Research on Human Action Recognition in Dance Video Images

Hua He\textsuperscript{1,*}

\textsuperscript{1}Nanfang College of Sun Yat-sen University, Guangzhou Guangdong, 510970

*Corresponding author e-mail: hehua1357091909@sysu.edu.cn

Abstract. This article mainly researches the recognition technology of dance video in human movement. Image preprocessing, codebook establishment, Zernike moments and support vector machines are used to classify and recognize human movements in dance videos. From the results of simulation experiments, using the recognition method proposed in this paper to recognize dance movements in dance videos in the database can effectively improve the recognition rate of human movements, and then better guide dancers’ dance movements.

Keywords: Dance Video; Image Segmentation; Human Action Recognition; Support Vector Machine

1. Introduction
One of the hotspots in computer research in our country is the recognition of human actions in videos. This technology uses image processing and recognition analysis techniques to extract and analyze the actions of people in the video to determine the actions and behaviors of the people in the video, obtain effective information, and its uses are extensive. The key to implementing this technology is to properly preprocess the original video, and then perform extraction operations on image features in the video, and classify and describe them.

2. Overall thinking research
Our country's recognition technology for actions in video images has just begun to develop, let alone combine it with dance art. Using this technology to recognize human movements in dance videos can effectively identify dance knowledge in videos. After the dance moves are recognized, they are compared with the standard dance moves, and the dancers' movements are evaluated objectively based on this, and the dancers are given corresponding suggestions for movement correction. It is a new type of dance movement auxiliary training method. This article applies the related technology of human action recognition, trains the relatively simple KTH database into an SVM classifier, and then uses the A-go-go dance video database SVM classifier for intensive training to improve the function of the SVM classifier. Finally, the trained classifier is used to recognize the human actions in the A-go-go dance video, so as to achieve the purpose of obtaining a better classification effect [1-3].

3. Data processing of images in video
3.1. Transform gray scale
Before proceeding with the video image, it is necessary to extract the image from the video first, and perform operations such as grayscale conversion, image thresholding, and image segmentation to reduce the amount of computer calculations and facilitate the extraction of more effective information. The most common video image in our daily life is a true color image, that is, an RGB image, that is, each pixel in the image is composed of three primary colors (R, G, B). Because of the complexity of true-color images, if the true-color images are processed directly, the amount of calculations on the computer will increase dramatically and the analysis efficiency will decrease. Therefore, first transform the true color image into a grayscale image. Then reduce the color information contained in the video image [4-6].

3.2. Thresholding moving images
In order to obtain a binary image of a moving image, the image needs to be thresholded first. The main method of thresholding the image is to select a reasonable threshold and divide the pixel gray value reasonably by the threshold. To segment the moving image, the general form of the threshold can be written as:

\[ T = T[x, y, f(x, y), p(x, y)] \]

In the above formula, the gray value of the pixel at \((x, y)\) is used to represent \(f(x, y)\); the gray gradient function of this point is used to represent \(p(x, y)\). After calculating this formula, the binarized image can be obtained.

3.3. Segmentation of moving images
In the above operation, the binarized image of the current moment of the image in the video is obtained, and then the scene and the motion area in the video need to be separated, which is related to the segmentation of the motion area. This article uses the binary image in Matable software to process the function, and then establishes a reasonable threshold to find the contour of the human body in motion.

The specific operation is: assuming the size is expressed as \(M \times N\), the current time is \(t\), and the frame at time \(t\) in the video is expressed as \(P(x, y, z)\), to obtain the binary image \(A(x, y, z)\).

At time \(t\), the background gray value of the binarized image \(A(x, y, z)\) is equal to 0, and the foreground gray value is equal to 255. Scan along the column vector of \(A\), and then count the number of foreground pixels in each column \(C(1 < C < n)\). Select the largest value of \(n\) frequency data, namely \(C_i\), and denote the column number corresponding to the maximum value by \(i\). If the ratio between \(C_i\) and the number of lines \(m\) generated between frames is greater than 1/6, it means that the frame contains the area of the moving human body. If 1/6 < \(C_i\) < 1/6, it means that the frame only includes the partial motion of human body; update the current time value, let \(t = t + 1\), all frames in the video are scanned, then the image after image thresholding and image segmentation is shown in the following figure:
3.4. Zernike moments extract overall features

After binarizing the image in the motion video, this paper uses Zernike moments to describe the characteristics of the binary image, and then classifies and recognizes the image. Zernike moment is an extremely effective orthogonal moment to describe the shape. It is widely used in various image processing. Its main function is to make the information extracted from the video more complete and less redundant information. As an image sequence, the calculation formula of Zernike moment is as follows:

\[
A_{mn\mu} = \frac{n+1}{\pi} \sum \sum \sum_{i=2}^{images} U(i, \mu, \gamma) [V_{mn}(r, \theta)] P(x, y)
\]

In the above formula, images represents all the numbers of images in the entire sequence; \(U(i, \mu, \gamma)\) represents the introduced third dimension, as shown below:

\[
U(i, \mu, \gamma) = (x_i - \bar{x}_{i-1})(y_i - \bar{y}_{i-1})
\]

In the above formula, \(x_i\) represents the center of gravity of the current image, then, \(x_{i-1}\) also represents the center of gravity of the previous image; in the same way, \(Y\) and \(X\) represent the same meaning; \(\mu\) and \(\gamma\) are the parameters set by the user, the sequence of Differences may cause differences in the number of images. Therefore, after calculating 3D Zernike, you need to use the following formula to normalize it:

\[
\bar{A}_{mn\mu} = \frac{A_{mn\mu}}{(A \times images)}
\]

In the above formula, \(A\) represents the number of pixels of the target (ie average area), and images represents all the numbers of images in the entire sequence. Use the above formula to process the image sequence of the human silhouette, and then obtain the corresponding 3D Zernike moment, which is the overall feature.

3.5. Create codebook

Combine the representative samples in the sample space to form a codebook. The samples in the codebook can easily distinguish this category from other categories. The method of creating codebooks in this paper is cluster analysis, defining similarity measures of various categories, and describing and analyzing the similarity measures between different categories, and then based on the similarity measures of 3D Zernike moments to complete the code.

Whether the directions of the two vectors are similar is the basis for judging the similarity measure.
Assuming that the vector of the 3D Zernike moments of the two pictures has been obtained and expressed by \( x, y \), the similarity coefficient can be expressed by the following formula:

\[
\cos(x, y) = x^T y / \left[ (x^T x) (y^T y) \right]^{1/2}
\]

In this paper, the pictures with similar coefficients \( \cos(x, y) > 0.8 \) are judged as the same action, and the action pictures represented by the similar coefficients \( \cos(x, y) > 0.9 \) are selected to make them the standard requirements of the descriptor matrix. This also represents how many actions there are in the database, and how many descriptors matrix can be selected from them [7-9].

3.6. Classification of SVM

After the codebook is created, any action selected from it is represented by a series of key gestures. When a frame of image corresponds to a certain descriptor, then the descriptor represents that it is very close to a certain key posture in the code book, so each action is composed of a series of key postures in the code book, and use SVM to achieve. SVM (ie Support Vector Machine) is a two-class classification model. It is a generalized linear classifier that performs binary classification of data in a supervised learning method, and uses a one-to-one method to expand it to multiple Class classifier. Its representation is as follows:

1) Suppose the training set is:

\[
T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\} \in (X \times Y)
\]

In the above formula, \( x_i \in X = \mathbb{R}^n, y_i \in Y = \{1, -1\}, i = 1, 2, \ldots, l, x_i \) all represent the feature vector.

2) Choose the appropriate parameter \( C \) and kernel function \( K(x, x') \), construct and solve the optimization problem.

\[
\min_{\alpha} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^{l} \alpha_j
\]

s.t. \( \sum_{j=1}^{l} y_j \alpha_j = 0, 0 \leq \alpha_i \leq C \)

And get the optimal solution:

\[
\alpha = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_l^*)^T
\]

3) Select the positive component \( 0 \leq \alpha_i^* \leq C \) of \( \alpha \), and use the following formula to calculate the threshold:

\[
\beta = y_j - \sum_{i=1}^{l} y_j \alpha_i^* K(x_i - x_j)
\]

4) Building a decision function

\[
f(x) = \text{sgn} \left( \sum_{i=1}^{l} \alpha_i^* K(x, x_i) + \beta^* \right)
\]

When using SVM for classification operations, first select part of the analysis samples from the existing analysis samples, and use them as learning samples to train the SVM classifier until a model with suitable accuracy is obtained. Finally, the remaining samples in the same sample space Samples are classified, identified and calculated.

4. Simulation test

In order to verify the effectiveness of the method proposed in this article, two different types of video databases, A-go-go dance video and KTH database, are used to verify the effect of this method.

In the KTH database, there are 6 different types of common actions, including jogging, running, walking, waving, punching, and clapping, and they are some simple and basic actions. Each behavior
has 100 different action videos in the KTH database, which are displayed by 25 completely different types of people in 4 different backgrounds.

Use the algorithm introduced in the previous chapter to preprocess the video frame by frame, extract features, and build codebooks. Then the many descriptor matrices obtained are reused to make them the training samples of the SVM classifier. After simulation and comparison, this paper decides to use the card house kernel function in the Matabl software as the decision function, and use the SVM classifier to make 6 Different types of common actions are classified, and the results are as follows:

![Figure 2. The recognition rate of actions in the KTH database](image)

It is not difficult to see from Figure 2 that using the SVM classifier to recognize the actions of different videos in the KTH database can effectively improve the recognition rate of human actions. The highest recognition rate is walking, which is 92%, and the lowest recognition rate is 85%, and the average recognition rate of all samples in the KTH database is 88.71%. From the results, it can be seen that the recognition rate of using SVM classifier to classify and recognize samples is higher than usual.

In order to verify the classification and recognition effect of this method on dance videos, the SVM classifier is used to classify and recognize the dance movements of the videos in the A-go-go dance video database. Unlike the KTH database, there are 19 different basic dance moves in the A-go-go dance video database, and make 5 dancers of different dance types collect 19 moves 3 times, then each type of movement can get 15 completely different sample data, so 19 moves have a total of 285 different human body motion sample data. In order to make full use of the SVM classifier, 200 sample data were extracted from 285 different sample data and divided into 200 groups. The SVM classifier obtained from the KTH database was used to train these 200 groups of samples. Finally, the SVM classifier is used to classify and recognize the dance movements in the A-go-go dance video database. The results are as follows:
Figure 3. The recognition rate of dance action video data in the A-go-go database

It can be seen from the figure above that the use of the enhanced SVM classifier to recognize the dance movements of the video in the A-go-go dance video database can effectively improve the recognition rate of dance movements, among which the lowest recognition rate actions are 2 and 15, both of them recognition rate are 86%, the highest recognition rate is action 3, which is 95%, and the average recognition rate of the entire action sample is 90.4%. From the results, we can see that this recognition method has a higher recognition rate for human actions in dance videos [10].

5. Conclusions
This article analyzes the recognition technology in the video image, and focuses on the method and effect of the recognition of the dancer’s body movements in the dance video. This article first extracts the video image from the video, and implements various operations such as binarization, grayscale and image segmentation on the video image. Then, the 3D Zernike moments in the binarized video pictures are sequentially classified according to the different features, and the similarity coefficient codebook based on the 3D Zernike moments is the standard requirement. The last step is to use the SVM classifier to classify and recognize the dance videos in the A-go-go dance video database and the videos in the KTH database one by one. From the results, it is not difficult to see that this method can obtain high accuracy of human action classification and recognition rate.

References
[1] Wei Yingzi, Shi Fuyu. A Local Description Method of Human Body Movement Classification. Electronic World, 2020(21): 102-104.
[2] Wu Qian, Wu Fei, Luo Lizhi. Human Skeleton Action Recognition Based on Deep Learning. Electronic Science and Technology, 2020, 33(11): 79-83.
[3] Zhang Hengxin, Ye Yingshi, Cai Xianzi, Wei Fuyi. High-efficiency Motion Recognition Algorithm Based on Human Joint Points. Computer Engineering and Design, 2020, 41(11): 3168-3174.
[4] Huang Haixin, Wang Ruipeng, Liu Xiaoyang. Overview of Human Behavior Recognition Technology Based on 3D Convolution. Computer Science, 2020, 47(S2): 139-144.

[5] Li Yibin, Guo Jiahao, Zhang Qin. Human Gait Recognition Method and Technology. Journal of Jilin University (Engineering Science Edition), 2020, 50(01): 1-18.

[6] Zhang Congcong, He Ning. Human Action Recognition Method Based on Key Frame Dual-Stream Convolutional Network. Journal of Nanjing University of Information Science & Technology (Natural Science Edition), 2019, 11(06): 716-721.

[7] Lv Jie, Li Hongqi, Zhao Yanhong, Sikandar Ali, Liu Yanfang. Human Behavior Recognition Based on Joint Space-Time Feature Fusion. Computer Engineering and Design, 2020, 41(01): 246-252.

[8] Tian Yuan, Li Fangdi. Survey of Human Body Gesture Recognition Based on Depth Information. Computer Engineering and Applications, 2020, 56(04): 1-8.

[9] Song Zhen, Zhou Yuanfeng, Jia Jingong, Xin Shiqing, Liu Yi. Local Feature Fusion Temporal Convolutional Network for Human Action Recognition. Journal of Computer Aided Design and Graphics, 2020, 32(03): 418-424.

[10] Jing Yuqin. Research on Feature Extraction Method of Human Behavior Recognition. Journal of Science & Technology Economics, 2020, 28(04): 149-150.