# DEPTH-BASED POSTURE RECOGNITION BY RADAR AND VISION FUSION FOR REAL-TIME APPLICATIONS

*I-Cheng Tsai and Ching-Te Chiu*

Department of Computer Science, National Tsing-Hua University, Hsinchu, Taiwan
s100062525@oz.nthu.edu.tw, ctchiu@cs.nthu.edu.tw

## ABSTRACT

A radar sensor can capture the distance and angle of an object. Mapping the radar distance and angle information to the coordinates of a video frame accelerates the speed of object identification. The distance information is used to calibrate the size of an object to help the recognition. To achieve real-time performance, we use only five center of gravity points (COG) and four feature sets. Two feature sets measure the displacement of the upper and lower body COG in the vertical and horizontal directions. The other two feature sets quantize the upper and lower body angular change rate. The simulation results show that our proposed approach achieve 98.02% to 80.20% recognition rates for various postures and actions in the KTH and ISIR databases.

***Index Terms***— posture recognition, action analysis, radar and vision fusion, center of gravity, video surveillance.

## 1. INTRODUCTION

The human posture and action analysis is an important research area in various applications such as pedestrian identification in vehicles, video surveillance, medical diagnostics, and human machine interaction systems [2]. Radar sensors become more and more popular in vehicles and medical detections for human breathing and heartbeat [1]. Human postures may affects the frequency of breathing and heartbeat. Currently there are few literatures on combining radar and vision sensors for human posture recognitions for vehicle or medical applications.

Several posture recognition approaches have been proposed. Oliver *et al*. [3] use hidden Markov models (HMMs) and a trajectory feature to develop a visual surveillance system to model and recognize human actions. Wren *et al*. propose a Pfinder system for tracking and recognizing human behavior based on a 2-D blob model [4]. Juang *et al*. [10] apply consecutive frame difference to extract silhouette, then use significant points in body silhouette to analyze human action. Hsieh *et al*. [5] propose deformable triangulation and centroid context for human action analyzation. Although these types of schemes are useful to analyze human actions, most of them have high computation complexity.

In this paper, we propose a real-time feature extraction algorithm that analyzes human actions by the radar and vision sensors. We adopt the distance and angle information from the radar sensor to accelerate the object identification in the vision sensor. The proposed approach adopts a centroid context based posture and action recognition using only five center of gravity points (COG) and four feature sets. Then we compute the COG of the human body and set this COG as the origin. With the vertical and horizontal lines through the origin, the body is divided into four parts and the COGs in each part are computed. Two feature sets measure the displacement of the upper and lower body COG in the vertical and horizontal directions. The other two feature sets quantize the upper and lower body angular change rates that can be used to identify the human actions.

The remainder of the paper is organized as follows. In Section II, we present the proposed depth-based posture recognition system by radar and vision fusion. The model-based feature sets and the posture classification scheme are shown in Section III. In Section IV, the experiment results and computational complexity analysis are presented. Finally, a conclusion is drawn in Section V.

## 2. A DEPTH-BASED POSTURE RECOGNITION SYSTEM WITH RADAR AND VISION FUSION

Our proposed depth-based posture recognition system with radar and vision fusion is shown in Fig. 1. The radar and vision sensors are setup to have the same distance to an object. The radar sensor detects objects with their distance and angle information. The camera sensor translates the distance and angle obtained from the radar to coordinates in the video frames that are used as references to detect objects. After the object is detected and extracted, we use the distance information to adjust the parameters in the recognition system. Then the model-based feature extraction and posture classification are adopted to identify the postures.

The radar sensor used in our system is LMS100-10000 with a maximum detection distance of 20 meter and maximum scanning angle of 270 degree. The radar measured distance $r$ and angle $\theta$ are corresponding to the coordinates ($x_r$,$y_r$)=($rcos\theta$, $rsin\theta$) in the radar domain. During calibra-

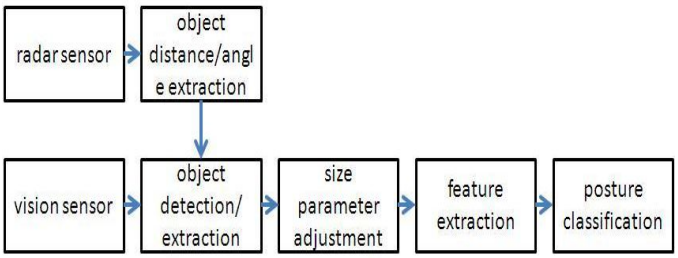



**Fig. 1**. Overall flow of the radar and vision based posture recognition system.

tion, we measure multiple radar vertical coordinates $y_r$ and multiple corresponding video frame vertical coordinates $y_v$. The relation between $y_r$ and $y_v$ can be expressed as below and Fig. 2(a).

$$y_v = a \times \ln(y_r) + b, \tag{1}$$

where $a$ and $b$ are constants. By using the similarity between an optical triangle and an image triangle, the relation between $x_r$ and $x_v$ can be expressed as below.

$$x_v = (1 - \frac{y_r}{c}) \times d \times x_r, \tag{2}$$

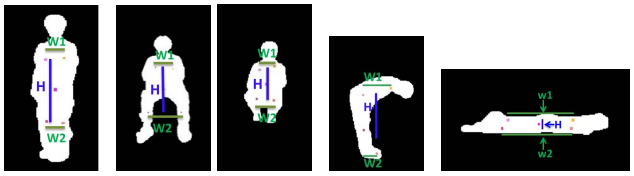
where $c$ and $d$ are constants.

After identifying the ($x_v$,$y_v$) in the video frame, we use the mapped coordinate as the center of a limit search region. Then we apply color saliency in this limited search region to find the silhouette of the object. Since the size of an object in a video frame depends on the distance of the vision sensor, we use the distance obtained from radar sensors to adjust the parameters for recognition. An example of two objects with measured radar distances and angles and the corresponding video frame coordinates is shown in Fig. 2.

(a) (b)

**Fig. 2**. (a) Radar and the $y$ ($y_v$(vertical index), $y_r$(horizontal index)) mapping curve, (b) an image example of two objects with radar measured distance $(4.5m, 6m)$ and angle $\theta(79, 98.75)$ in degree.

In order to reduce the computational complexity, we propose a centroid context based posture and action recognition approach using only five center of gravity points (COG) and four feature sets. Two feature sets measure the displacement of the upper and lower body COG in the vertical and horizontal directions. For example, Fig. 3 shows the variation of vertical height at five different postures. When a human being is in the stand posture, the vertical height value is the highest and the values of vertical height decrease in the other postures. The other two feature sets quantize the upper and lower




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

**Fig. 3**. Variation of height (H) at four different postures: (a) stand, (b) sit, (c) squat, (d) Bending, and (e) laying.

body angular change rate. With the feature sets and a classification model, our proposed approach is able to recognize five different static postures including stand, laying, bend, sit and squat and two actions, walking and handwaving.

## 3. MODEL-BASED FEATURE EXTRACTION AND CLASSIFICATION

According to the concept that human body has different features in different postures, we propose a model-based feature extraction method. In this work, we use the trunk ratio and the upper and lower body angles to distinguish postures. The proposed method is simple and fast that can be applied in real-time applications.

The center of gravity of the human body is used as a feature point in many posture recognition methods. The following equation shows the computation of COG of a human body region from the extracted silhouette.

$$(X_{cog}, Y_{cog}) = (\frac{1}{N}\sum_{x_i=1}^{N} x_i, \frac{1}{N}\sum_{y_i=1}^{N} y_i), \tag{3}$$

$$(x_i, y_i) \in \textit{human body region}$$

where $(X_{cog}, Y_{cog})$ is the $x$ and $y$ coordinates of the computed COG. When the body region is the extracted human body silhouette, we call this as the origin COG and is denoted as $COG_o$. From this origin COG, we divide the body into four parts using the vertical and horizontal lines passing through the $COG_o$.

Then we compute the COGs in each part and they are denoted as $COG_1$, $COG_2$, $COG_3$ and $COG_4$.

### 3.1. Definition of the Features

In this subsection, we extract three feature ratios $R_1$, $R_2$ and $R_3$ respectively. At first, we use $COG_o$, $COG_1$, and $COG_2$

to construct an upper triangle then use $COG_o$, $COG_3$, and $COG_4$ to construct a lower triangle. From these two triangles, we define three values $W_1$, $W_2$, and $H$ as follows.

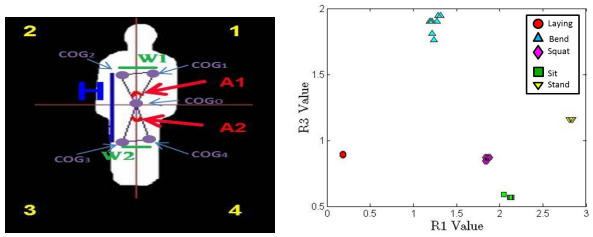
$$W_1 = |X_{cog1} - X_{cog2}| \quad (4)$$

$$W_2 = |X_{cog3} - X_{cog4}|, \quad (5)$$

where

$$H = \sqrt{(\frac{X_{cog12} - X_{cog34}}{2})^2 + \frac{Y_{cog12} - Y_{cog34}}{2})^2}. \quad (6)$$

The $W_1$ is the horizontal distance of the $COG_1$ and $COG_2$, and $W_2$ is the horizontal distance of the $COG_3$ and $COG_4$. The $H$ is the Euclidean distance between the midpoint of $COG_1$ and $COG_2$ and the midpoint of $COG_3$ and $COG_4$. According to these values, we define the three feature ratios $R_1 = \frac{H}{W_1}$, and $R_3 = \frac{W_1}{W_2}$. $R_1$ is used to distinguish postures including the stand, sit, squat and laying. $R_3$ is used to identify the bend posture. Fig. 4(a) shows an example of the body feature values definition. Fig. 4(b) shows the postures distribution of different ratio value.




(a) (b)

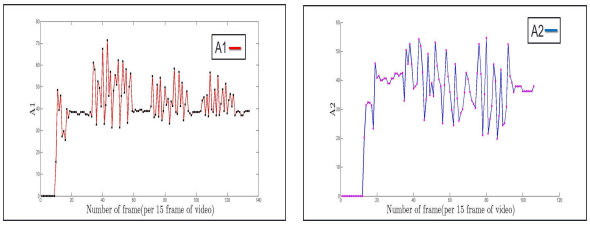
**Fig. 4**. (a) Human body feature value definition, (b) Postures distribution of different ratio value.

tures distribution of different $R_1$ and $R_3$ values and it shows that the stand, sit, squat, laying and bending can be differentiated from these two values. Next, we present angle features to distinguish action postures of handwaving and walking. The upper $COG_1$ and $COG_2$ change their positions for a human handwaving action. For a walking action, the lower $COG_3$ and $COG_4$ change their positions.

Here we define two angle features $A_1$ and $A_2$ based on the law of cosines as below. We can find $A_1$ and $A_2$ as shown in Fig. 4(a). $A_1$ is the bottom angle of the triangle in the upper body. The $A_1$ value changes up and down repeatedly during handwaving. $A_2$ is the top angle of the triangle in the lower body and the $A_2$ changes repeatedly for the walking. Fig. 5(a)(b) shows the angle change rate versus frame for handwaving and walking.

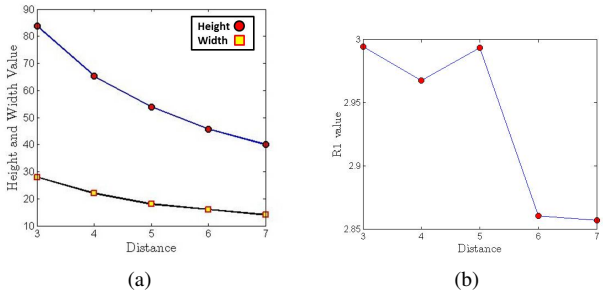
### 3.2. Threshold Definition and Computation

In this paper, we use a static camera to observe human postures. Human have different ground truth size in different depth distance in the static scenes. Fig. 6 shows the $H$, $W_1$, and $R_1$ values at different distances. The decreasing of the $H$ values is larger than the $W_1$ values when the distance increases. The $R_1$ has higher values when the distance is smaller than five meter then the value becomes much smaller after this distance. In order to reduce the scaling problem for these features, we restrict the initial posture in a test video is the stand posture and use it as a reference. We also use the distance obtained from radar to adjust the classification parameters. When the stand posture have the highest $R_1$ value and the $R_1$ value changes depending on postures. At first, we perform a recognition step for a static stand posture. If this stand posture persists static with $T_R$ seconds that $T_R$ is 3 here, then we compute the current $R_1$ value for the stand position. Based on the computed $R_1$ value, we can define stand threshold $T_S = R_1$. A human body has very large $W_1$ and $W_2$ values and a small $H$ value at the laying posture as shown in Fig. 3(e). In this paper, we assign the laying posture threshold $T_L$ is 0.5. Then we define an equation to calculate a threshold $T_{SS} = (T_S - T_L) * \frac{4}{5} + T_L$ to distinguish the sit and squat postures. In order to distinguish the bend posture, we use the $R_3$ feature. The reason of using the $R_3$ feature is beacuse the upper $W_1$ is larger than $W_2$ in the bend posture. Based on this feature, we assign the bend threshold $T_B$ is $\frac{3}{2}$. We adopt these features and thresholds to model the recognition flow. The recognition flow is shown in Fig .7.




(a) (b)

**Fig. 5**. Angle change rate versus frame for: (a) handwaving, and (b) walking.




(a) (b)

**Fig. 6**. Height, width and $R_1$ comparison in different distances: (a) Height ($H$) and width ($W_1$) variation in different distances, and (b) $R_1$ variation in different distances.

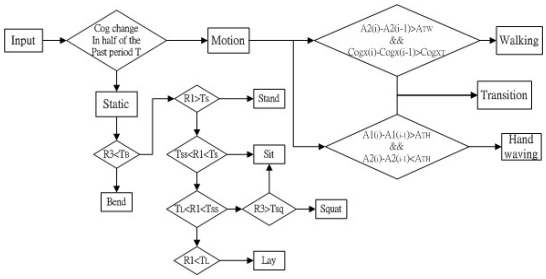



**Fig. 7**. Recognition flow of the proposed system.

## 4. EXPERIMENTAL RESULTS AND COMPUTATIONAL COMPLEXITY ANALYSIS

We evaluate our approach on the KTH database [7] and ISIR database [11]. We also adopt Hsieh *et al* [5]'s approach for comparison.

The KTH human action dataset, originally created by [7], consists of 600 videos (160 $\times$ 120) with 25 persons performing human action in four different scenarios: outdoors s1, outdoors with scale variation s2, outdoors with different clothes s3, and indoors s4. Due to the background limitation, we perform the experiments for the walking sequence here. The result is shown in Fig. 8. It shows that the proposed method can

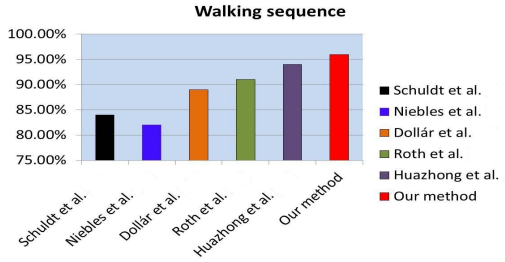



**Fig. 8**. Experimental results from the KTH video sequence with walking posture.

recognize most walking sequences and a recognition rate of 96% is obtained. A sequence database ISIR comprising eight actions is considered: crouch down, stand up, sit down, sit up, walk, bend down, get up from bending, and jump. Various viewpoints are acquired for each action. The face, 45°and 90°views are captured while others are synthesized from the recorded sequences already by symmetry. The number of the test images for the squat postures is 165 and the correct recognized number is 162. The recognition rate is 98.2%. The number of the test images of the sit posture is 112 and the correct recognized number is 102. The recognition rate is 91.07%. The number of the test images of the bend posture is 126 and the correct recognized number is 101. The recognition rate is 80.20%. The recognition rate results are shown in Fig. 9.

The experimental results show that our proposed method


**Fig. 9**. Experimental results from the ISIR database with squat, sit, and bend postures.

has good recognition rate at squat and sit postures. For the bend postures, we have lower recognition rate because the direction of bend is face to the camera.

The computation complexity of the proposed approach, Chen's [5], and Juang's [10] methods are listed in Table I. The parameter $n$ means the number of pixels in an image; the parameter $k$ denotes the number of contour points in a silhouette and the parameter $t$ means the number of triangle in a silhouette. From Table I, we show that the required arithmetic operations of our proposed approach are lower than the other two compared methods. Especially the number of multiplication and division operation required in our method is independent of the image size and a constant. The numbers in parentheses in the Table I are the gate counts when the value of $n$, $k$, and $t$ are 131648, 895, and 58 respectively. It shows that our approach use the least amount of add/subtract operations and the multiplication/division operation is significantly lower than others.

**Table 1**. Computational Complexity Analysis

| Method | operation numbers of (+, -) | operation numbers of (*, /) |
|---|---|---|
| Our proposed | 5n + 13 (658253) | 29 (29) |
| Hsieh *et al.* [5] | 5n + 10k + 12t (667886) | 2n + 4k + 12t (267572) |
| Juang *et al.*[10] | 12n + 26k (1603046) | 4n + 3k + 2 (529279) |

$n$: the number of pixels in a image.
$k$: the number of contour in a silhouette.
$t$: the number of triangle.

## 5. CONCLUSION

We have proposed a model-based feature recognition scheme to analyze human postures. The proposed method using five COG points to compute the ratio variation of vertical and horizontal direction and the upper and lower body angular change rate. With the feature sets, we present a classification flow for static and action postures. The proposed method only uses five feature sets to achieve high recognition rates. This method not only has low computation complexity but also has low implementation cost so it is suitable to be applied in real-time handheld devices.

## 6. REFERENCES


[1] S.I. Ivashov, V.V. Razevig, A.P. Sheyko, and I.A. Vasilyev, "Detection of Human Breathing and Heartbeat by Remote Radar," *Progress in Electromagnetic Research Symposium*, Pisa, Italy, 2004.

[2] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans,* vol. 30, no. 3, pp. 286-297, May 2000.

[3] N. Oliver, B. Rosario, and A. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 831-843, Aug. 2000.

[4] C. R. Wren *et al.*, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780-785, Jul. 1997.

[5] C. C. Chen, J. W. Hsieh, Y. T. Hsu, and H. Y. Mark Liao, "Video-Based Human Movement Analysis and Its Application to Surveillance Systems," *IEEE Trans. on Multimedia*, vol. 10, no. 3, pp. 372-384, 2008.

[6] I. Haritaoglu, D. Harwood, and L. S. Davis, "$W^4$-Who? When? Where? What? A Real-Time System for Detecting and Tracking People," *International Conference on Face and Gesture Recognition*, April, 14-16, 1998.

[7] C. Schuldt, I. Laptev and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," in *Proc. ICPR'04*, Cambridge, UK.

[8] J. Barron, D. Fleet, and S. Beauchemin, "Performance of Optical Flow Techniques," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 42-77, 1994.

[9] C. Anderson, P. Burt, and G. van der Wall, "Change Detection and Tracking Using Pyramid Transformation Techniques," in *Proceedings of SPIE-Intelligent Robots and Computer Vision*, vol. 579, pp. 72-78, 1985.

[10] C. F. Juang, C. M. Chang, J. R. Wu, and D. Lee, "Computer Vision-Based Human Body Segmentation and Posture Estimation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans,* vol. 39, no. 1, pp. 119-133, 2009.

[11] A. Mokhber, C. Achard, and M. Milgram, "Recognition of human behavior by space-time silhouette characterization," *Pattern Recognition Letters, 2008.*

[12] Peter M. Roth, Thomas Mauthner, Inayatullah Khan, and Horst Bischof, "Efficient Human Action Recognition by Cascaded Linear Classification," 2009 *IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops.*

[13] Huazhong Ning, Tony X. Han, Dirk B. Walther, M. Liu and Thomas S. Huang, "Hierarchical Space-Time Model Enabling Efficient Search for Human Actions," *IEEE Trans. on Circuit and Syst. for video tech.*, vol. 19, no. 6, pp. 808-820, June 2009.

[14] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, Behavior recognition via sparse spatio-temporal features, in *VS-PETS*, 2005.

[15] J. Niebles and L. Fei-Fei. "Unsupervised learning of human action categories using spatial-temporal words," In *BMVC* 2006.

[16] C. Achard, X. Qu, A. Mokhber, and M. Milgram, "A novel approach for recognition of human actions with semi-global features," *Machihne Vision and Applications*, vol. 19, no. 1, pp. 27-34, 2008.