Hindawi
Complexity
Volume 2018, Article ID 6271348, 12 pages
https://doi.org/10.1155/2018/6271348

WILEY | Hindawi

Research Article

Human Pose Recognition Based on Depth Image Multifeature Fusion

Haikuan Wang, Feixiang Zhou, Wenju Zhou, and Ling Chen

School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 201900, China

Correspondence should be addressed to Wenju Zhou; zhouwenju@shu.edu.cn

Received 13 July 2018; Accepted 9 September 2018; Published 2 December 2018

Guest Editor: Liang Hu

The recognition of human pose based on machine vision usually results in a low recognition rate, low robustness, and low operating efficiency. That is mainly caused by the complexity of the background, as well as the diversity of human pose, occlusion, and self-occlusion. To solve this problem, a feature extraction method combining directional gradient of depth feature (DGoD) and local difference of depth feature (LDoD) is proposed in this paper, which uses a novel strategy that incorporates eight neighborhood points around a pixel for mutual comparison to calculate the difference between the pixels. A new data set is then established to train the random forest classifier, and a random forest two-way voting mechanism is adopted to classify the pixels on different parts of the human body depth image. Finally, the gravity center of each part is calculated and a reasonable point is selected as the joint to extract human skeleton. The experimental results show that the robustness and accuracy are significantly improved, associated with a competitive operating efficiency by evaluating our approach with the proposed data set.

1. Introduction

Human perception to the external world is mainly through the sense organs such as sight, touch, hearing, and smell, of which about 80% of information is obtained by the vision. It is important for the next generation intelligent computers to mount visual functions on computers so that they can automatically recognize and analyze the activities of people in the surrounding environment [1–3].

At present, pose and action recognition is widely used in the many fields like advanced human-computer interaction, intelligent monitoring system, motion analysis, and medical rehabilitation [4–6]. Pose recognition is a challenging research in motion analysis. The core target is to infer the posture parameters from the image sequence on various parts of the human body, such as the actual position in the three-dimensional space or the angle between the various joints. Human body motion can be reconstructed in three-dimensional space through posture parameters mentioned above. At present, the human pose recognition algorithms based on machine vision are mainly divided into two categories: one is based on traditional RGB images and the other is based on depth images. The biggest difference between them is that pixels in the RGB image record the color information of the object, while pixels in the depth image record the distance between the object and the camera. Human pose recognition based on RGB images mainly utilizes the apparent features on the image, such as HOG (histogram of oriented gradient) features [7] and contour features [8]. However, these methods are usually affected by the external environment and are particularly vulnerable to the light, resulting in low detection accuracy. In addition, due to the large differences in the size of the human body, these algorithms are only suitable for the limited environments and people. In recent years, with the development of depth sensors, especially the Kinect developed by Microsoft which has color and depth information (RGB-D), the recognition rate of human pose has been greatly improved compared with ordinary sensors [9–13]. The main reason is that the depth images have many advantages over the RGB images.

First, depth images have robustness against changing of color and illumination. Also, the depth image, which is 3D, has more information than the RGB image. Human pose recognition methods can be divided into two categories: model-based method and feature learning. In model-based human pose detection, the human body is divided into multiple components which are combined into a model and then the human pose is estimated by inverting the kinematics or solving optimization problems. Pishchulin et al. proposed a new articulated posture model based on image morphology [14]. Sun and Savarese proposed APM (articulated part-based model) based on the joint detection [15], and Sharma et al. proposed an EPM (expanded parts' model) based on a collection of body parts [16]. Siddiqui and Medioni used a Markov chain Monte Carlo (MCMC) framework with head, hand, and forearm detectors to estimate the body [17].

Feature learning tries to get advanced features from depth images through analyzing each pixel, and uses various machine learning algorithms to realize human pose recognition [12, 18–23]. Shotton et al. proposed two different methods for estimating human body poses [18]. One of the methods uses a random forest to classify each pixel in the depth image. Another method predicts the position of a human joint. Both methods are based on random forest classifiers that train a large number of synthetic and real human depth images. Hernández-Vela et al. proposed graph cuts' optimization based on Shotton's method [24]. Kim et al. proposed another human pose estimation method based on SVM (support vector machine) and superpixel [25]. In addition, deep learning algorithms are also used to solve the pose estimation of the target [26–28], and the convolution neural network (CNN) is used for large-scale data set processing [29–32].

In general, the advantage of model-based human pose recognition is that there is no need to build a large data set; instead, it only establishes some models. It has a higher recognition rate for the pose as the model matched. However, this method also has some disadvantages. For example, it is difficult to construct complex human body models mainly because of the diversity of human postures in actual situations.

The main merit of feature learning is that it does not need to establish a complex human body model, so it is not restricted to the model and can be applied to various situations. However, this method also has disadvantages. On one hand, it has to build a huge data set to fit in different environments. On the other hand, many feature extraction methods have high complexity and cannot meet real-time requirements. Therefore, a human pose recognition method based on depth image multifeature fusion is proposed in this paper. First, the body parts were encoded with number in depth images and a data set was constructed. Afterwards, the LDoD and DGoD features are extracted for training to get a random forest classifier. Finally, the gravity center is calculated and possible joints are screened out. The LDoD and DGoD have lower computational complexity than other algorithms, so they can satisfy the real-time requirement. Moreover, the recognition rate of human pose improves by combining LDoD and DGoD.

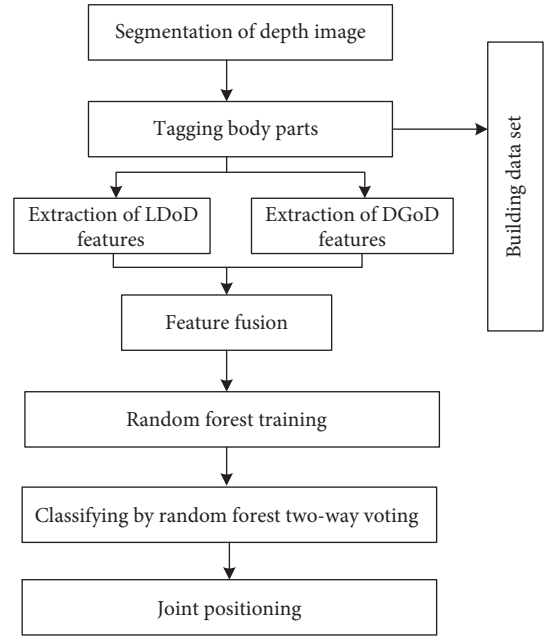

Figure 1: Flow chart of human pose recognition based on depth image multifeature fusion.

The rest of this paper is organized as follows: Section 2 introduces the algorithm flow about depth image multifeature fusion for recognizing human pose. Section 3 details each step of pose recognition and related algorithms. Section 4 constructs a random forest classifier. Section 5 describes the positioning of the joints in the human body image. Section 6 analyzes the experimental results. Section 7 is the conclusion.

2. Algorithm Overview

The flowchart of human pose recognition based on depth image multifeature fusion is shown in Figure 1. Firstly, the original depth image is segmented to extract the human target so that the different parts of segmented body is easily tagged with a specific code. And then LDoD and DGoD features are extracted for training multiple decision trees to obtain a random forest classifier. The classifier is used to classify the body parts of the test samples. Finally, the position of the joints in the human body image is calculated.

3. Human Pose Recognition

3.1. Depth Image Segmentation. In image processing, we often focus on special areas which are called ROI (region of interest) [33–35]. Usually, these areas have rich feature information. Therefore, in order to identify and analyze the target, the area, where the target is located, needs to be separated from the background. On this basis, feature extraction and human body recognition can be performed.

Due to the fact that the actual scene is fixed in this paper, depth background difference is used to segment the human body. The depth value is quantized to a grayscale space of 0–255, that is, little number is corresponding to large depth. Therefore, the 3D image can be displayed as a 2D gray image, where pixel value represents a different meaning from the conventional RGB image.

Because the camera shoots downwards from head top, the leg information of the human body cannot be considered. The depth range is controlled between D_{near} and D_{far}, and it can be expressed as $[D_{\text{near}}, D_{\text{far}}]$. First, Gaussian filtering is performed on the original depth data to filter out noise and suppress the drift of depth data. Then, the original depth image is subtracted from the background image, and the foreground target is extracted according to the threshold T, shown as follows:

$$\begin{aligned} D(x,y) &= |I(x,y) - B(x,y)|, \\ T(x,y) &= \begin{cases} 1, & D(x,y) \geq T, \\ 0, & D(x,y) < T, \end{cases} \end{aligned} \tag{1}$$

where $B(x,y)$ is the background image, $I(x,y)$ is the original image, and $T(\mathrm{x},y)$ is the binary image. Then, the depth image of the corresponding area is extracted, shown as follows:

$$S(x,y) = \begin{cases} I(x,y), & D(x,y) \geq T, \\ 0, & D(x,y) < T, \end{cases} \tag{2}$$

where $S(x,y)$ is the effective depth area and $S(x,y) \subseteq [D_{\text{near}}, D_{\text{far}}]$.

3.2. Tagging Body Parts. Since there is no standard human pose depth image library, we builds a data set, including common human actions such as running, jumping, lifting, bending, knee flexion, and interaction. Random forest learning algorithm belongs to supervised learning; the data samples are a known category, and these samples need to be tagged [36–39]. The tagging method is to divide the human body into 11 parts, and the rest is the background; the approximate position of each part of the human body in the depth image is observed, and then, the position is tagged with the corresponding color. As shown in Figure 2, the valid points inside the rectangle of the head area are all marked in red.

The tagging result is shown in Table 1. This paper divides the waist above the human body into the head, the left shoulder, the right shoulder, the left upper arm, the right upper arm, the left lower arm, the lower right arm, the left hand, the right hand, the left body, the right body, and the background.

3.3. LDoD Feature Extraction. According to the depth image of the human body that has been manually tagged, the features of 12 parts need to be extracted. This paper uses the local difference feature as a feature representation of the pixel, which can reflect the neighborhood information of the pixel. It uses the difference between two pixels among the eight neighborhood points to represent the characteristics of the pixel. The location of the eight neighborhood pixels is shown in Figure 3.

LDoD feature can be represented as

$$T_{i,j}(I,p) = \left| d_s(p_i) - d_s\left(p_j\right) \right|, \tag{3}$$

where $i, j \in \{1,2,3,4,5,6,7,8\}$, $i \neq j$, and $d_s(p_i)$ is the depth value of p_i.

Assuming $\theta = \{\theta_1, \theta_2, \theta_3, \ldots, \theta_{28}\}$, $T_{\theta_k}(S,p)$ is replaced by $T_{i,j}(S,p)$ and $k \in \{1,2,3,\ldots,28\}$. The feature vector of a point can be expressed as

$$T_\theta(I,p) = \left\{ T_{\theta_1}(I,p), T_{\theta_2}(I,p), \ldots, T_{\theta_{28}}(I,p) \right\}. \tag{4}$$

According to LDoD feature, features of the same type of pixels are mostly similar and features of different pixel types have large differences. Therefore, for various parts of the human body, this feature has a good division degree. Figure 4(a) shows the divided depth image, and Figure 4(b) is an enlarged image of the left lower arm. As can be seen from the figures, pixel p_6 and pixel p_7 are in the body area and pixel p_4 is out of body area and its value is 0. Therefore, the value of $T_{6,7}(S,p)$ is smaller and $T_{4,7}(S,p)$ is larger. Figure 4(c) is an enlarged image of the right lower arm. The value of $T_{6,7}(S,p)$ is larger, and the value of $T_{4,7}(S,p)$ is smaller. Therefore, these two values can distinguish the left and right lower arms of the human body.

The computational complexity of this feature is very low. Formula (3) only uses subtraction between values. In addition, it also has space translation and rotation invariance and can be applied to people's changes in postures.

3.4. DGoD Feature Extraction. Due to that, the depth information represents the distance between the object and the depth camera; the angle relationship between the plane where the pixel is located and the plane where the depth camera is located can be obtained by simply calculating the arctangent of the gradient, that is, the DGoD feature, which can be calculated as follows:

$$\begin{aligned} \mathrm{DGoD}_{S(x,y)} &= \tan^{-1}\frac{dy}{dx}, \\ dy &= S(x,y+i) - S(x,y-i), \\ dx &= S(x+i,y) - S(x-i,y), \end{aligned} \tag{5}$$

where $i = 1, 2, 3$. Three DGoD features were selected, which are represented as $G_{\theta 1}(S,x,y)$, $G_{\theta 2}(S,x,y)$, and $G_{\theta 3}(S,x,y)$.

The range of directional gradients is $[0^\circ, 360^\circ]$. When the pixel points are on the same plane, the direction gradients are

(a) (b)

Figure 2: (a) Original depth images. (b) Tagged depth images.

Table 1: The tagging result.

Body part no.	12 parts	Tagged colors
1	Background	Black
2	Head	Red
3	Left shoulder	Orange
4	Right shoulder	Yellow
5	Left upper arm	Light blue
6	Right upper arm	Blue
7	Left lower arm	Brown
8	Right lower arm	Light brown
9	Left hand	Dark green
10	Right hand	Light green
11	Left body	Green
12	Right body	Purple

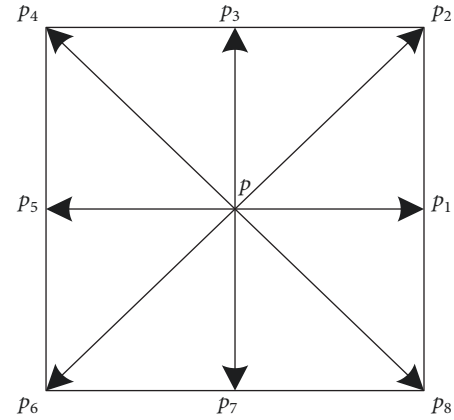

Figure 3: Neighborhood distribution of a pixel.

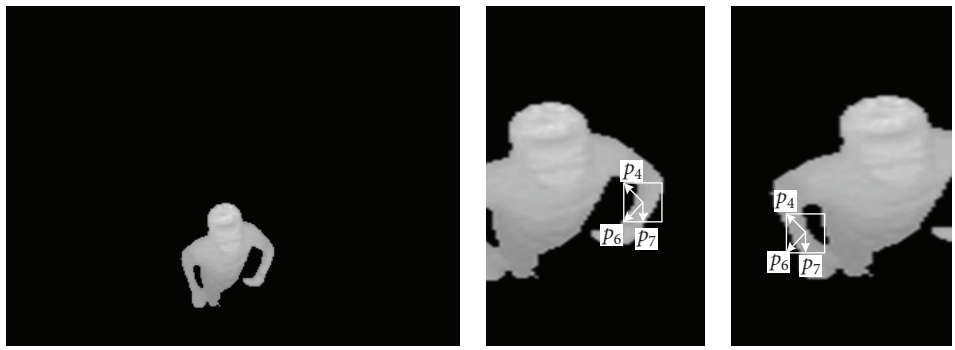

Figure 4: (a) Depth image of human body. (b) LDoD characteristic component of the left lower arm. (c) LDoD characteristic component of the right lower arm.

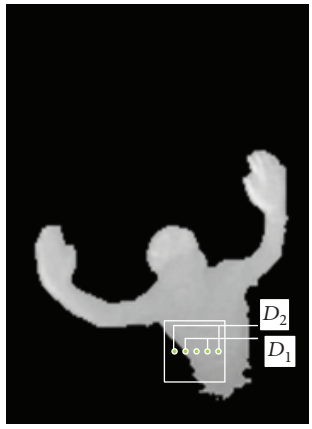

Figure 5: The diagram of DGoD feature.

similar in size; otherwise, the direction gradients are quite different. The diagram of DGoD feature is shown in Figure 5. The green dots from left to right are p_1, p_2, p, p_3, and p_4. It can be seen from formula (5) that the value of D_1 is smaller and the value of D_2 is larger, which means that p_2 and p_3 are in the same plane, while p_1 and p_4 are in different planes.

4. The Design of the Random Forest Classifier

4.1. Random Forest Model Construction. Decision tree is one of the most widely used inductive inference algorithms at present. Its generated rules are simple and easy to understand. Pixels of depth images can be classified quickly and efficiently by decision tree, so it can be widely used in target detection and recognition. However, a single decision tree can easily lead to overfitting causing wrong classification. The random forest is composed of multiple decision trees [40, 41], and the decision tree is trained with different data sets and parameters, which cannot only reduce the degree of overfitting but also the classification accuracy can be improved because its output is voted by multiple decision trees.

The classification effect of random forest classifiers is affected by many factors, including the size of the training set D, the dimension of the sample feature vector N, the number of the decision tree K, the maximum depth of each tree d, the eigenvector dimension n, and the termination condition for growth of each tree.

In the previous sections, the human body was divided into 12 different parts and then LDoD and DGoD features were extracted as the input of the random forest classifier. All of preliminary works are prepared for the design of the classifier model. The set of attributes can be represented as

$$S = \left\{ T_{\theta_1}(I,p), T_{\theta_2}(I,p), \ldots, T_{\theta_{28}}(I,p), \mathrm{DGoD}_{S(x,y)} \right\}. \quad (6)$$

ID3 decision tree algorithm is used to train each decision tree in a random forest. Training sample set can be expressed as

$$Q = \{P, C\}, \quad (7)$$

where $P = \{p_1, p_2, p_3, p_4, \ldots, p_i, \ldots, p_n\}$ is a set of training pixels and $C = \{c_1, c_1, c_1, \ldots, c_{12}\}$ is a collection of categories to which a pixel belongs, that is, 12 parts of the human body.

The set of parameters can be expressed as

$$\varphi = (\theta, \tau_1, \tau_2), \quad (8)$$

where θ is the attribute parameter and τ_1 and τ_2 are the thresholds.

The flow chart of the construction of a single decision tree is shown in Figure 6. First, putting back is adopted in the extraction method and the training set D_i, which is the same size of D, is extracted from D to get K subsets. Then, a tree node is created, and if it reaches the termination condition, the process is stopped and the current node is set as a leaf node. Otherwise, n features is extracted from the N

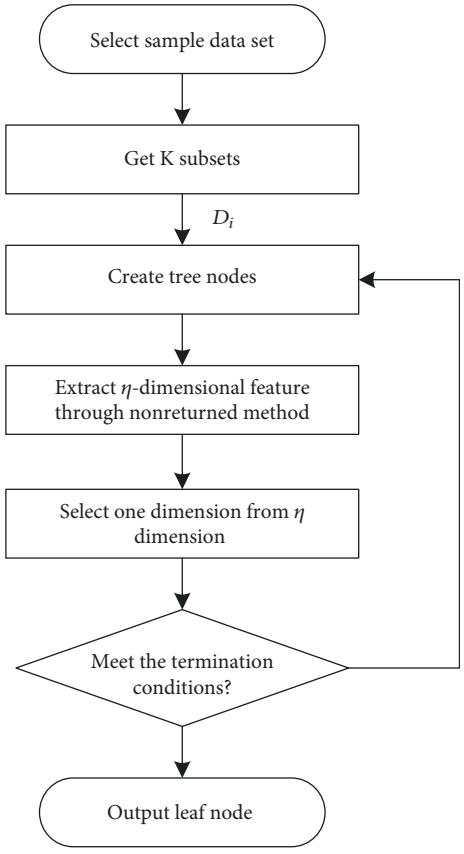

Figure 6: Flow chart of the construction of a single decision tree.

-dimensional feature set using a fixed-scale and nonreturned extraction method. The one-dimensional feature is determined according to the metric of the feature attribute, and the current node is split into the left subset $D_{\mathrm{l}}(\varphi)$ and the right subset $D_{\mathrm{r}}(\varphi)$:

$$\begin{cases} p_i \in D_{\mathrm{l}}(\varphi), & T_\theta \le \tau_1, G_\theta \le \tau_2, \\ p_i \in D_{\mathrm{r}}(\varphi), & T_\theta > \tau_1, G_\theta > \tau_2. \end{cases} \tag{9}$$

The information gain is used to select the partitioning property of the decision tree, which can be calculated by as follows:

$$\begin{aligned} \varphi_* &= \arg\max \mathrm{Gain}(D, \varphi), \\ \mathrm{Gain}(D, \varphi) &= \mathrm{Ent}(D) - \sum_{v \in \{\mathrm{l},\mathrm{r}\}} \frac{|D^v(\varphi)|}{|D|} \mathrm{Ent}(D^v(\varphi)), \end{aligned} \tag{10}$$

where $\mathrm{Ent}(D)$ is information entropy.

4.2. Random Forest Two-Way Voting. In the traditional random forest classification [42–44], the sample is judged by every decision tree and voted by every tree. Every tree has equal decision right. The random forest two-way voting mechanism is adopted with different decision rights in this paper. Data set is divided into in-of-bag and out-of-bag data. The data subset is called in-of-bag data when it is used to build a random forest. Otherwise, a data subset is called out-of-bag. The decision right of a tree is gained according to the result of testing out-of-bag data. When the result is true, then the tree is voted. If a decision tree has more votes, the weight will be higher. The basic algorithm steps for two-way voting are as follows.

Step 1. Create K decision trees. And in-of-bag data and out-of-bag data are generated for every tree.

Step 2. Perform a performance evaluation. That is, a tree is evaluated by a certain amount of out-of-bag data. If the decision tree's classification result is true, the tree is voted.

Step 3. Assign the total number of votes to the decision tree as weight and normalize the weights of all decision trees.

Step 4. Input the test sample to the sorted random forest model, and the obtained classification result multiplies the weight to get the final classification result, shown as follows:

$$R(x) = \sum_{i=1}^{K} T_i r_i(x), \tag{11}$$

where $R(x)$ is the final classification result, T_i is the weight coefficient corresponding to i-th decision tree, and $T_1 + T_2 + T_3 + \cdots + T_k = 1$. $r_i(x)$ is the evaluation result of i-th decision tree.

5. Human Joint Positioning

Determining the human body joints is the final step in human pose recognition [45]. The above chapters have used the random forest classifier to classify 12 parts in the human body image, but the joint position has not still been determined. The joints will be determined by calculating the gravity center of the 12 body parts in this paper.

For the depth image with size $(M \times N)$, the $p + q$ moment m_{pq} and central moment at the pixel $I(x, \mathrm{y})$ can be calculated by formula (12) and (13), respectively.

$$m_{pq} = \sum_{y=1}^{N} \sum_{x=1}^{M} x^p y^q \cdot I(x, y), \tag{12}$$

$$\mu_{pq} = \sum_{y=1}^{N} \sum_{x=1}^{M} (x - x_{\mathrm{c}})^p \cdot (y - y_{\mathrm{c}})^q \cdot I(x, y), \tag{13}$$

where $(x_{\mathrm{c}}, y_{\mathrm{c}})$ is the gravity center, which can be calculated by formula (14) and (15).

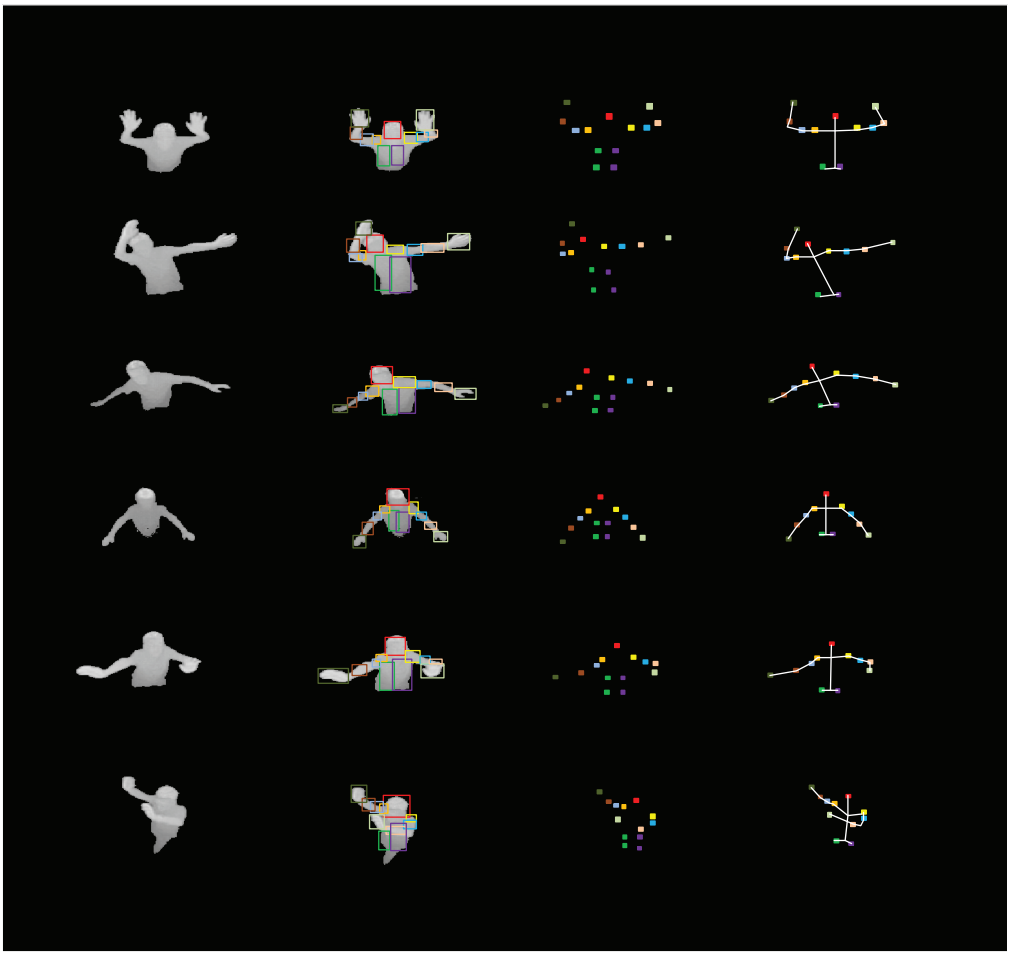

Figure 7: Results of human body part recognition and joint point positioning in different postures.

The gravity center of the upper arm and the gravity center of the lower arm are calculated to obtain the joint of the left elbow or the right elbow, as shown in formula (16).

$$\begin{aligned} x_c &= \frac{m_{10}}{m_{00}}, \\ y_c &= \frac{m_{01}}{m_{00}}, \end{aligned} \tag{14}$$

$$\begin{aligned} x_c &= \frac{\sum_{y=1}^{N}\sum_{x=1}^{M} x \cdot I(x,y)}{\sum_{y=1}^{N}\sum_{x=1}^{M} I(x,y)}, \\ y_c &= \frac{\sum_{y=1}^{N}\sum_{x=1}^{M} y \cdot I(x,y)}{\sum_{y=1}^{N}\sum_{x=1}^{M} I(x,y)}, \end{aligned} \tag{15}$$

$$\begin{aligned} x_{\text{elbow}} &= \frac{\left(\sum_{y=1}^{N_{\text{up}}}\sum_{x=1}^{M_{\text{up}}} x \cdot I(x,y) / \sum_{y=1}^{N_{\text{up}}}\sum_{x=1}^{M_{\text{up}}} I(x,y)\right) + \left(\sum_{y=1}^{N_{\text{low}}}\sum_{x=1}^{M_{\text{low}}} x \cdot I(x,y) / \sum_{y=1}^{N_{\text{low}}}\sum_{x=1}^{M_{\text{low}}} I(x,y)\right)}{2} + \Delta x, \\ y_{\text{elbow}} &= \frac{\left(\sum_{y=1}^{N_{\text{up}}}\sum_{x=1}^{M_{\text{up}}} y \cdot I(x,y) / \sum_{y=1}^{N_{\text{up}}}\sum_{x=1}^{M_{\text{up}}} I(x,y)\right) + \left(\sum_{y=1}^{N_{\text{low}}}\sum_{x=1}^{M_{\text{low}}} y \cdot I(x,y) / \sum_{y=1}^{N_{\text{low}}}\sum_{x=1}^{M_{\text{low}}} I(x,y)\right)}{2} + \Delta y, \end{aligned} \tag{16}$$

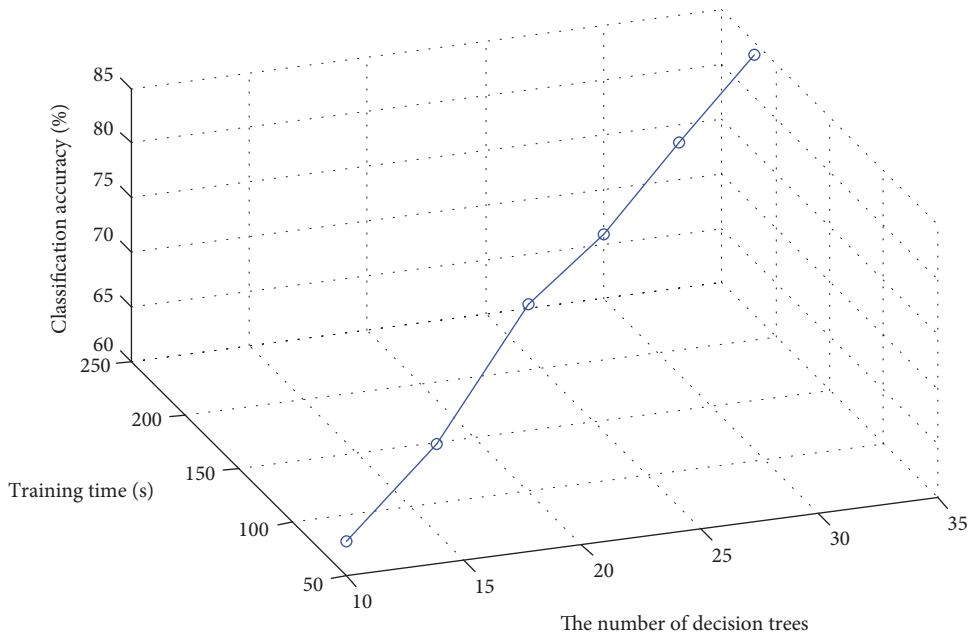

Figure 8: Influence of the number of decision trees on the experimental results.

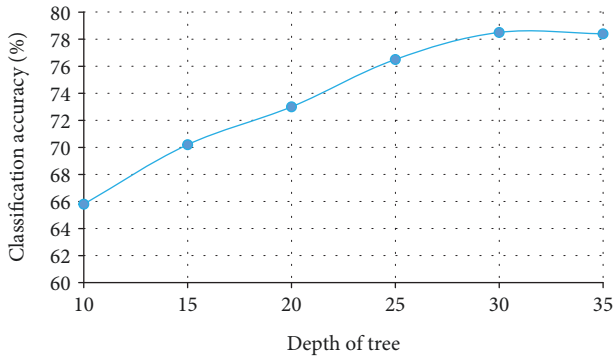

Figure 9: Influence of decision tree depth on experimental results.

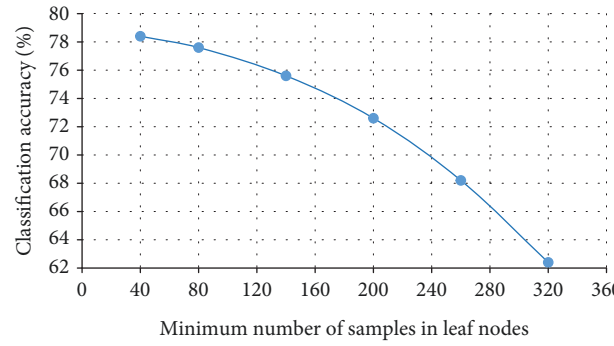

Figure 10: Influence of the minimum number of samples in the leaf nodes on the experimental results.

where the size of area of the upper arm is $M_{up} \times N_{up}$, the size of the area of the lower arm is $M_{low} \times N_{low}$, and Δx and Δy are the offsets.

6. Experimental Results and Analysis

In this paper, 1000 depth images are used for the training of the classifier model and 100 images are used for the test, including the poses of 10 different people. The algorithm is programmed in C++ and compiled in Visual Studio 2013. The test computer uses an Intel Core i5-4570 processor clocked at 3.20 GHz. The ToF (time of flight) depth camera is used with the resolution of 320×240 in this paper.

6.1. Qualitative Analysis. The results of human body part recognition and joint positioning with 6 postures are shown in Figure 7. The first column is the segmented depth images, the second column is the outputs of the random forest classifier, the third column is the gravity center of each part, and the last column is the skeletons composed of the joints. As can be seen from Figure 7, the random forest classifier can correctly classify most of the pixels in the human body image, such as the body and the head. Incorrect classification often happens at the intersection of the two parts. Fortunately, the joints are almost positioned accurately and a reasonable human skeleton can be obtained. Finally, in the sixth picture, one of the hands blocks the body, and according to the positioning result, the posture based on the fusion of DGoD and LDoD features proposed in this paper can solve the mutual fusion and occlusion.

6.2. Quantitative Analysis

6.2.1. Comparison of experimental results of different parameters of classifiers. When constructing a random forest model, the number of decision trees K, the maximum depth of numbers d, and the minimum number of samples in leaf nodes N_{node} can affect the classifier performance. The experiment first determines the optimal classifier parameters by training five sample images. Figures 8–10 compare

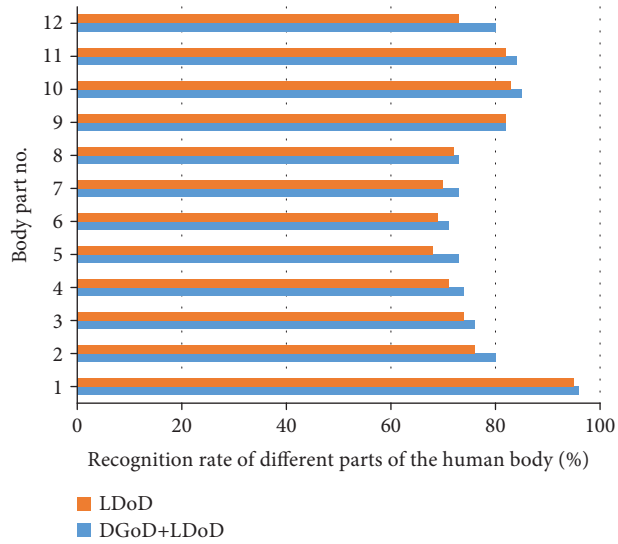

Figure 11: Recognition rates of different parts under different algorithms.

the classification accuracy and the training time of the algorithm with different parameters.

From Figure 8, we can see that with the other parameters fixed, as the number of decision tree K increases, the training time consumed and the accuracy of the classification show an increasing trend. When d is 20, the classification accuracy of the test sample reaches 77.2% and the training time is 100 s. When d is 25, the correct rate of classification only increased by 1% but the required training time is increased to 140 s. Therefore, the optimal K in this paper is 20.

As shown in Figure 9, when the other parameters are fixed, the greater the depth of the tree d is, the higher the accuracy is. When the value of d is 30, the correct rate reaches its maximum and then the correct rate is almost constant with the increase of the depth. So the optimal number of depth is 30.

The minimum number of samples in the leaf nodes can be used as the termination condition for the growth of the decision tree. When it is too large, the structure of the tree will stop prematurely, which will affect the classification accuracy. When it is too small, the structure of the tree will become more complicated and will consume too much time. In Figure 10, when other parameters are fixed, the classification accuracy of the test sample reaches 78.4% when $N_{\text{node}} = 40$. When $N_{\text{node}} = 80$, the test sample classification accuracy rate drops to 77.6%. So in this paper, $N_{\text{node}} = 40$.

6.2.2. Comparison of the recognition rate of various algorithms. This paper compares the recognition rate of each part with a single feature LDoD and the recognition rate with the combination of two features of DGoD and LDoD, as shown in Figure 11. The recognition rate of the random forest algorithm with multifeature fusion is obviously improved, reaching about 80%. Among these 12 parts, the recognition rates of the left and right arms are lower, mainly because of the complex movements of the upper limbs. In addition, as the amount of samples collected increases, the recognition rate will increase.

The traditional voting mechanism of random forests and the two-way voting mechanism are compared in this paper, as shown in Figure 12. It can be seen from the figure that the classification accuracy of the random forest two-way voting mechanism is significantly higher than the traditional one-way voting mechanism.

Finally, we also compare our algorithm with the popular algorithms in other literatures, as shown in Table 2; our classification method is superior to that of Shotton and Kim. In addition, the computation time is about 54.9% of Shotton's algorithm. Therefore, the proposed method is more suitable for high real-time and high recognition rate occasions.

7. Conclusion

In this work, we propose a human pose recognition algorithm based on the fusion of LDoD and DGoD features. In human pose recognition, we first establish our own sample data set including depth images with a specific code. Then, we extract the LDoD and DGoD features from the sample. It is simple to calculate the above two features. Thus, the computation is greatly reduced. In the next step, these two features are used to train the random forest classifier. In order to improve the accuracy of classification, a random forest two-way voting mechanism is used to detect and classify different parts of the human body. Finally, according to the classification results, the gravity center of different body parts is calculated, so that accurate joints and skeleton can be obtained.

The experimental results show that the random forest classifier has higher classification accuracy and robustness. In addition, our method has low computation cost compared with the other methods and meets the real-time requirements. However, no method is perfect in terms of human

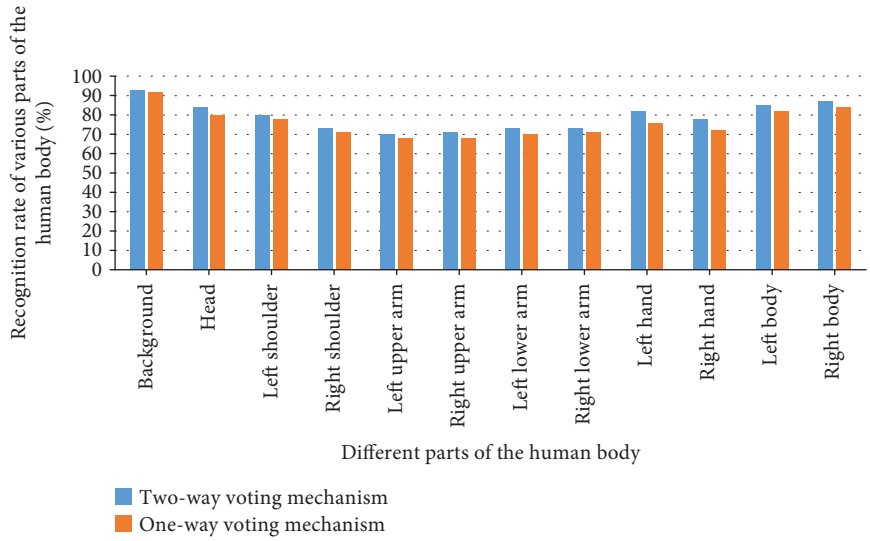

Figure 12: Comparison of the accuracy of a one-way voting mechanism and a two-way voting mechanism in random forests.

Table 2: Comparison of classification accuracy and efficiency of different algorithms.

Algorithm	Average body part classification accuracy (%)	Computational cost for each frame (GFlops)
Shotton's algorithm	70.2	1.44
Kim's algorithm	78.3	0.81
Our algorithm	80.7	0.79

body pose recognition, so it is necessary to research the following aspects.

(i) Extracting better features for body part recognition and classification

(ii) Using other classification algorithms or classifiers for body part recognition and classification, such as some efficient deep learning methods

(iii) Studying body part recognition with more complex human poses, such as lying on the ground

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by National Science Foundation of China (nos. 61877065 and 61473182).

References

[1] V. Gatteschi, F. Lamberti, P. Montuschi, and A. Sanna, "Semantics-based intelligent human-computer interaction," *IEEE Intelligent Systems*, vol. 31, no. 4, pp. 11–21, 2016.

[2] Y. Cao, D. Barrett, A. Barbu et al., "Recognize human activities from partially observed videos," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2658–2665, Portland, Oregon, June 2013.

[3] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

[4] G. Plouffe and A.-M. Cretu, "Static and dynamic hand gesture recognition in depth data using dynamic time warping," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 2, pp. 305–316, 2016.

[5] K. Kurita and S. Ueta, "A new motion control method for bipedal robot based on non-contact and non-attached human motion sensing technique," *IEEE Transactions on Industry Applications*, vol. 47, no. 2, pp. 1022–1027, 2011.

[6] W. Wang, X. Chen, G. Zhang et al., "Precision security: integrating video surveillance with surrounding environment changes," *Complexity*, vol. 2018, Article ID 2959030, 10 pages, 2018.

[7] J. Konečný and M. Hagara, "One-shot-learning gesture recognition using HOG-HOF features," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2513–2532, 2017.

[8] C. Rasche, "Rapid contour detection for image classification," *IET Image Processing*, vol. 12, no. 4, pp. 532–538, 2018.

[9] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2659–2668, Venice, Italy, October 2017.

[10] D. Michel, C. Panagiotakis, and A. A. Argyros, "Tracking the articulated motion of the human body with two RGBD cameras," *Machine Vision and Applications*, vol. 26, no. 1, pp. 41–54, 2015.

[11] F. Moreno-Noguer, "3D human pose estimation from a single image via distance matrix regression," in *2017 IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1561–1570, Honolulu, Hawaii, July 2017.

[12] J. S. Supančič, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan, "Depth-based hand pose estimation: methods, data, and challenges," *International Journal of Computer Vision*, vol. 126, no. 11, pp. 1180–1198, 2018.

[13] K. Nishi and J. Miura, "Generation of human depth images with body part labels for complex human pose recognition," *Pattern Recognition*, vol. 71, pp. 402–413, 2017.

[14] L. Pishchulin, A. Jain, M. Andriluka, T. Thormahlen, and B. Schiele, "Articulated people detection and pose estimation: reshaping the future," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3178–3185, Providence, Rhode Island, June 2012.

[15] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in *2011 International Conference on Computer Vision*, pp. 723–730, Barcelona, Spain, November 2011.

[16] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for semantic description of humans in still images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 87–101, 2017.

[17] M. Siddiqui and G. Medioni, "Human pose estimation from a single view point, real-time range sensor," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 1–8, San Francisco, CA, USA, June 2010.

[18] J. Shotton, T. Sharp, A. Kipman et al., "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[19] A. Nanjappa, L. Cheng, W. Gao, C. Xu, A. Claridgechang, and Z. Bichler, *Mouse Pose Estimation from Depth Images*, Computer Science, 2015.

[20] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[21] A. Toshev and C. Szegedy, "DeepPose: human pose estimation via deep neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, Columbus, Ohio, June 2014.

[22] J. Shotton, R. Girshick, A. Fitzgibbon et al., "Efficient human pose estimation from single depth images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2821–2840, 2013.

[23] Y. Liu, Z. Lu, J. Li, C. Yao, and Y. Deng, "Transferable feature representation for visible-to-infrared cross-dataset human action recognition," *Complexity*, vol. 2018, Article ID 5345241, 20 pages, 2018.

[24] A. Hernández-Vela, N. Zlateva, A. Marinov et al., "Graph cuts optimization for multi-limb human segmentation in depth maps," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 726–732, Providence, Rhode Island, June 2012.

[25] H. Kim, S. Lee, D. Lee, S. Choi, J. Ju, and H. Myung, "Real-time human pose estimation and gesture recognition from depth images using superpixels and SVM classifier," *Sensors*, vol. 15, no. 6, pp. 12410–12427, 2015.

[26] G. L. Oliveira, A. Valada, C. Bollen, W. Burgard, and T. Brox, "Deep learning for human part discovery in images," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1634–1641, Stockholm, Sweden, May 2016.

[27] Z. Yan, Y. Zhan, Z. Peng et al., "Multi-instance deep learning: discover discriminative local anatomies for bodypart recognition," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1332–1343, 2016.

[28] l.-M. Ramanathan, W. Y. Yau, and E. K. Teoh, "Improving human body part detection using deep learning and motion consistency," in *2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pp. 1–5, Phuket, Thailand, November 2016.

[29] H. Su, C. R. Qi, Y. Li, and L. Guibas, "Render for CNN: viewpoint estimation in images using CNNs trained with rendered 3D model views," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2686–2694, December 2015.

[30] F. Massa, M. Aubry, and R. Marlet, "Convolutional neural networks for joint object detection and pose estimation: a comparative study," *Computer Science*, vol. 19, no. 2a, pp. 412–417, 2015.

[31] C. S. Chin, J. T. Si, A. S. Clare, and M. Ma, "Intelligent image recognition system for marine fouling using Softmax transfer learning and deep convolutional neural networks," *Complexity*, vol. 2017, 9 pages, 2017.

[32] P. Witoonchart and P. Chongstitvatana, "Application of structured support vector machine backpropagation to a convolutional neural network for human pose estimation," *Neural Networks*, vol. 92, pp. 39–46, 2017.

[33] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern Recognition*, vol. 26, no. 9, pp. 1277–1294, 1993.

[34] Y. Guo and T. Chen, "Semantic segmentation of RGBD images based on deep depth regression," *Pattern Recognition Letters*, vol. 109, pp. 55–64, 2018.

[35] A. Hynes and S. Czarnuch, "Human part segmentation in depth images with annotated part positions," *Sensors*, vol. 18, no. 6, 2018.

[36] G. Carneiro and N. Vasconcelos, "Formulating semantic image annotation as a supervised learning problem," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 163–168, San Diego, CA, USA, 2005.

[37] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.

[38] Z. Li, Z. Tang, W. Zhao, and Z. Li, "Combining generative/discriminative learning for automatic image annotation and retrieval," *International Journal of Intelligence Science*, vol. 2, no. 3, pp. 55–62, 2012.

[39] S. Zhu, X. Sun, and D. Jin, "Multi-view semi-supervised learning for image classification," *Neurocomputing*, vol. 208, pp. 136–142, 2016.

[40] C. Lindner, S. Thiagarajah, J. Wilkinson, T. Consortium, G. Wallis, and T. Cootes, "Fully automatic segmentation of the proximal femur using random forest regression voting," *IEEE Transactions on Medical Imaging*, vol. 32, no. 8, pp. 1462–1472, 2013.

[41] M. J. Kang, J. K. Lee, and J. W. Kang, "Combining random forest with multi-block local binary pattern feature selection for multiclass head pose estimation," *PLoS One*, vol. 12, no. 7, article e0180792, 2017.

[42] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, "Robust and accurate shape model matching using random forest regression-voting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1862–1874, 2015.

[43] C. Yang, G. Wu, K. Ding, T. Shi, Q. Li, and J. Wang, "Improving land use/land cover classification by integrating pixel unmixing and decision tree methods," *Remote Sensing*, vol. 9, no. 12, p. 1222, 2017.

[44] J. Xia, N. Falco, J. A. Benediktsson, P. Du, and J. Chanussot, "Hyperspectral image classification with rotation random forest via KPCA," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 4, pp. 1601–1609, 2017.

[45] M. A. Gowayyed, M. Torki, M. E. Hussein, and M. El-Saban, "Histogram of oriented displacements (HOD): describing trajectories of human joints for action recognition," in *International Joint Conference on Artificial Intelligence*, pp. 1351–1357, Beijing, China, 2013.