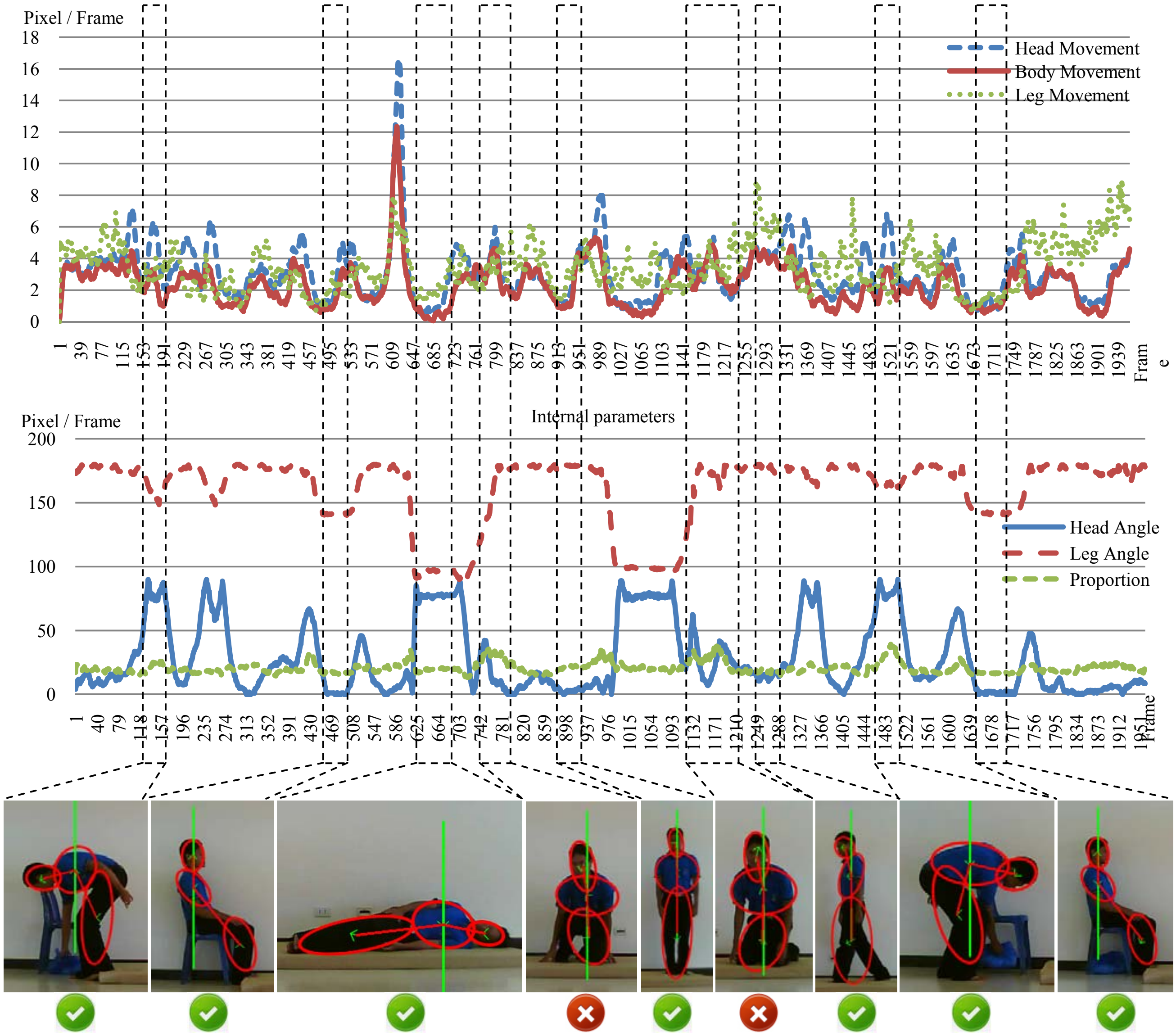


Figure 6. Example of action features with detected results.