Research Article
Multiview Layer Fusion Model for Action Recognition Using RGBD Images

Pongsagorn Chalearnnetkul and Nikom Suvonvorn

Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand

Correspondence should be addressed to Pongsagorn Chalearnnetkul; pongsagorn.ch@gmail.com

Received 3 January 2018; Revised 27 April 2018; Accepted 20 May 2018; Published 20 June 2018

Copyright © 2018 Pongsagorn Chalearnnetkul and Nikom Suvonvorn. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Vision-based action recognition encounters different challenges in practice, including recognition of the subject from any viewpoint, processing of data in real time, and offering privacy in a real-world setting. Even recognizing profile-based human actions, a subset of vision-based action recognition, is a considerable challenge in computer vision which forms the basis for an understanding of complex actions, activities, and behaviors, especially in healthcare applications and video surveillance systems. Accordingly, we introduce a novel method to construct a layer feature model for a profile-based solution that allows the fusion of features for multiview depth images. This model enables recognition from several viewpoints with low complexity at a real-time running speed of 63 fps for four profile-based actions: standing/walking, sitting, stooping, and lying. The experiment using the Northwestern-UCLA 3D dataset resulted in an average precision of 86.40%. With the i3DPost dataset, the experiment achieved an average precision of 93.00%. With the PSU multiview profile-based action dataset, a new dataset for multiple viewpoints which provides profile-based action RGBD images built by our group, we achieved an average precision of 99.31%.

1. Introduction
Since 2010, action recognition methods have been increasingly developed and have been gradually introduced in healthcare applications, especially for monitoring the elderly. Action analysis plays an important role in the investigation of normal or abnormal events in daily-life activities. In such applications, privacy and convenience of usage of chosen technologies are two key factors that must be thoroughly considered. The pattern of recognized actions is an important function of a system for monitoring complex activities and behaviors which consist of several brief actions constituting a longer-term activity outcome. For example, a sleeping process involves standing/walking, sitting, and lying actions; and a falling process includes all actions mentioned above except sitting.

Recently, two main approaches have been studied and proposed for determining these actions: a wearable sensor-based technique and a vision-based technique.

Wearable inertial sensor-based devices have been used extensively in action recognition due to their small size, low power consumption, low cost, and the ease with which they can be embedded into other portable devices, such as mobile phones and smart watches. An inertial sensor used for performing navigation commonly comprises motion and rotation sensors, (e.g. accelerometers and gyroscopes). It provides the path of movement, viewpoint, velocity, and acceleration of the tracked subject. Some research studies have used wearable sensors [1–3], mobile phones [4–7], and smart watches [8] for recognizing different actions. In some research, the focus was on detection of abnormal actions, such as falling [9–11], or on reporting status for both normal and abnormal situations [12]. To recognize complex actions, moreover, several sensors must be embedded at different positions on the body. The only limitation of inertial sensors is the inconvenience presented because sensors must eventually be attached to the body, which is uncomfortable and cumbersome.
For vision-based techniques, many studies emphasize using either a single-view or multiview approach for recognizing human actions.

In a single-view approach, four types of feature representation have been used: (1) joint-based/skeleton-based, (2) motion/flow-based, (3) space-time volume-based, and (4) grid-based:

1. Joint-based/skeleton-based representation defines the characteristics of human physical structure and distinguishes its actions, for example, multilevel of joints and parts from posing features [13], the Fisher vector using skeletal quads [14], spatial-temporal feature of joints-mHOG [15], Lie vector space from a 3D skeleton [16], invariant trajectory tracking using fifteen joints [17], histogram bag-of-skeleton-codewords [18], masked joint trajectories using 3D skeletons [19], posture features from 3D skeleton joints with SVM [20], and star skeletons using HMMs for missing observations [21]. These representations result in clear human modeling, although the complexity of joint/skeleton estimation requires good accuracy from tracking and prediction.

2. Motion/flow-based representation is a global feature-based method using the motion or flow of an object, such as invariant motion history volume [22], local descriptors from optical-flow trajectories [23], KLT motion-based snippet trajectories [24], Divergence-Curl-Shear descriptors [25], hybrid features using contours and optical flow [26], motion history and optical-flow images [27], multilevel motion sets [28], projection of accumulated motion energy [29], pyramid of spatial-temporal motion descriptors [30], and motion and optical flow with Markov random fields for occlusion estimation [31]. These methods do not require accurate background subtraction but make use of acquired, inconstant features that need strategy and descriptors to manage.

3. Volume-based representations are modeled by stacks of silhouettes, shapes, or surfaces that use several frames to build a model, such as space-time silhouettes from shape history volume [32], geometric properties from continuous volume [33], spatial-temporal shapes from 3D point clouds [34], spatial-temporal features of shaplets from 3D binary cube spacetime [35], affine invariants with SVM [36], spatial-temporal micro volume using binary silhouettes [37], integral volume of visual-hull and motion history volume [38], and saliency volume from luminance, color, and orientation components [39]. These methods acquire a detailed model but must deal with high dimensions of features which require accurate human segmentation without the background.

4. Grid-based representations divide the observation region of interest into cells, a grid, or overlapped blocks to encode local features, for example, a grid or histogram of oriented rectangles [40], flow descriptors from spatial-temporal small cells [41], histogram of local binary patterns from a spatial grid [42] and rectangular optical-flow grid [43], codeword features for histograms of oriented gradients and histograms of optical flow [44], 3D interest points within multisize windows [45], histogram of motion gradients [46], and combination of motion history, local binary pattern, and histogram of oriented gradients [47]. This method is simple for feature modeling in the spatial domain, but it must deal with some duplicate and insignificant features.

Although the four types of representation described in the single-view approach are generally good, in monitoring a large area, one single camera will lose its ability to determine continuous human daily-life actions due to view variance, occlusion, obstruction, and lost information, among others. Thus, a multiview approach is introduced to lessen the limitations of a single-view approach.

In the multiview approach, methods can be categorized into 2D and 3D methods.

Examples of the 2D methods are layer-based circular representation of human model structure [48], bag-of-visual-words using spatial-temporal interest points for human modeling and classification [49], view-invariant action masks and movement representation [50], R-transform features [51], silhouette feature space with PCA [52], low-level characteristics of human features [53], combination of optical-flow histograms and bag-of-interest-point-words using transition HMMs [54], contour-based and uniform local binary pattern with SVM [55], multifeatures with key poses learning [56], dimension-reduced silhouette contours [57], action map using linear discriminant analysis on multiview action images [58], posture prototype map using self-organizing map with voting function and Bayesian framework [59], multiview action learning using convolutional neural networks with long short term memory [60], and multiview action recognition with an autoencoder neural network for learning view-invariant features [61].

Examples of the 3D method, where the human model is reconstructed or modeled from features between views, are pyramid bag-of-spatial-temporal-descriptors and part-based features with induced multitask learning [62], spatial-temporal logical graphs with descriptor parts [63], temporal shape similarity in 3D video [64], circular FFT features from convex shapes [65], bag-of-multiple-temporal-self-similar-features [66], circular shift invariance of DFT from movement [67], and 3D full body/pose dictionary features with convolutional neural networks [68]. All of these 3D approaches attempt to construct a temporal-spatial data model that is able to increase the model precision and, consequently, raise the accuracy of the recognition rate.

The multiview approach, however, has some drawbacks. The methods need more cameras and hence are more costly. It is a more complex approach in terms of installation, camera calibration between viewpoints, and model building and hence is more time-consuming. In actual application, however, installation and setup should be simple, flexible, and as easy as possible. Systems that are calibration-free or automatically self-calibrating between viewpoints are sought.
Computational Intelligence and Neuroscience

Multi-view Fusion

Layer Feature Fusion

Action

Single-view 1

Preprocessing → Layer Human Feature Extraction

Multi-view Fusion

Layer Feature Fusion

Action

Single-view N

Preprocessing → Layer Human Feature Extraction

Figure 1: Overview of layer fusion model.

Figure 2: Preprocessing and human depth profile extraction; (a) 8-bit depth image acquired from depth camera; (b) motion detected output from mixture-based Gaussian model for background subtraction; (c) blob position: consists of top-left position \((X_b, Y_b)\) and width \((W_b)\) and height \((H_b)\); (d) rejected arm parts of human blob; (e) depth profile and position of human blob: \(X_h, Y_h, W_h,\) and \(H_h\).

2. Layer Fusion Model

Our layer fusion model is described in three parts: (1) preprocessing for image quality improvement; (2) human modeling and feature extraction using a single-view layer feature extraction module; and (3) fusion of features from any view into one single model using layer feature fusion module and classifying to actions. The system overview is shown in Figure 1.

2.1. Preprocessing. The objective of preprocessing is to segregate human structure from the background and to eliminate arm parts before extracting the features, as depicted in Figure 2. In our experiment, the structure of the human in the foreground is extracted from the background by applying motion detection using mixture of Gaussian model segmentation algorithms \[69\]. The extracted motion image \((I_m)\) (Figure 2(b)), from the depth image \((I_d)\) (Figure 2(a)), is assumed to be the human object. However, \(I_m\) still contains noise from motion detection, which has to be reduced by morphological noise removal operator. The depth of the extracted human object, defined by its depth values inside the object, is then improved \((I_{mo})\) as determined by intersecting \(I_d\) and \(I_m\) using the AND operation: \(I_{mo} = I_m \& I_d\).

The human blob, with bounding rectangle coordinates \(X_h\) and \(Y_h\), width \(W_h\), and height \(H_h\) as shown in Figure 2(c), is located using a contour approximation technique. However, our action technique emphasizes only the profile structure of the figure, while hands and arms are excluded, as seen in Figure 2(d), and the obtained structure is defined as the figure depth profile \((I_{mo})\) (Figure 2(e)) in further recognition steps.

One problem facing a person within camera view, be it one single camera or a multitude of cameras, is that of privacy and lighting conditions. Vision-based and profile-based techniques involve the use of either RGB or non-RGB. The former poses a serious problem to privacy. Monitoring actions in private areas using RGB cameras make those under surveillance feel uncomfortable because the images expose more clearly their physical outlines. As for lighting conditions, RGB is also susceptible to intensity; images often deteriorate in dim environments. The depth approach helps solve both problems; a coarse depth profile of the subject is adequate for determining actions, and depth information can prevent illumination change issues, which are a serious problem in real-life applications of round-the-clock surveillance.

The depth approach that is adopted in our research together with a multiview arrangement is considered worthy of more costly installation than the single-view approach.

A gap that needs attention for most multiview, non-RGB results is that of perspective robustness, or viewing-orientation stability, and model complexity. Under a calibration-free setup, our research aims to contribute to the development of a fusion technique that is robust and simple in evaluating the depth profile of human action recognition. We have developed a layer fusion model in order to fuse depth profile features from multiviews and to test our technique on a triple dataset of validation and efficiency. The three datasets tested are the Northwestern-UCLA dataset, the i3DPost dataset, and the PSU dataset for multiview action from various viewpoints.

The following sections detail our model, its results, and comparisons.
2.2. Layer Human Feature Extraction. We model the depth profile of a human in a layered manner that allows extraction of specific features depending on the height level of the human structure, which possesses different physical characteristics. The depth profile is divided vertically into odd-numbered layers (e.g., 5 layers, as shown in Figure 3) with a specific size, regardless of the distance, perspective, and views, which would allow features of the same layer from all views to be fused into reliable features.

The human object in the bounding rectangle is divided into equal layers to represent features at different levels of the structure. The total number of layers is 2N+1, where N is the maximum number of upper or lower layers. For example, in Figure 3, the human structure is divided into five layers. The horizontal boundaries of all layers, the red vertical lines, are defined by the left and right boundaries of the human object. The vertical boundaries of each layer, shown as yellow horizontal lines, are defined by the top $y_T[k]$ and the bottom $y_B[k]$ values that can be computed as follows:

$$y_T[k] = \frac{H_k (k + N)}{2N + 1} + 1$$  

(1)

$$y_B[k] = \frac{H_k (k + N + 1)}{2N + 1}$$  

(2)

The region of interest in layer $k$ is defined as $x = 0$ to $W_h$ and $y = y_T[k]$ to $y_B[k]$.

According to the model, features from the depth profile human object can be computed along with layers as concatenated features of every segment using basic and statistical properties (e.g., axis, density, depth, width, and area). Depth can also be distinguished for specific characteristic of actions.

In our model, we define two main features for each layer, including the density ($\rho[k]$) and weighted depth density ($Z[k]$) of layers. In addition, the proportion value ($P_v$) is defined as a global feature to handle the horizontal action.

2.2.1. Density of Layer. The density of layer ($\rho[k]$) indicates the amount of object at a specific layer, which varies distinctively according to actions, and can be computed as the number of white pixels in the layer, as shown in the following equation:

$$\rho'[k] = \sum_{y=y_T[k]}^{y_B[k]} \sum_{x=0}^{W_h} l_m(x, y)$$  

(3)

In multiple views, different distances between the object and the cameras would affect $\rho'[k]$. Objects close to a camera certainly appear larger than when further away, and thus $\rho'[k]$ must be normalized in order for these to be fused. We use the maximum value of the perceived object and normalize it, employing the following equation:

$$\rho[k] = \frac{\rho'[k]}{\text{arg max} (\rho'[k])}$$  

(4)

2.2.2. Weighted Depth Density of Layer. An inverse depth density is additionally introduced to improve the pattern of density feature. The procedure is comprised of two parts: inverse depth extraction and weighting for density of the layer.

At the outset, depth extraction is applied to the layer profile. The depth profile reveals the surface of the object that indicates rough structure ranging from 0 to 255 or from near to far distances from the camera. According to perspective projection varying in a polynomial form, a depth value at a near distance, for example, from 4 to 5, has a much smaller real distance than a depth value at a far distance, for example, from 250 to 251. The real-range depth of the layer ($D'[k]$) translates the property of 2D depth values to real 3D depth values in centimeters. The real-range depth better distinguishes the depth between layers of the object—different
parts of the human body—and increases the ability to classify actions. A polynomial regression [70] has been used to convert the depth value to real depth, as described in the following equation:

\[ f_c' (x) = 1.2512e^{-10} x^6 - 1.0370e^{-7} x^5 + 3.5014e^{-5} x^4 - 0.0061 x^3 + 0.5775 x^2 - 27.6342 x + 5.6759 e^3 \]  

In the next step, we apply the inverse real-range depth (D*[k]*), hereafter referred to as the inverse depth, for weighting the density of layer (\( \rho[k] \)) that increases the probability of classification for certain actions, such as sitting and stooping. We establish the inverse depth, for employing the following equation:

\[ D'[k] = f_c' \left( \frac{\sum_{y=y_1[k]} \sum_{x=0} W_h I_{mb} (x,y)}{\rho [k]} \right) \]  

In addition, the (D*[k]*) value of each layer is represented by the converted value of every depth value averaged in that layer, as defined in the following equation:

\[ D'[k] = \frac{D'[k]}{\arg \max \left( D'[k] \right)} \]  

In the next step, we apply the inverse real-range depth (D*[k]*) hereafter referred to as the inverse depth, for weighting the density of the layer in order to enhance the feature (\( \rho[k] \)) that increases the probability of classification for certain actions, such as sitting and stooping. We establish the inverse depth, as described in (8), to measure the hidden volume of body structure which distinguishes particular actions from others: for example, in Table 1, in viewing stooping from the front, the upper body is hidden but the value of D*[k]* can reveal the volume of this zone; and in viewing sitting from the front, the depth of the thigh will reveal the hidden volume compared to other parts of the body.

\[ D_1[k] = \left( \frac{D[k] - \arg \max(D[k])}{\arg \max(D[k])} \right) \]  

2.2.3. Proportion Value. Proportion value (\( P_v \)) is a penalty parameter of the model to indicate roughly the proportion of object vertical actions distinguished from horizontal actions. It is the ratio of the width \( W_h \) and the height \( H_h \) of the object in each view (see the following equation):

\[ P_v = \frac{W_h}{H_h} \]  

Table 1 mentioned earlier shows the depth profile of action and its features in each view. In general, the feature patterns of each action are mostly similar, though they do exhibit some differences depending on the viewpoints. It should be noted here that the camera(s) are positioned at 2 m above the floor, pointing down at an angle of 30° from the horizontal line.

Four actions in Table 1, standing/walking, sitting, stooping, and lying, shall be elaborated here. For standing/walking, the features are invariant to viewpoints for both density (\( \rho[k] \)) and inverse depth (D*[k]*) The D*[k]* values slope equally for every viewpoint due to the position of the camera(s). For sitting, the features vary for \( \rho[k] \) and D*[k]* according to their viewpoints. However, the patterns of sitting for front and slant views are rather similar to standing/walking. D*[k]* from some viewpoints indicates the hidden volume of the thigh. For stooping, the \( \rho[k] \) patterns are quite the same from most viewpoints, except for the front view and back view, due to occlusion of the upper body. However, D*[k]* reveals clearly the volume in stooping. For lying, the \( \rho[k] \) patterns vary depending on the viewpoints and cannot be distinguished using layer-based features. In this particular case, the proportion value (\( P_v \)) is introduced to help identify the action.

2.3. Layer-Based Feature Fusion. In this section, we emphasize the fusion of features from various views. The pattern of features can vary or self-overlap with respect to the viewpoint. In a single view, this problem leads to similar and unclear features between actions. Accordingly, we have introduced a method for the fusion of features from multiviews to improve action recognition. From each viewpoint, three features of action are extracted: (1) density of layer (\( \rho[k] \)), (2) weighted depth density of layers (Z*[k]*), and (3) proportion value (\( P_v \)).

We have established two fused features as the combination of width, area, and volume of the body structure from every view with respect to layers. These two fused features are the mass of dimension (\( \omega[k] \)) and the weighted mass of dimension (\( \omega[k] \)).

\[ Z[k] = (1 - \alpha) (Q[k]) + (\alpha) (\rho[k]) \]  

As can be deduced from Table 2, the weighted depth density of layers (Z*[k]*) improves the feature pattern for better differentiation for 13 out of 20 features, is the same for 5 features (I through V on standing-walking), and is worse for 2 features (VIII on side-view sitting and X on back-view sitting). Thus, Z*[k]* is generally very useful, though the similar and worse outcomes would require a further multiview fusion process to distinguish the pattern.
| Action/view       | $\rho_k$ | $D_k$ | $I_{max}$ |
|------------------|---------|-------|----------|
| Standing-walking/front |  |   |   |
| Standing-walking/slant |  |   |   |
| Standing-walking/side |  |   |   |
| Standing-walking/backslant |  |   |   |
| Standing-walking/back |  |   |   |
| Sitting/front |  |   |   |
| Sitting/slant |  |   |   |
| Lying/front |  |   |   |
| Lying/slant |  |   |   |

Table 1: Sampled human profile depth in various views and actions.
| Action/view   | $I_{mo}$ | $\rho[k]$ | $D[k]$ | $D_i[k]$ | Action/view   | $I_{mo}$ | $\rho[k]$ | $D[k]$ | $D_i[k]$ |
|--------------|---------|-----------|--------|----------|--------------|---------|-----------|--------|----------|
| Sitting/side | ![Image] | ![Image]  | ![Image] | ![Image] | Lying/side   | ![Image] | ![Image]  | ![Image] | ![Image] |
| Sitting/back slant | ![Image] | ![Image]  | ![Image] | ![Image] | Lying/back slant | ![Image] | ![Image]  | ![Image] | ![Image] |
| Sitting/back | ![Image] | ![Image]  | ![Image] | ![Image] | Lying/back   | ![Image] | ![Image]  | ![Image] | ![Image] |

Note: (I) $\rho[k]$ is density of layer. (II) $D[k]$ is real-range depth of layer. (III) $D_i[k]$ is inverse depth of layer.
### Table 2: Sampled features in single view and fused features in multiview.

| Action/view                     | Camera 1 | Camera 2 | $\bar{z}[k]$ |
|---------------------------------|----------|----------|--------------|
| (I) Standing-walking/front      | $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
| (II) Standing-walking/slant     | $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
| (III) Standing-walking/side     | $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
| (IV) Standing-walking/back slant| $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
| (V) Standing-walking/back       | $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
| (VI) Sitting/front              | $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
| (VII) Sitting/slant             | $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
| (VIII) Sitting/side             | $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
| (IX) Sitting/back slant         | $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
| (X) Sitting/back                | $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
| (XI) Stooping/front             | $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
| (XII) Stooping/slant            | $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
| (XIII) Stooping/side            | $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
| (XIV) Stooping/back slant       | $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
| (XV) Stooping/back              | $I_{mx}$ | $Z[k] \mid \alpha = 0.9$ |              |
Table 2: Continued.

| Action/view | Camera 1 | Camera 2 |
|------------|----------|----------|
| (XVI) Lying/front | $I_{mn}$ | $Z[k] \mid \alpha = 0.9$ | $I_{mn}$ | $Z[k] \mid \alpha = 0.9$ |
| (XVII) Lying/slant | $\bar{o}_1$ | $\bar{o}_2$ | $\bar{o}_3$ | $\bar{o}_4$ |
| (XVIII) Lying/side | $\bar{o}_5$ | $\bar{o}_6$ | $\bar{o}_7$ | $\bar{o}_8$ |
| (XIX) Lying/back slant | $\bar{o}_9$ | $\bar{o}_{10}$ | $\bar{o}_{11}$ | $\bar{o}_{12}$ |
| (XX) Lying/back | $\bar{o}_{13}$ | $\bar{o}_{14}$ | $\bar{o}_{15}$ | $\bar{o}_{16}$ |

Note: (I) $Z[k]$ is weighted depth density of layers which uses $\alpha$ for adjustment between $Q[k]$ and $\rho[k]$. (II) $\bar{o}_k$, fused features from multiview, called the weighted mass of dimension feature.

The depth profiles, particularly in front views, affect the feature adjustment. However, the classifier can still distinguish appearances, because generally patterns of features in the upper layers are shorter than those in the lower layers.

3. Experimental Results

Experiments to test the performance of our method were performed on three datasets: the PSU (Prince of Songkla University) dataset, the NW-UCLA (Northwestern-University of California at Los Angeles) dataset, and the i3DPost dataset.

3.1. Experiments on the PSU Dataset.

The PSU dataset [76] contains 328 video clips of human profiles with four basic actions recorded in two views using RGBD cameras (Kinect...
The videos were simultaneously captured and synchronized between views. The profile-based actions consisted of standing/walking, sitting, stooping, and lying. Two scenarios, one in a work room for training and another in a living room for testing, were performed. Figure 4 shows an example of each scenario, together with the viewpoints covered.

Two Kinect cameras were set overhead at 60° to the vertical line, each at the end of a 2 m pole. RGB and depth information from multiviews was taken from the stationary cameras with varying viewpoints to observe the areas of interest. The operational range was about 3-5.5 m from the cameras to accommodate a full body image, as illustrated in Figure 5. The RGB resolution for the video dataset was 640×480, while depths at 8 and 24 bits were also of the same resolution. Each sequence was performed by 3-5 actors, having no less than 40 frames of background at the beginning to allow motion detection using any chosen background subtraction technique. The frame rate was about 8-12 fps.

(i) Scenario in the Work Room (Training Set). As illustrated in Figure 6(a), the two cameras’ views are perpendicular to each other. There are five angles of object orientation: front (0°), slant (45°), side (90°), rear-slant (135°), and rear (180°), as shown in Figure 6(b). A total of 8,700 frames were obtained in this scenario for training.

(ii) Scenario in the Living Room (Testing Set). This scenario, illustrated in Figure 7, involves one moving Kinect camera at four angles: 30°, 45°, 60°, and 90°, while another Kinect camera remains stationary. Actions are performed freely in various directions and positions within the area of interest. A total of 10,720 frames of actions were tested.

3.1.1. Evaluation of the Number of Layers. We determine the appropriate number of layers by testing our model with different numbers of layers (L) using the PSU dataset. The numbers of layers for testing are 3, 5, 7, 9, 11, 13, 15, 17, and 19. The alpha value (α) is instinctively fixed at 0.7 to find the optimal value of the number of layers. Two classification methods are used for training and testing: an artificial neural network (ANN) with a back-propagation algorithm over 20 nodes of hidden layers and a support vector machine (SVM) using a radial basis function kernel along with C-SVC. The results using ANN and SVM are shown in Figures 8 and 9, respectively.

Figure 8 shows that the 3-layer size achieves the highest average precision of 94.88% using the ANN and achieves 92.11% using the SVM in Figure 9. Because the ANN performs better, in the tests that follow, the evaluation of our
model will be based on this layer size together with the use
of the ANN classifier.

3.1.2. Evaluation of Adjustable Parameter $\alpha$. For the weighted mass of dimension feature, $(\omega[k])$, the adjustment parameter alpha ($\alpha$)—the value for the weight between the inverse depth density of layers ($Q[k]$) and the density of layer ($\rho[k]$)—is employed. The optimal $\alpha$ value is determined to show the improved feature that uses inverse depth to reveal hidden volume in some parts and the normal volume. The experiment is carried out by varying alpha from 0 to 1 at 0.1 intervals.

Figure 10 shows the precision of action recognition using a 3-layer size and the ANN classifier versus the alpha values. In general, except for the sitting action, one may note that as the portion of inverse depth density of layers ($Q[k]$) is augmented, precision increases. The highest average precision is 95.32\% at $\alpha = 0.9$, meaning that 90\% of the inverse depth density of layers ($Q[k]$) and 10\% of the density of layers ($\rho[k]$) are an optimal proportion. When $\alpha$ is above 0.9, all precision values drop. The trend for sitting action is remarkably different from others in that precision always hovers near the maximum and gradually but slightly decreases when $\alpha$ increases.

Figure 11 illustrates the multiview confusion matrix of action precisions when $L = 3$ and $\alpha = 0.9$, using the PSU dataset. We found that standing/walking action had the highest precision (99.31%), while lying only reached 90.65\%. The classification error of the lying action depends mostly on its characteristic that the principal axis of the body is
aligned horizontally, which works against the feature model. In general, the missed classifications of standing/walking, stooping, and lying actions were mostly confused with the sitting action, accounting for 0.69%, 4.82%, and 8.83% of classifications, respectively. Nevertheless, the precision of sitting is relatively high at 98.59%.

3.1.3. Comparison of Single View/Multiview. For the sake of comparison, we also evaluate tests in single-view recognition for the PSU dataset for $L = 3$ and $\alpha = 0.9$ in the living room scene, similar to that used to train the classification model for the work room. Figure 12 shows the results from the single-view Kinect 1 (stationary camera), while Figure 13 shows those from the single-view Kinect 2 (moving camera).

Results show that the single-view Kinect 1, which is stationary, performs slightly better than the single-view Kinect 2, which is moving (average precision of 92.50% compared to 90.63%). The stationary camera gives the best results for sitting action, while for the moving camera, the result is best for the standing/walking action. It is worth noting that the stationary camera yields a remarkably better result for lying action than the moving one.

Figure 14 shows the precision of each of the four postures, together with that of the average, for the multiview and the two single views. On average, the result is best accomplished with the use of the multiview and is better for all postures other than the lying action. In this regard, single-view 1 yields a slightly better result, most probably due to its stationary viewpoint toward the sofa, which is perpendicular to its line of sight.

3.1.4. Comparison of Angle between Cameras. As depicted earlier in Figure 7, the single-view Kinect 1 camera is stationary, while the single-view Kinect 2 camera is movable, adjusted to capture viewpoints at $30^\circ$, $45^\circ$, $60^\circ$, and $90^\circ$ to the stationary camera. We test our model on these angles to assess
the robustness of the model on the four postures. Results are shown in Figure 15.

In Figure 15, the lowest average precision result occurs at $30^\circ$ — the smallest angle configuration between the two cameras. This is most probably because the angle is narrow and thus not much additional information is gathered. For all other angles, the results are closely clustered. In general, standing/walking and sitting results are quite consistent at all angles, while lying and stooping are more affected by the change.

### 3.1.5. Evaluation with NW-UCLA Trained Model

In addition, we tested the PSU dataset in living room scenes using the model trained on the NW-UCLA dataset [63]. Figure 16 illustrates that the 9-layer size achieves the highest average precision at 93.44%. Sitting gives the best results and the highest precision, up to 98.74% when $L = 17$. However, sitting at low layers also gives good results, for example, $L = 3$ at 97.18%, while the highest precision for standing/walking is 95.40%, and the lowest precision is 85.16%. The lowest precision for stooping is 92.08% when $L = 5$.

Figure 17 shows the results when $L = 9$, which illustrates that standing/walking gives the highest precision at 96.32%, while bending gives the lowest precision at 88.63%.

In addition, we also compare precision of different angles between cameras, as shown in Figure 18. The result shows that the highest precision on average is 94.74% at $45^\circ$, and the lowest is 89.69% at $90^\circ$.
Table 3: Time consumption testing on different numbers of layers and different classifiers.

| Classifier | L=3 Min | L=3 Max | L=3 Avg | L=11 Min | L=11 Max | L=11 Avg | L=19 Min | L=19 Max | L=19 Avg | L=3 Avg | L=11 Avg | L=19 Avg |
|------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| ANN        | 13.95   | 17.77   | 15.08   | 14.12   | 20.58   | 15.18   | 14.43   | 20.02   | 15.85   | 66.31   | 65.88   | 63.09   |
| SVM        | 14.01   | 17.40   | 15.20   | 14.32   | 18.97   | 15.46   | 14.34   | 19.64   | 15.71   | 65.79   | 64.68   | 63.65   |

Note: L stands for the number of layers. ANN, artificial neural network; SVM, support vector machine.

3.1.6. Evaluation of Time Consumption. Time consumption evaluation, excluding interface and video showing time, is conducted using the OpenMP wall clock. Our system is tested on a normal PC (Intel® Core™ i5 4590 at 3.30 GHz with 8 GB DDR3). We use the OpenCV library for computer vision, the OpenMP library for parallel processing, and CLNUI to capture images from the RGBD cameras. The number of layers and the classifier are tested using 10,720 action frames from the living room scene.

On average, the time consumption is found to be approximately 15 ms per frame or a frame rate of around 63 fps. As detailed in Table 3, the number of layers and the type of classifier affect the performance only slightly. In addition, we compare serial processing with parallel processing, which divides a single process into threads. The latter is found to be 1.5507 times faster than the former. It was noted that thread initialization and synchronization consume a portion of the computation time.

3.2. Experiment on the NW-UCLA Dataset. The NW-UCLA dataset [63] is used to benchmark our method. This dataset is similar to our work for the multiview action 3D PSU dataset taken at different viewpoints to capture the RGB and depth images. The NW-UCLA dataset covers nearly ten actions including stand up, walk around, sit down, and bend to pick up item, but it lacks a lie down action. Actions in this dataset are marked in colors. To test our method, the motion detection step for segmenting movement is replaced by a specific color segmentation to obtain the human structure. The rest of the procedure stays the same.

Only actions of interest are selected for the test—its transition actions are excluded. For example, standing/walking frames are extracted from stand up and walk around; sitting frames are selected from sit down and stand up; and stooping is extracted from pick up with one hand and pick up with both hands.

Our method employs the model learned using the PSU dataset in the work room scenario to test the NW-UCLA dataset. All parameters in the test are the same except that alpha is set to zero due to variations in depth values. The experiment is performed for various numbers of layers from L = 3 to L = 19. Test results on the NW-UCLA dataset are shown in Figure 19.

From Figure 19, the maximum average precision (86.40%) of the NW-UCLA dataset is obtained at layer L = 11, in contrast to the PSU dataset at L = 3. Performance for stooping is generally better than other actions and peaks at 95.60%. As detailed in Figure 20, standing/walking gives the lowest precision at 76.8%. The principal cause of low
Table 4: Comparison between NW-UCLA and our recognition systems on the NW-UCLA dataset.

| NW-UCLA recognition action [63] | Precision (%) | Our recognition action | Precision (%) |
|---------------------------------|--------------|------------------------|--------------|
| Walk around                     | 77.60        | Walking/standing       | 76.80        |
| Sit down                        | 68.10        | Sitting                | 86.80        |
| Bend to pick item (1 hand)      | 74.50        | Stooping               | 95.60        |
| Average                         | 73.40        | Average                | 86.40        |

Figure 19: Precision of action by layer size on NW-UCLA dataset.

Figure 20: Confusion matrix of test results on the NW-UCLA 3D dataset when L = 11.

precision is that the angle of the camera and its captured range are very different from the PSU dataset. Compared with the method proposed by NW-UCLA, our method performs better by up to 13 percentage points, from an average of 73.40% to 86.40%, as shown in Table 4. However, to be fair, many more activities are considered by the NW-UCLA than ours which focuses only on four basic actions and hence its disadvantage by comparison.

3.3. Experiment on the i3DPost Dataset. The i3DPost dataset [77] is an RGB multiview dataset that contains 13 activities. The dataset was captured by 8 cameras from different viewpoints with 45° between cameras, performed by eight persons for two sets at different positions. The background images allow the same segmentation procedure to build the nondepth feature vector for recognizing profile-based action.

The testing sets extract only target actions from temporal activities, including standing/walking, sitting, and stooping from sit-standup, walk, walk-sit, and bend.

3.3.1. Evaluation of i3DPost Dataset with PSU-Trained Model. Firstly, i3DPost is tested by the PSU-trained model from 2 views at 90° between cameras on different layer sizes.

Figure 21 shows the testing results of the i3DPost dataset using the PSU-trained model. The failed prediction for sitting shows a precision of only 28.08% at L = 9. By observation, the mistake is generally caused by the action of sitting that looks like a squat in the air, which is predicted as standing. In the PSU dataset, sitting is done on a bench/chair. On the other hand, standing/walking and stooping are performed with good results of about 96.40% and 100% at L = 11, respectively.
Figure 22: Confusion matrix of 2 views for i3DPost dataset using PSU-trained model when L = 9.

Figure 23: Precision of i3DPost dataset with new trained model by layer size.

Figure 24: Confusion matrix of test result on the 2-view i3DPost dataset when L = 17.

Figure 25: Precision comparison graph for different angles in the i3DPost.

Figure 26: Precision of i3DPost dataset according to the number of views.
Table 5: Comparison between our recognition systems and [59] on the i3DPost dataset.

| Recognition of similar approach [59] | Precision (%) | Our recognition action | Precision (%) |
|--------------------------------------|---------------|------------------------|---------------|
| Walk                                 | 95.00         | Walking/standing       | 98.28         |
| Sit                                  | 87.00         | Sitting                | 81.03         |
| Bend                                 | 100.0         | Stooping               | 99.68         |
| Average                              | 94.00         | Average                | 93.00         |

Table 6: Precision results of our study and others emphasizing profile-based action recognition using different methods and datasets.

| Method                              | Standing | Walking | Sitting | Stooping | Lying | Average |
|-------------------------------------|----------|---------|---------|----------|-------|---------|
| P. Chawalitsittikul et al. [70]     | 98.00    | 98.00   | 93.00   | 94.10    | 98.00 | 96.22   |
| N. Noorit et al. [71]               | 99.41    | 80.65   | 89.26   | 94.35    | 100.0 | 92.73   |
| M. Ahmad et al. [72]                | -        | 89.00   | 85.00   | 100.0    | 91.00 | 91.25   |
| C. H. Chuang et al. [73]            | -        | 92.40   | 97.60   | 95.40    | -     | 95.80   |
| G. I. Parisi et al. [74]            | 96.67    | 90.00   | 83.33   | -        | 86.67 | 89.17   |
| N. Sawant et al. [75]               | 91.85    | 96.14   | 85.03   | -        | -     | 91.01   |
| Our method on PSU dataset            | 99.31    | 99.31   | 98.59   | 92.72    | 90.65 | 95.32   |

Figure 26: Precision in i3DPost dataset with new trained model by the number of views.

L = 17, L = 17, L = 7, and L = 13, respectively. We noticed that the highest precision is for 2 views. In general, the performance increases when the number of views increases, except for sitting. Moreover, the number of layers that give maximum precision is reduced as the number of views increases. In conclusion, only two or three views from different angles are necessary for obtaining the best performance.

We compared our method with a similar approach [59], based on a posture prototype map and results voting function with a Bayesian framework for multiview fusion. Table 5 shows the comparison results. The highest precisions of our method and the comparison approach are 99.68% and 100% for stooping and bend, respectively. Likewise, the lowest precisions are 81.03% and 87.00% for the same actions. However, for walking/standing, our approach obtains better results. On average, the comparison approach performs slightly better than our method.

4. Comparison with Other Studies

Our study has now been presented with other visual profile-based action recognition studies and also compared with other general action recognition studies.

4.1. Precision Results of Different Studies. Table 6 shows precision results of various methods emphasizing profile-based action recognition. Note that the methods employed different algorithms and datasets; thus results are presented only for the sake of studying their trends with respect to actions. Our method is tested on the PSU dataset. Our precision is highest on walking and sitting actions, quite good on the standing action, quite acceptable on the lying action, but poor on the stooping action. It is not possible to compare on precision due to lack of information on the performance of other methods. However, we can note that each method performs better than others on different actions; for example, [74] is good for standing, and [73] is a better fit for sitting.

4.2. Comparison with Other Action Recognition Studies. Table 7 compares the advantages and disadvantages of some previous studies concerning action recognition acquired according to the criteria of their specific applications. In the inertial sensor-based approach, sensors/devices are attached to the body and hence monitoring is available everywhere, for the inconvenience of carrying them around. This approach gives high privacy but is highly complex. In an RGB vision-based approach, sensors are not attached to the body, and hence it is less cumbersome. Though not so complex, the main drawback of this method is the lack of privacy. In this regard, depth-based views may provide more privacy. Although rather similar in comparison in the table, the multiview approach can cope with some limitations, such
Table 7: Comparison of some action recognition approaches based on ease of use, perspective robustness, and real-world application conditions.

| Approach                          | Flexible calibration | View scalability | Perspective robustness | Real-world Application Condition |
|-----------------------------------|----------------------|------------------|------------------------|----------------------------------|
| Inertial sensor-based [1–12]      | N/A                  | N/A              | High                   | Highly complex                   |
| RGB vision-based [13, 17, 21–47, 71–73, 75] | Low-moderate         | Low-moderate     | Low-moderate           | Simple-moderate                  |
| Depth-based Single-view [14–16, 18–20, 70, 74] | Low-moderate         | Low-moderate     | Low-moderate           | Depends                          |
| RGB/depth-based multiview [48–68] | Low-moderate         | Low-moderate     | Low-moderate           | Not cumbersome                   |
| Our work, depth-based multiview   | Moderate-high        | Moderate-high    | High                   | Simple                           |
|                                   |                      |                  |                        | Not cumbersome                   |

*Complexity: Highly complex, Simple-moderate, Depends
Cumbersomeness from sensor/device attachment: Cumbersome, Not cumbersome
Privacy: High, Lacking, Moderate-high, Depends
as vision coverage and continuity or obstruction, which are common in the single-view approach, as described in some detail in Introduction.

(i) As for our proposed work, it can be seen from the table that the approach is in general quite good in comparison; in addition to being simple, it offers a high degree of privacy and other characteristics such as flexibility, scalability, and robustness are at similar levels, if not better, and no calibration is needed, which is similar to most of the other approaches.

However, the two cameras for our multiview approach still have to be installed following certain specifications, such as the fact that the angle between cameras should be more than 30°.

5. Conclusion/Summary and Further Work

In this paper, we explore both the inertial and visual approaches to camera surveillance in a confined space such as a healthcare home. The less cumbersome vision-based approach is studied in further detail for single-view and multiview RGB depth-based and for non-RGB depth-based approaches. We decided on a multiview, non-RGB depth-based approach for privacy and have proposed a layer fusion model, which is representation-based model that allows fusion of features and information by segmenting parts of the object into vertical layers.

We trained and tested our model on the PSU dataset with four postures (standing/walking, sitting, stooping, and lying) and have evaluated the outcomes using the NW-UCLA and i3DPost datasets on available postures that could be extracted. Results show that our model achieved an average precision of 95.32% on the PSU dataset—on a par with many other achievements, if not better, 93.00% on the i3DPost dataset, and 86.40% on the NW-UCLA dataset. In addition to flexibility of installation, scalability, and noncalibration of features, one advantage over most other approaches is that our approach is simple, while it contributes good recognition on various viewpoints with a high speed of 63 frames per second, suitable for application in a real-world setting.

In further research, a spatial-temporal feature is interesting and should be investigated for more complex action recognition, such as waving and kicking. Moreover, reconstruction of 3D structural and bag-of-visual-word models is also of interest.

Data Availability

The PSU multiview profile-based action dataset is available at [74], which is authorized only for noncommercial or educational purposes. The additional datasets to support this study are cited at relevant places within the text as references [63] and [75].

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partially supported by the 2015 Graduate School Thesis Grant, Prince of Songkla University (PSU), and Thailand Center of Excellence for Life Sciences (TECLS), and the authors wish to express their sincere appreciation. The first author is also grateful for the scholarship funding granted for him from Rajamangala University of Technology Srivijaya to undertake his postgraduate study, and he would like to express gratitude to Assistant Professor Dr. Nikom Suvonvorn for his guidance and advice. Thanks are extended to all members of the Machine Vision Laboratory, PSU Department of Computer Engineering, for the sharing of their time in the making of the PSU profile-based action dataset. Last but not least, special thanks are due to Mr. Wiwat Sutiwipakorn, a former lecturer at the PSU Faculty of Engineering, regarding the use of language in the manuscript.

References

[1] Z. He and X. Bai, “A wearable wireless body area network for human activity recognition,” in Proceedings of the 6th International Conference on Ubiquitous and Future Networks, ICUFN 2014, pp. 115–119, China, July 2014.

[2] C. Zhu and W. Sheng, “Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 41, no. 3, pp. 569–573, 2011.

[3] J.-S. Sheu, G.-S. Huang, W.-C. Jheng, and C.-H. Hsiao, “Design and implementation of a three-dimensional pedometer accumulating walking or jogging motions,” in Proceedings of the 2nd International Symposium on Computer, Consumer and Control, IS3C 2014, pp. 828–831, Taiwan, June 2014.

[4] R. Samiei-Zonouz, H. Memarzadeh-Tehran, and R. Rahmani, “Smartphone-centric human posture monitoring system,” in Proceedings of the 2014 IEEE Canada International Humanitarian Technology Conference, IHTC 2014, pp. 1–4, Canada, June 2014.

[5] X. Yin, W. Shen, J. Samarabandu, and X. Wang, “Human activity detection based on multiple smart phone sensors and machine learning algorithms,” in Proceedings of the 19th IEEE International Conference on Computer Supported Cooperative Work in Design, CSCWD 2015, pp. 582–587, Italy, May 2015.

[6] C. A. Siebra, B. A. Sa, T. B. Gouveia, F. Q. Silva, and A. L. Santos, “A neural network based application for remote monitoring of human behaviour,” in Proceedings of the 2015 International Conference on Computer Vision and Image Analysis Applications (ICCVIA), pp. 1–6, Sousse, Tunisia, January 2015.

[7] C. Pham, “MobiRAR: real-time human activity recognition using mobile devices,” in Proceedings of the 7th IEEE International Conference on Knowledge and Systems Engineering, KSE 2015, pp. 144–149, Vietnam, October 2015.

[8] G. M. Weiss, J. L. Timko, C. M. Gallagher, K. Yoneda, and A. J. Schreiber, “Smartwatch-based activity recognition: A machine learning approach,” in Proceedings of the 3rd IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2016, pp. 426–429, USA, February 2016.

[9] Y. Wang and X.-Y. Bai, “Research of fall detection and alarm applications for the elderly,” in Proceedings of the 2013 International Conference on Mechatronic Sciences, Electric Engineering
and Computer, MEC 2013, pp. 615–619, China, December 2013.

[10] Y. Ge and B. Xu, “Detecting falls using accelerometers by adaptive thresholds in mobile devices,” Journal of Computers in Academy Publisher, vol. 9, no. 7, pp. 1553–1559, 2014.

[11] W. Liu, Y. Luo, J. Yan, C. Tao, and L. Ma, “Falling monitoring system based on multi axial accelerometer,” in Proceedings of the 2014 11th World Congress on Intelligent Control and Automation (WCICA), pp. 7–12, Shenyang, China, June 2014.

[12] J. Yin, Q. Yang, and I. J. Pan, “Sensor-based abnormal human-activity detection,” IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 8, pp. 1082–1090, 2008.

[13] B. X. Nie, C. Xiong, and S.-C. Zhu, “Joint action recognition and pose estimation from videos,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, pp. 1293–1301, Boston, Mass, USA, June 2015.

[14] G. Evangelidis, G. Singh, and R.Horaud, “Skeletal quads: Human action recognition using joint quadruples,” in Proceedings of the 22nd International Conference on Pattern Recognition, ICPAR 2014, pp. 4513–4518, Sweden, August 2014.

[15] E. Ohn-Bar and M. M. Trivedi, “Joint angles similarities and HOG2 for action recognition,” in Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2013, pp. 465–470, USA, June 2013.

[16] R. Vemulpalli, F. Arrate, and R. Chellappa, “Human action recognition by representing 3D skeletons as points in a lie group,” in Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR ’14), pp. 588–595, Columbus, Ohio, USA, June 2014.

[17] V. Parameswaran and R. Chellappa, “View invariance for human action recognition,” International Journal of Computer Vision, vol. 66, no. 1, pp. 83–101, 2006.

[18] G. Lu, Y. Zhou, X. Li, and M. Kudo, “Efficient action recognition via local position offset of 3D skeletal body joints,” Multimedia Tools and Applications, vol. 75, no. 6, pp. 3479–3494, 2016.

[19] A. Tejero-de-Pablos, Y. Nakashima, N. Yokoya, F.-J. Díaz-Pernas, and M. Martínez-Zarruela, “Flexible human action recognition in depth video sequences using masked joint trajectories,” Eurasip Journal on Image and Video Processing, vol. 2016, no. 1, pp. 1–12, 2016.

[20] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, “A Human activity recognition system using skeletal data from RGBD sensors,” Computational Intelligence and Neuroscience, vol. 2016, Article ID 4351435, 2016.

[21] P. Peursum, H. H. Bui, S. Venkatesh, and G. West, “Robust recognition and segmentation of human actions using HMMs with missing observations,” EURASIP Journal on Applied Signal Processing, vol. 2005, no. 13, pp. 2101–2126, 2005.

[22] D. Weinland, R. Ronfard, and E. Boyer, “Free viewpoint action recognition using motion history volumes,” Journal of Computer Vision and Image Understanding, vol. 104, no. 2-3, pp. 249–257, 2006.

[23] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, “Dense trajectories and motion boundary descriptors for action recognition,” International Journal of Computer Vision, vol. 103, no. 1, pp. 60–79, 2013.

[24] P. Matikainen, M. Hebert, and R. Sukthankar, “Trajectories: Action recognition through the motion analysis of tracked features,” in Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 2009, pp. 514–521, Japan, October 2009.

[25] M. Jain, H. Jegou, and P. Bouthemy, “Better exploiting motion for better action recognition,” in Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, pp. 2555–2562, Portland, OR, USA, June 2013.

[26] S. Zhu and L. Xia, “Human action recognition based on fusion features extraction of adaptive background subtraction and optical flow model,” Journal of Mathematical Problems in Engineering, vol. 2015, pp. 1–11, 2015.

[27] D.-M. Tsai, W.-Y. Chiu, and M.-H. Lee, “Optical flow-motion history image (OF-MHI) for action recognition,” Journal of Signal, Image and Video Processing, vol. 9, no. 8, pp. 1897–1906, 2015.

[28] L. Wang, Y. Qiao, and X. Tang, “MoFAP: a multi-level representation for action recognition,” International Journal of Computer Mathematics, vol. 119, no. 3, pp. 254–271, 2016.

[29] W. Kim, J. Lee, M. Kim, D. Oh, and C. Kim, “Human action recognition using ordinal measure of accumulated motion,” EURASIP Journal on Advances in Signal Processing, vol. 2010, Article ID 219190, 2010.

[30] W. Zhang, Y. Zhang, C. Gao, and J. Zhou, “Action recognition by joint spatial-temporal motion feature,” Journal of Applied Mathematics, vol. 2013, pp. 1–9, 2013.

[31] H.-B. Tu, L.-M. Xia, and Z.-W. Wang, “The complex action recognition via the correlated topic model,” The Scientific World Journal, vol. 2014, Article ID 810185, 10 pages, 2014.

[32] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp. 2247–2253, 2007.

[33] A. Yilmaz and M. Shah, “A differential geometric approach to representing the human actions,” Computer Vision and Image Understanding, vol. 109, no. 3, pp. 335–351, 2008.

[34] M. Grundmann, F. Meier, and I. Essa, “3D shape context and distance transform for action recognition,” in Proceedings of the 2008 19th International Conference on Pattern Recognition, ICPR 2008, pp. 1–4, USA, December 2008.

[35] D. Batra, T. Chen, and R. Sukthankar, “Space-time shapelets for action recognition,” in Proceedings of the 2008 IEEE Workshop on Motion and Video Computing, WMVC, pp. 1–6, USA, January 2008.

[36] S. Sadek, A. Al-Hamadi, G. Krell, and B. Michaelis, “Affine-invariant feature extraction for activity recognition,” International Scholarly Research Notices on Machine Vision, vol. 2013, pp. 1–7, 2013.

[37] C. Achar, X. Qu, A. Mokhber, and M. Milgram, “A novel approach for recognition of human actions with semi-global features,” Machine Vision and Applications, vol. 19, no. 1, pp. 27–34, 2008.

[38] L. Díaz-Más, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and R. Medina-Carnicer, “Three-dimensional action recognition using volume integrals,” Pattern Analysis and Applications, vol. 15, no. 3, pp. 289–298, 2012.

[39] K. Rapantzikos, Y. Avrithis, and S. Kollias, “Spatiotemporal features for action recognition and salient event detection,” Cognitive Computation, vol. 3, no. 1, pp. 167–184, 2011.

[40] N. Ikizler and P. Duygulu, “Histogram of oriented rectangles: A new pose descriptor for human action recognition,” Image and Vision Computing, vol. 27, no. 10, pp. 1515–1526, 2009.

[41] A. Fathi and G. Mori, “Action recognition by learning mid-level motion features,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR ’08), pp. 1–8, Anchorage, Alaska, USA, June 2008.
[42] V. Kellokumpu, G. Zhao, and M. Pietikainen, "Human activity recognition using a dynamic texture based method," in *Proceedings of the British Machine Vision Conference 2008*, pp. 885–894, Leeds, England, UK, 2008.

[43] D. Tran, A. Sorokin, and D. A. Forsyth, "Human activity recognition with metric learning," in *Proceedings of the European Conference on Computer Vision (ECCV ’08)*, Lecture Notes in Computer Science, pp. 548–561, Springer, 2008.

[44] B. Wang, Y. Liu, W. Wang, W. Xu, and M. Zhang, "Multi-scale locality-constrained spatiotemporal coding for local feature based human action recognition," *The Scientific World Journal*, vol. 2013, Article ID 405645, 11 pages, 2013.

[45] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined hierarchical compound features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 883–897, 2011.

[46] I. C. Dutta, J. R. R. Uijlings, B. Ionescu, K. Aizawa, A. G. Hauptmann, and N. Sebe, "Efficient human action recognition using histograms of motion gradients and VLAD with descriptor shape information," *Multimedia Tools and Applications*, vol. 76, no. 21, pp. 22445–22472, 2017.

[47] M. Ahad, M. Islam, and I. Jahan, "Action recognition based on binary patterns of action-history and histogram of oriented gradient," *Journal on Multimodal User Interfaces*, vol. 10, no. 4, pp. 335–344, 2016.

[48] S. Pehlivan and P. Duygulu, "A new pose-based representation for recognizing actions from multiple cameras," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 140–151, 2011.

[49] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR ’11)*, pp. 3209–3216, June 2011.

[50] N. Gklealis, N. Nikolaidis, and I. Pitas, "View independent human movement recognition from multi-view video exploiting a circular invariant posture representation," in *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, ICME 2009*, pp. 394–397, USA, July 2009.

[51] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, USA, June 2008.

[52] M. Ahmad and S. W. Lee, "HMM-based human action recognition using multiview image sequences," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR’06)*, pp. 263–266, Hong Kong, China, 2006.

[53] A. Yao, J. Gall, and L. Van Gool, "Coupled action recognition and pose estimation from multiple views," *International Journal of Computer Vision*, vol. 100, no. 1, pp. 16–37, 2012.

[54] X. Ji, Z. Ju, C. Wang, and C. Wang, "Multi-view transition HMMs based view-invariant human action recognition method," *Multimedia Tools and Applications*, vol. 75, no. 19, pp. 11847–11864, 2016.

[55] A. K. S. Kushwaha, S. Srivastava, and R. Srivastava, "Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns," *Journal of Multimedia Systems*, vol. 23, no. 4, pp. 451–467, 2017.

[56] S. Spurlock and R. Souvenir, "Dynamic view selection for multi-camera action recognition," *Machine Vision and Applications*, vol. 27, no. 1, pp. 53–63, 2016.

[57] A. A. Chaaraoui and F. Flórez-Revueltas, "A low-dimensional radial silhouette-based feature for fast human action recognition fusing multiple views," *International Scholarly Research Notices*, vol. 2014, pp. 1–11, 2014.

[58] A. Iosifidis, A. Tefas, and I. Pitas, "View-independent human action recognition based on multi-view action images and discriminant learning," in *Proceeding of the IEEE Image, Video, and Multidimensional Signal Processing Workshop 2013*, pp. 1–4, 2013.

[59] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 412–424, 2012.

[60] R. Kavi, V. Kulathumani, F. Rohit, and V. Kecojevic, "Multiview fusion for activity recognition using deep neural networks," *Journal of Electronic Imaging*, vol. 25, no. 4, Article ID 043010, 2016.

[61] Y. Kong, Z. Ding, J. Li, and Y. Fu, "Deeply learned view-invariant features for cross-view action recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 3028–3037, 2017.

[62] A.-A. Liu, N. Xu, Y.-T. Su, H. Lin, T. Hao, and Z.-X. Yang, "Single/multi-view human action recognition via regularized multi-task learning," *Journal of Neurocomputing*, vol. 151, no. 2, pp. 544–553, 2015.

[63] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning, and recognition," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*, pp. 2649–2656, USA, June 2014.

[64] P. Huang, A. Hilton, and J. Starck, "Shape similarity for 3D video sequences of people," *International Journal of Computer Vision*, vol. 89, no. 2-3, pp. 362–381, 2010.

[65] A. Veeraraghavan, A. Srivastava, A. K. Roy-Chowdhury, and R. Chellappa, "Rate-invariant recognition of humans and their activities," *IEEE Transactions on Image Processing*, vol. 18, no. 6, pp. 1326–1339, 2009.

[66] I. N. Junejo, E. Dexter, I. Laptev, and P. Pérez, "View-independent action recognition from temporal self-similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 172–185, 2011.

[67] A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 347–360, 2012.

[68] H. Rahmani and A. Mian, "3D action recognition from novel viewpoints," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 1506–1515, USA, July 2016.

[69] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Proceeding of Advanced Video-Based Surveillance Systems Springer*, pp. 135–144, 2001.

[70] P. Chawalitsittikul and N. Suvonvorn, "Profile-based human action recognition using depth information," in *Proceedings of the IASTED International Conference on Advances in Computer Science and Engineering, ACSE 2012*, pp. 376–380, Thailand, April 2012.

[71] N. Noorit, N. Suvonvorn, and M. Karnchanadecha, "Model-based human action recognition," in *Proceedings of the 2nd International Conference on Digital Image Processing, Singapore*, February 2010.

[72] M. Ahmad and S.-W. Lee, "Human action recognition using shape and CLG-motion flow from multi-view image
sequences,” *Journal of Pattern Recognition*, vol. 41, no. 7, pp. 2237–2252, 2008.

[73] C.-H. Chuang, J.-W. Hsieh, L.-W. Tsai, and K.-C. Fan, “Human action recognition using star templates and delaunay triangulation,” in *Proceedings of the 2008 4th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP 2008*, pp. 179–182, China, August 2008.

[74] G. I. Parisi, C. Weber, and S. Wermter, “Human action recognition with hierarchical growing neural gas learning,” in *Proceedings of Artificial Neural Networks and Machine Learning on Springer – ICANN 2014*, vol. 8681 of *Lecture Notes in Computer Science*, pp. 89–96, Springer International Publishing, Switzerland, 2014.

[75] N. Sawant and K. K. Biswas, “Human action recognition based on spatio-temporal features,” in *Proceedings of the Pattern Recognition and Machine Intelligence Springer*, Lecture Notes in Computer Science, pp. 357–362, Springer, Heidelberg, Berlin, 2009.

[76] N. Suwonvorn, *Prince of Songkla University (PSU) Multi-view profile-based action RGBD dataset*, 2017, http://fivedots.coe.psu.ac.th/~kom/?p=1483 (accessed 20 December 2017).

[77] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, “The i3DPost multi-view and 3D human action/interaction database,” in *Proceeding of the 6th European Conference for Visual Media Production (CVMP ’09)*, pp. 159–168, London, UK, November 2009.