Research on Human Behavior Recognition in Static Images Based on Attitude Adaptive

Haijing Zhou[1], Shaofeng Han[2]

[1]Chongqing Vocational Institute Of Tourism, Chongqing, 409099, China

[2]Vocational Education Center Of Qianjiang Chongqing, Chongqing, 409099, China

Keywords: Attitude Recognition, Fourier Descriptors, Support Vector Machine, Feature Extraction

Abstract: human behavior recognition in static images is one of the hot research directions in the field of image processing. Correct recognition of human behavior in static images is helpful for image classification, retrieval, video monitoring and human tracking applications. In this paper, firstly, the adaptive gaussian mixture background modeling and morphological method are used to detect the edge using canny operator to realize the feature extraction of the target human body contour. Then, the human body is divided into pose parts by combining fourier descriptors based on centroid edge distance, k-means clustering algorithm and svm classifier. Image regions with similar structures are found by distance measurement method as positive data samples, and the classification model is obtained by training. Experimental results show that svm classification is more accurate and does not require repeated iterative training. It can effectively identify complex human motion behaviors in video and improve the recognition rate of behaviors.

1. Introduction

Attitude estimation and behavior recognition is a very hot research topic in computer vision, which includes many applications, such as human computer interface, full video search, robots, video surveillance and other fields. Behavior recognition has a wide application prospect in video surveillance, virtual reality, human-computer interaction, video retrieval, etc. It is easily affected by background changes, illumination changes, moving cameras, shooting angles, occlusion, clothing and other factors [1]. In life, the images that people use electronic equipment to acquire, store and transfer are all digital images, and they try to process and understand the images by using certain computer and signal processing methods, which constitutes a new science related to artificial intelligence and pattern recognition-computer vision [2]. Human behavior can be divided into group behavior, individual behavior, normal behavior and abnormal behavior under normal circumstances [3]. Group events are small probability events in daily life, and the application occasions are relatively limited, so the research in this area is relatively few. Therefore, this paper proposes to use local features and global features to mix, and then use k-means algorithm and svm classifier to distinguish the five postures of human body, namely standing upright, standing sideways, lying down, bending down and crouching down.

2. Contour Extraction and Processing

2.1 Background Subtraction

Moving object detection methods are mainly divided into background subtraction method, frame difference method and optical flow method. In this paper, the background subtraction method and the adaptive Gaussian mixture model are combined to detect moving objects. The background subtraction method is actually to subtract each image in the video sequence from the background image in the video [5]. Each pixel in the depth image records the depth value of the scene, not the light intensity. The introduction of the depth camera expands the ability of the computer system to perceive the 3D visual world, and to a certain extent makes up for the lack of dimensional information when capturing 3D object information into 2D visual information during perception.

  DOI: 10.25236/iiicec.2019.060

The background subtraction method is relatively simple to implement and has fast operation speed, and has good detection effect for scenes with relatively fixed background. However, due to the need to build a background model in advance, the good or bad background extraction will directly affect the final detection effect, so it is very important to build a background model that meets the requirements [6]. When using background subtraction method to detect moving objects in a scene, it is easily affected by external light, sound and other changes.

In this paper, a modeling algorithm based on an improved adaptive Gaussian mixture model is used to detect moving human targets. Binarize the moving target, that is, the threshold T, and set the current frame image to f_k and the background image to f_{bk}. Differentiate the background image from the current frame: $D_k(x,y) = |f_k(x,y) - f_{bk}(x,y)|$, and then binarize the difference $D_k(x,y)$ to get moving target area [7]:

$$R_k(x,y)\begin{cases} 1, R_k(x,y) > T \\ 0, R_k(x,y) \le T \end{cases} \quad (1)$$

Where $f_k(x,y)$ is the current frame image, $f_{bk}(x,y)$ is the background image, Dk (x, y) is the difference image, and T is the set threshold. If the difference $D_k(x,y)$ is greater than the threshold T, then the pixel point at that position in the image must be set to 1, and this pixel point is classified as the foreground moving target of the image frame. Otherwise, the pixel value at this position is set to 0 as the background image of the image frame. Where $R_k(x,y)$ is the target image obtained by binarizing the original image.

2.2 Normalized Target Image

The extracted binarized image is morphologically processed to remove noise and other interference factors, and the empty points in the target area are filled and redundant noise points are removed by opening and closing operations. Whether based on color images or depth images, there is no essential difference between the two only in terms of object classification. Therefore, even the same background point may present different color values at different time points. Therefore, it is impossible to accurately detect the background image using only a single Gaussian model. At this time, the mixed Gaussian distribution is used to construct the background model of the video sequence. For example, the quality of equipment performance and the influence of light intensity often lead to noise and insufficient contrast in the collected input images [8]. The purpose of normalization processing in this paper is to make images with different imaging conditions (illumination intensity, direction, etc.) consistent. After detecting the peripheral contour of the moving target in the video sequence, the height of the human body contour can be normalized to a uniform height. The extraction and representation of human behavior features based on depth images is also bound to be realized around geometrical features. For each pixel point in the picture area, several Gaussian distribution models are used to describe the distribution of pixels, thus constructing a Gaussian model and finally realizing the extraction and update of the video sequence background. The width of the contour is scaled according to a certain proportion, as shown in the human body contour image extracted in Figure 1.

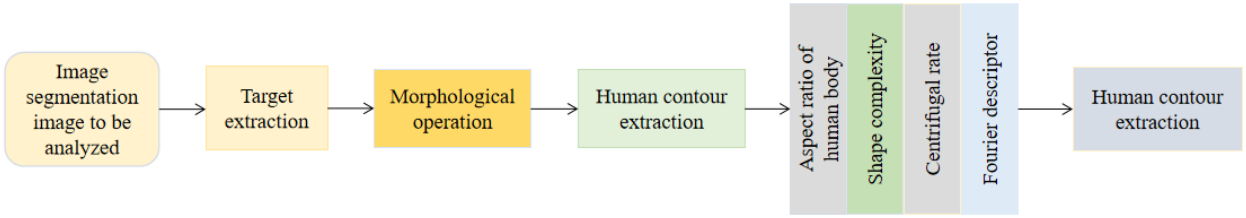

Fig.1 Overall Flow Chart of Human Body Identification

2.3 Background Model Update

As the background of the scene is affected by light and some moving objects, the background is constantly changing. This paper adopts an improved Gaussian mixture model based on dynamic adaptation to update the algorithm. It describes human motion information in the form of some unrelated points. Points of interest provide a compact representation of image content by describing the local part of the scene, thus enhancing robustness against chaos, occlusion and intra-class differences. The background model can well restrain the interference of shaking, sudden change of illumination and other external environments on the background model. It can not only completely segment larger moving objects, but also better detect smaller moving objects. The algorithm processes the image after difference. If a pixel in the image is judged as a foreground image, then the pixel is the pixel of the image. Bilinear product operation is carried out on the distinguishing features guided by body joints to capture the spatio-temporal features of behaviors. This method obtains the best recognition results when obtaining the real joint positions of the human body.

3. Human Behavior Recognition

The feature model in the algorithm uses a two-dimensional contour model, which includes human body aspect ratio k, shape complexity change rate r, and eccentricity e. The basic postures in this article include five postures: standing on the front, standing on the side, squatting down, bending down and lying down. Including static posture features, continuous motion features and offset features. Static posture features represent the position difference between joint points in the current frame, and continuous motion features represent the position difference between joint points in the current frame and the previous frame. Then the inter-frame difference method can be selected to extract the moving object. The basic principle of two-frame difference is that the gray level between adjacent pixels will change during the continuous movement of the target. By comparing two frames of images at different times under the same background, the movement of the moving target under this background can be seen. The root filter covers approximately the entire target object at a coarser resolution, while the component filter covers smaller components in the object at a higher resolution. Generally, the resolution of the component filter is twice that of the root filter.

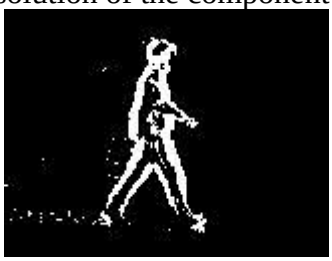

Fig.2 Normalized Image

The normalized minimum circumscribed rectangle of the human body has a height H and a width W, and the average width of the human contour is $\overline{W}$, as shown in Figure 2. The aspect ratio k of the human body is:

$$k = \frac{\overline{W}}{H} \quad (2)$$

In some postures of the human body, such as standing and lying down, the aspect ratio is very different, so this feature can be used to distinguish the two postures.

The complex shape change rate r is the ratio of the contour Area ar of the moving target to its circumscribed rectangular area:

$$r = \frac{A_r}{W \cdot H} \quad (3)$$

According to the complexity of the shape, the two postures of bending and squatting can be

better distinguished.

In the characteristic model, the eccentricity e can be well used as a characteristic, and the eccentricity is:

$$e = \sqrt{1 - \frac{b^2}{a^2}} \quad r = \frac{A_r}{W \cdot H} \tag{4}$$

Where a is the major half axis of the moving target's peripheral contour ellipse and b is the minor half axis of the peripheral contour ellipse.

The inter-frame difference method has a good detection effect on moving targets with small speed changes, but when targets with acceleration movement or deceleration movement are encountered, the detection result may have multiple detections or missed detections. Firstly, a standard pyramid of image features should be calculated, which is realized through continuous smoothing and downsampling. Secondly, the pyramid of each layer of pyramid features should be calculated, thus specifying a feature map on different scales. The offset feature is formed by calculating the position difference between the joint points in the current frame and the initial frame. The combination of these three channel features forms a preliminary feature representation [9]. Equal interval length sampling refers to selecting the coordinate points on the contour with the same curve length and dividing the contour of the human body into n equal parts (n is generally taken as the power of 2, where N=128).

Calculate the centroid (x_c, y_c) of the human target profile:

$$x_c = \frac{1}{N_b} \sum_{i=1}^{N_b} x_i \tag{5}$$

$$y_c = \frac{1}{N_b} \sum_{i=1}^{N_b} y_i \tag{6}$$

Where N_b describes the number of target pixels, and (x_c, y_c) is a pixel of the target image. The distance from any point above the human contour to the center of mass is r_i:

$$r_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \tag{7}$$

Then, you can use the Fourier transform on this set of data to get:

$$a_n = \frac{1}{N} \sum_{i=1}^{N} r_i \exp(-j2\pi i / N) \tag{8}$$

Where i = 1,2, ···, N.

4. Training and Recognition

4.1 Support Vector Machine Svm Classifier

Support vector machine (SVM) [10] is a hot topic in machine learning research and has achieved success in many fields, such as face recognition, license plate character recognition, handwritten numeral recognition, etc. Firstly, a joint point is selected as a base point for the extracted static pose features, and then the position difference between each joint point and the base point is calculated. When doing differential calculation, there will be many pixel points of non-moving targets, thus causing the detected moving targets to be larger than the actual moving targets, and dragging phenomenon will occur. However, because the behavior image region does not consider whether the region's judgment on this behavior category is relevant during model training, the processing

capacity of the training model is virtually increased, and uncertain feature information is also introduced, thus reducing the accuracy of the behavior recognition model. Since the objective of SVM is the principle of risk minimization, an objective function should be established to separate these two-level mode areas as much as possible. The method also describes the characteristics of the two channels through a weighted graph, which can deal with the instability of joint points, different sequence lengths and other issues. Finally, the method of sequence matching is used for classification.

4.2 Training of Samples

In this paper, the posture of the elderly living alone indoors is mainly recognized. Therefore, the static posture of a single person is studied here, which mainly includes five postures: standing upright, standing sideways, bending down, crouching down and lying down. The orientation of limbs is calculated for each frame as static characteristics, and then the characteristics of these limbs are compiled by Markov random fields to reduce the gap within the class. Finally, multi-channel multi-instance learning algorithm is used to learn discriminative skeletal movements. The image characteristics refer to the symbolic characteristics of moving objects in the image. Image features are the basis for distinguishing different moving objects. as factors of image features, they need to be able to cope with changes in brightness, rotation, size, etc. the first classification is based on the width-to-height ratio k value of human posture. as shown in Figure 2, there are obvious differences in k value, and three postures of human body can be judged: standing (side and front), bending (bending and squatting), and lying.

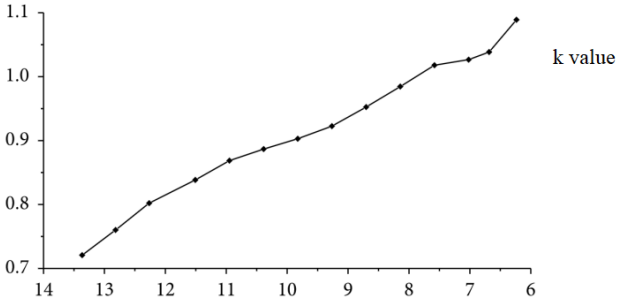

Fig.3 K Values of Five Postures

The k value cannot completely distinguish the actions of standing (front and side), bending (bending and crouching), so when the k value is in a certain range and is not exactly which posture, the eccentricity e should be used to judge which posture belongs to carry out the second classification. Assuming that people are still, human posture and human interaction can be roughly described by calculating the occupancy information of each grid. An excellent feature should have good distinction. The more obvious the difference between the eigenvalues of different attributes, the better, and the obvious difference between the eigenvalues, so as to facilitate the identification of the classifier. Attitude activation vector refers to a detector that trains body parts through positioning and annotation data of human joint points in an image, and annotations here are used to find similar attitude spaces in a given joint point configuration. It can be seen that the contour curves of the two postures of bending and crouching are very similar, but the shape complexity rate k of the two postures is very different, so k can be taken as the weight of Fourier descriptors of the two postures. Experiments show that posture features are superior to other features, and the feature information near human joints is more effective and distinguishable than the structural feature information of the whole image.

4.3 Human Posture Recognition

Each feature extraction method has its own advantages and is independent of each other. If different features can be effectively fused to obtain a more discriminative feature vector, the recognition performance will definitely be improved. The pixels of the target area or background area in the image have their own effects on the color features. Compared with other features, color

features have good stability, are not easy to change due to changes in the size or direction of moving objects, and have higher robustness. The feature extraction method is relatively simple and has been widely used. First of all, it is recognized according to the aspect ratio k. Through Table 1, it is found that the three postures of human body, standing, bending and lying down, can be distinguished according to the aspect ratio k of human body.

Table 1 Attitude Feature

Attitude	Threshold
station	k ⩽0.24
Stand or bend	0.33< k ⩽0.27
Bending	0.39< k ⩽1.2
Lie down	1.3< k <3.6

When the width-to-height ratio of the human body contour is k⩽0.24, the posture is standing. When the human body is in a bent state, the aspect ratio of the human body will change accordingly. The height becomes smaller, the width becomes larger, and the value of k will increase accordingly. When 0.39 <k⩽1.2, it belongs to bending over. status. Its advantage is that it is not easy to be affected by changes in image translation, rotation, size, etc. However, since the specific spatial position of color cannot be determined, it is only suitable for images without considering position and division. On the contrary, when the European distance difference between the two features is large, it is believed that the semantics expressed by the two features are also very different, and the semantics expressed in the image content corresponding to nature are also quite different. Extract the temporal and spatial features of the pose sequence, where the spatial feature depicts the position and mutual position of joint points on the same frame of image, and the temporal feature depicts the change in joint position due to pose changes. When 1.3 <k <3.6 The human body is in a lying state. At this time, the width is similar to the height when standing, and the height is similar to the width when standing.

4.4 Experimental Results and Analysis

The experiments in this paper identify five different poses of the human body. Dahua cameras are used to collect video data and establish a video database. The intercepted video frame rate is 30f / s. We collected a total of 6 videos for experiments. Each video There are five simplest postures of the human body, which are standing at 0 °, standing at 90 °, bending over, squatting and lying. Generally, the similarity of data objects is expressed by the distance between data objects. The further the distance between data objects, the greater the difference between data objects. The closer the distance between data objects, the more similar the data objects are. It can avoid the pre-processing of the image and directly use the original image as input. At the same time, the extracted description features are different because the video shooting starting nodes of the same action type may be different. Therefore, the spatio-temporal feature of posture cannot be directly used as the classification feature of behavior recognition. The experiment in this paper uses C++ and Opencv to implement the experiment. For each action, 30 frames of images are selected as training set samples. The experimental data of the recognition rates of the five human postures are shown in Table 2.

Table 2 Svm Classification Results

Numbering	Attitude	Recognition rate(%)
1	Stand upright	92.01
2	Stand sideways	90.11
3	bend over	89.36
4	Squat	93.51
5	Lie down	98.22

From the experimental results in Table 2, it can be seen that SVM classifier has a good result for human posture recognition, and the recognition rate is as high as 91.042%, SVM has a good classification performance for small samples.

5. Conclusion

Human behavior recognition in static images is a complicated recognition process. In this paper, according to the principle of visual perception mechanism, the image preprocessing stage is selected to analyze the semantically significant regions of the image, and a feature comparison algorithm based on sliding window is proposed to find the pixels and their regions that contribute more to the semantic of the image. In this paper, Fourier descriptors and some local features are used as feature vectors, and SVM classifier combined with k-means is used to identify simple gestures of human body. The foreground image extracted by the Gaussian mixture model and the foreground object extracted by the improved frame-to-frame four-frame difference algorithm are combined, and then morphological processing is carried out to finally obtain an ideal foreground object. Finally, the camera is used to collect the video database by itself, and then the video database is identified. Finally, good results are obtained. In short, human behavior recognition is a complex and very important research direction. To improve the performance of human behavior recognition, we need to consider not only the model itself, but also a variety of large-scale public databases.

Acknowledgment

The 2019 Municipal Education Commission Science and Technology Research Program Project, Project Name: Research on Attitude Adaptive Static Image Human Behavior Recognition, Project Number: KJQN201904601

References

[1] Bi Xiaojun, Feng Xueyun. (2017). Improvement of Feature Extraction Model of Human Behavior Recognition Combined with CNN. Science and Technology Innovation, no. 4, pp. 79-81.

[2] Chen Yuping, Qiu Weigen. (2019). Review of Research on Vision-based Human Behavior Recognition Algorithms. Application Research of Computers, no. 7, pp. 1927-1934.

[3] Tang Chao, Wang Wenjian, Wang Xiaofeng, et al. (2019). Human behavior recognition based on multi-view semi-supervised learning. Pattern Recognition and Artificial Intelligence, no. 4, pp. 376-384.

[4] Jiang Xueying, Su Chengli, Xu Yapeng, et al. (2018). Adaptive Backstepping Sliding Mode Control for Flight Attitude of Quadrotor UAV (English). Journal of Central South University, vol. 25, no. 3, pp. 616-631 .

[5] Han Xinxin, ye Qiling. (2019). Human behavior recognition method based on SIFT and hog feature fusion. Computer technology and development, no. 6, pp. 71-73

[6] Ye Dan, Li Zhi, Wang Yongjun. (2019). Research on behavior recognition method based on splda reduced dimension xgboost classifier. Microelectronics and computer, no. 6, pp. 35-39

[7] Zhu Yaolin [1], Tian Li [1], Wan Taoruan [2]. (2017). Real time human posture tracking based on template model. Computer engineering and application, vol. 44, no. 12, pp. 147-151

[8] Fan Zijian, Xu Wei, Liu Feifan, et al. (2017). Kinect-based dynamic recognition method of learner's head pose. Computer and Digital Engineering, vol. 45, no. 2, pp. 360-366.

[9] Zhang Man, Xian Hequn, Zhang Shuguang, et al. (2017). Research on authentication technology based on gravity sensor. Information Network Security, no. 9, pp. 58-62.

[10] Chen Hao, Xiao Lixue, Li Guang, et al. (2019). Aggressive behavior recognition based on human joint point data. Journal of Computer Applications, vol. 39, no. 8, pp. 2235-2241.