**Journal of Cases on Information Technology**
Volume 24 • Issue 5

# Implementation of a Human Motion Capture System Based on the Internet of Things Machine Vision

Fang Yu, Jilin Sport University, China*

## ABSTRACT

The classification of the stereo matching comprehensive analysis related algorithm model can be subdivided into local stereo matching based on the entire acquisition and global stereo matching based on the entire local. But it can have a higher capture efficiency because the log-likelihood variance cost calculation function can have a faster feature convergence capture speed than the ordinary log-mean-square error cost function. Through the combination of gray channel and frame difference channel, a better network structure and parameters on the KTH data set are obtained, which can ensure the classification effect while greatly reducing the number of parameters, improving training efficiency and improving classification accuracy. The article uses dual-channel 3D convolutional human neural network technology to achieve 92.5% accuracy of human feature capture, which is significantly better than many traditional feature extraction techniques proposed in the literature.

### KEYWORDS

Human Motion Capture, Internet of Things, Machine Vision, Stereo Matching Algorithm

## 1. INTRODUCTION

Visual science is an important way to help human individuals perceive the external environment world and obtain psychological information of the external environment. Due to its unique characteristics of external space-time and space, visual environment perception often plays an extremely important leading role in the human body perception processing system of all kinds of people, and the environmental information that can be obtained and obtained from the human body's visual perception system is also more intuitive and rich. With the rapid and in-depth development of mobile Internet and multimedia information technology, the amount of visual information in the real-life world continues to grow rapidly. If directly making a computer itself have the ability to calculate motion perception based on human body vision, then it can directly fill all of these needs. Although individual human beings are the main body of social relations, their behaviors and psychological actions often have specific social meanings (Ding, 2019; Chang 2019). Accurate recognition of various human behaviors can often directly help to better understand the social and psychological behaviors of various people. Therefore, accurate recognition of human actions has developed into one of the most active research topics in the field of visual science in the current computer era.

Automatic tracking and surveillance video automatic surveillance technology based on human-computer vision is a multi-disciplinary academic interdisciplinary research topic that has attracted

DOI: 10.4018/JCIT.302245 *Corresponding Author

much attention from all walks of life in recent years. The positioning detection and automatic tracking technology of the facial image of the moving robot under the background of complex information is the current academic hotspot and research difficulty in the development of various fields of computer automatic vision technology research in the world. There are extensive research applications in various fields such as navigation, traffic management, multimedia teaching, target perspective recognition and automatic tracking, and security automatic monitoring. At the same time, it can also be the theoretical basis for various image subsequent advanced image processing technologies such as image target perspective classification, behavior pattern recognition and logical understanding. The main research results of this research topic can not only be widely used in solving various illegal motion captures, but also can be applied to safety automatic monitoring (Wan, 2019; Wei, 2019). And further in-depth research on the automatic vision of the robot, the creation of the map library for automatic positioning and tracking of the image of the mobile robot in the unknown information environment, the question bank and the tracking of the target person are of important research significance. This research is based on the GMM algorithm and the stereo matching algorithm to encode the human motion capture algorithm of the video, which makes the human motion capture algorithm more sensitive.

The research on human motion capture based on visual sensors is still in the theoretical stage of the laboratory, and it still need to make a certain amount of reference to other people's theories. JL Jiménez Bascones has constructed a high-speed and intelligent visual image sensor that integrates visual image information acquisition and signal processing. The extensive use of DSP also enables the visual sensor not only to directly have its own image cpu, but also to directly perform some complex mathematical operations, but due to its greater flexibility, it is not suitable for general conditions (Jiménez, 2019). In order to realize the image measurement of the height and accuracy of the digital car interior, Ma H has steadily improved the technical level of the automotive visual image detection system and studied the technical content of the core component of the automotive visual image sensor. But the research did not enter into deeper theoretical discussion (Ma, 2019). Xuan T N system adopts depfga algorithm to realize the automatic algorithm of automatic bottom visual image data processing, which obviously greatly accelerates the processing speed of bottom image data and enhances the processing performance of image sensor. The system design has successfully realized the information intelligence and information networking of the automatic vision image sensor, which has great practical application value. However, due to the excessive development of the quality of this design and no attention to the speed of transmission, the theory is still to be studied (Xuan, 2019). Y Li carried out in-depth research on various related application technical characteristics of digital distributed visual image sensors suitable for intelligent applications. He used technology to promote the data detection management system from the two-way transmission of distributed analog signals, the centralized processing of data by the central management computer to the two-way transmission of distributed digital signals, and the two-way conversion of distributed data processing methods between the detection nodes. This further greatly improves the system's data stability, real-time performance and data measurement processing accuracy, but this theory is difficult to put into practice (Li, 2019). Holt C mainly studies the hardware design of fpga and odvs, embedded linuxodvs system structure, odvs automatic video data server and other odvs video application system management software design and development. He elaborated on the specific technical application of the multi-channel target automatic data tracking system based on fpga and odvs, but the application cost is too high and there is no actual development value (Holt, 2019). The non-linear convolution scanning algorithm used by Heydari M J can obtain the information scanned by the virtual object machine, and then aggregate and classify it according to the features in the information structure scanned by the machine to generate a highly laser feature image. Based on the principle of a highly symmetrical laser feature image detection system, the false image information points obtained after matching are eliminated, and the combination of these remaining false information points is output as the result of machine recognition. But the final test result failed to reach the expected theoretical effect (Heydari, 2019). In this paper, Ashhar K proposes an algorithm svmn that uses svmb-lbp and combines texture

number and feature cardinality with radial cardinality as the core basis function of the system to perform intelligent face recognition. Algorithm analysis can quickly complete various face recognition based on vision using intelligent face robots. For various face recognition images with large changes in visual light intensity, reducing the number of superimposed substitution conversions through the algorithm greatly reduces the operation time and complexity of face recognition, but it reduces the accuracy of the algorithm that should have (Ashhar, 2019).

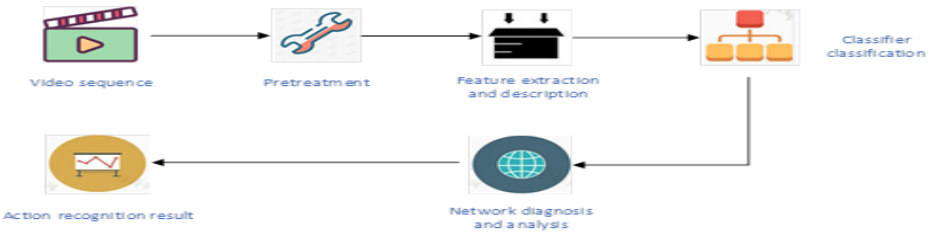
The paper mainly encodes the human body capture system through the GMM algorithm, and then uses the three-frame difference method to detect whether there is a moving target entering the video range. If a foreground target enters, the startup program, which is the target detection program described above, is started to detect. That is, the target binary image obtained after processing by the three-frame difference method, filtering and mathematical morphology. In the intelligent analysis algorithm of intelligent human body movement posture type classification action recognition, its design and successful research has realized a cloud-based multi-channel human facial feature recognition information database fusion. And the method to effectively improve the similarity of human images, intelligent movement, human facial movement target posture type classification, action recognition, intelligent analysis and detection algorithm. This is used to quickly and effectively classify and intelligently recognize motion postures based on various human features. It constructs the volumetric volume in the three-dimensional space-time dual-channel 3d volume, that is, the recognition and matching operation of the convolutional dimension is performed between the dimensions of the two convolutional layers in different three-dimensional space-time, and various intelligent human facial gesture recognition actions are extracted.

## 2. PRINCIPLE AND OVERVIEW

### 2.1 Human Body Motion Extraction Method

At present, the realization of vision-based human-computer interaction technology generally only targets limited simple body movements, such as running, walking, jumping in place, and so on. And it has strict conditions, such as it must be facing the camera, multiple human bodies cannot appear, and it is only suitable for simple background environments. In addition, related algorithms usually have problems such as large calculation amount, serious memory consumption, and not suitable for the development of embedded miniaturization. Human body video action feature recognition technology broadly refers to through practical training and theoretical learning, so that professional computer personnel can correctly understand the recognition meaning of a person's video action in a specific scene in a human body video. Now Figure 1 shows several basic operation processes of human video action feature recognition, which are called video data feature acquisition, preprocessing, feature extraction and data description, and classification feature recognition. Then, the video classifier is used to identify and analyze the data types in these video features. Figure 1 shows the basic flow of human video action feature recognition:

**Figure 1. The basic process of human action recognition**



The representation of human motion video data consists of a series of continuous images and the source video data consists of a series of continuous images, and each image contains the RGB value of a point. The RGB color mode is a color standard in the industry. It is achieved by changing the three color channels of red (R), green (G), and blue (B) and superimposing them with each other. RGB is the color representing the three channels of red, green, and blue. This standard includes almost all the colors that human vision can perceive, and it is one of the most widely used color systems. Only by correctly mapping the source data can the granularity and sensitive attributes contained in the data be produced. The current current behavior methods can be divided into human body structure models, from the perspective of the structure and characteristics of motion, according to the human body model and behavior behavior of motion (Bernd, 2018; Kettering, 2016).

Pattern matching is one of the most common basic pattern recognition algorithms, which refers to the alignment of two images. Image matching technology for the same target (or multiple images) in space is widely used in navigation, aerospace technology and other industries and fields. Common gray-scale image matching methods such as target recognition and scene intrusion include gray-scale matching methods and feature-based matching methods. Gray matching methods are generally point-to-point matching, so the relative accuracy is relatively high, and it is very sensitive to changes in interference. When the external light changes, the detection result will be affected. At the same time, it is very complicated from a computational point of view, and rotation, deformation, and occlusion are all subtle. The selected attributes include points, lines, borders, textures, etc. Resist external light changes by isolating certain characteristics, scale changes and rotation changes then use these features to match methods (Zidek, 2018; Poornima, 2020). To make it very stable in the coordinate distribution in the image, first preset the exposure I (x, y, t) of a coordinate point (x, y) in the image at time t. Let u (x, y) and v (x, y) be regarded as the points (x, y) when decomposing the motion decomposition vectors in the x and y axis directions respectively. In a very short time differential dt, the point (x, y) is shifted to the pixel coordinates (x + dx, y + dy), and dx = udt, dy = vdt. This represents the relationship between the spatial grayscale of the grayscale in the image and the optical flow velocity. Table 1 is a list of parameters taken under different video sensors under the matching method:

**Table 1. Comparison of detection at different width thresholds**

| Width threshold | Fixed threshold (5 Pixels) | Template width Of 1/3 | Template width Of 1/2 | Template width 2/3 |
|---|---|---|---|---|
| Foreign body (<=2cm) | 6589 | 5489 | 5784 | 5995 |
| Foreign body (2~8cm) | 1520 | 1621 | 1425 | 1236 |
| Foreign body (>8cm) | 1868 | 1896 | 1758 | 1851 |

In the optical flow constraint equation, both u and v are unknown. And there is only one constraint equation, so the luminous flux cannot be uniquely determined, so other constraints must be added to define u and v. Moreover, the introduction of different constraints will lead to different methods of calculating luminous flux. The most classic is the gradient optical flow algorithm, also known as the differential method, most of which are based on the gradient function of the gray image to get the motion vector of each pixel in the image (Shanthi, 2020; Choi, 2016). The length of the gradient function is this maximum rate of change. More strictly speaking, the gradient of the function from Rn to R in Euclidean space is the best linear approximation at a certain point in Rn. In this sense, gradient is a special case of Jacobi matrix. In the process of image template matching, due to some perspective distortion between the image to be matched and the template image, there are some phenomena of

partial non-matching and complete non-matching. How to filter out these situations, and how to measure the similarity between the image to be detected and the template. The distance matrix D is:

$$Q = q_1, q_2 \cdots\cdots\cdots, q_n \tag{1}$$

$$C = c_1, c_2, \cdots\cdots\cdots\cdots c_n \tag{2}$$

$$D = \begin{matrix} d\left(q_1, c_1\right) & d\left(q_1, c_2\right)\cdots\cdots\cdots\cdots & d\left(q_1, c_n\right) \\ d\left(q_2, c_1\right) & d\left(q_2, c_2\right)\cdots\cdots\cdots\cdots & d\left(q_2, c_n\right) \\ d\left(q_n, c_1\right) & d\left(q_n, c_2\right)\cdots\cdots\cdots\cdots & d\left(q_n, c_n\right) \end{matrix} \tag{3}$$

The path of the curve is defined as: at two distances of different time series D, the path of curve W is defined as a series of continuous matrix elements with different relationships between time series:

$$W = \left\{w_1, w_2, \cdots\cdots\cdots\cdots w_n\right\} \tag{4}$$

The curved path satisfies the following constraints:
Boundedness

$$\mathrm{Max}\left(m, n\right) \le k \le m + n - 1 \tag{5}$$

Boundary conditions

$$w_1 = D\left(1,1\right), w_k = D\left(n, m\right) \tag{6}$$

Continuity

$$w_1 = D\left(i, j\right), w_{k-1} = D\left(O, P\right), I - O \le 1, J - P \le 1 \tag{7}$$

Monotonicity
Not allowed to appear at the same time

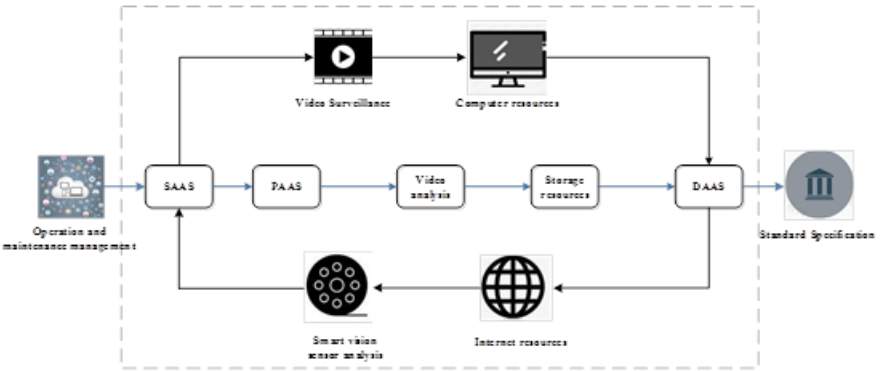
$$I - O \le 1, J - P \le 1 \tag{8}$$

Therefore, after the above definition, find the iterative formula of the optimal path. The iterative method, also known as the toss and turns method, is a process of continuously using the old value of the variable to recurse the new value. The direct method (or called the one-time solution method) corresponds to the iterative method, that is, the problem is solved at once. Iterative algorithm is a basic method of solving problems with computers. It uses the characteristics of fast computing speed

and suitable for repetitive operations to allow the computer to repeatedly execute a set of instructions (or certain steps). Each time this set of instructions (or these steps) is executed, a new value of the variable is derived from the original value of the variable. The iteration method is divided into precise iteration and approximate iteration.

Human motion capture is developed on the basis of intelligent visual sensors, using the combination of network real-time monitoring and visual sensors. The framework diagram designed is shown in Figure 2:

**Figure 2. Framework diagram of the action collection system**




After introducing the feature extraction content of human motion, the subsequent computer needs to use a series of algorithms to capture the human motion in the video content. The specific capture algorithm uses a stereo matching algorithm and section 2 for specific details.

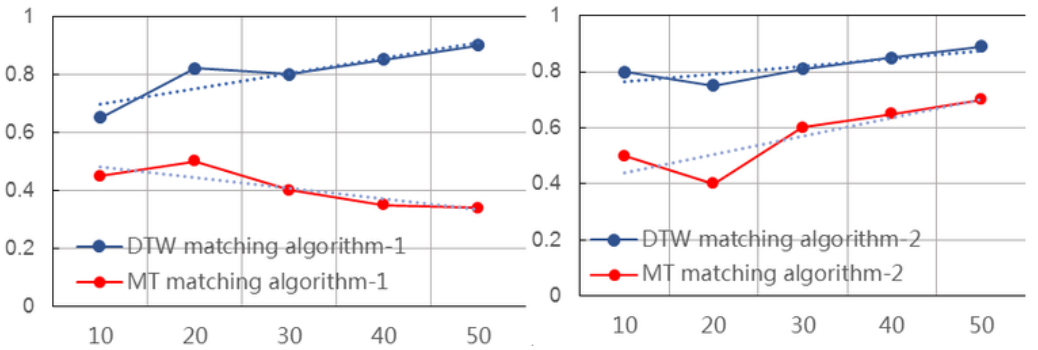
## 2.2 Human Motion Capture Method

Whether the stereo matching algorithm can achieve better results mainly depends on three factors, including matching primitives, matching criteria and matching algorithms. Among them, the stereo matching analysis algorithm is also a crucial part. The stereo matching analysis algorithm is actually a calculation process used to find the optimal disparity solution. It finally determines the dense parallax from the pixel to the focus by solving the dense extreme parallax of a function pixel with an energy function cost. The analysis algorithm benchmark of stereo matching can realize different stereo matching standard classification according to different stereo matching comprehensive standard classification criteria. Based on the basic element criterion of the stereo matching comprehensive analysis algorithm, the classification of the stereo matching standard is classified according to the criterion. Generally speaking, it can subdivide its categories into three-dimensional matching based on the entire collection part, stereo matching based on the entire local global and stereo matching based on the entire local collection feature. The stereo matching algorithm mainly establishes an energy cost function, and estimates the pixel disparity value by minimizing the energy cost function. The essence of the stereo matching algorithm is an optimization problem. By establishing a reasonable energy function, adding some constraints, and using the method of optimization theory to solve the equation, this is also the method for solving all ill-conditioned problems.

Using this algorithm to analyze and optimize the dense and sparse size matching degree of the generated image depth range is the benchmark of the stereo matching algorithm. Stereo matching can be comprehensively analyzed and related algorithm model categories, and then subdivided into local stereo matching based on the entire acquisition and global stereo matching based on the entire local.

In addition, if this algorithm is used to optimize the density and sparse size of the depth map of the dense disparity image, the matching benchmark can also subdivide its algorithm category into sparse dense dense disparity stereo matching and dense sparse dense disparity stereo matching (Harfmann, 2016; Shimada,2021). This article introduces the types and categories of the commonly used stereo matching comprehensive analysis algorithms listed below. The main stereo type categories include: stereo matching based on the entire local acquisition feature, stereo matching based on the entire acquisition local, and stereo matching based on the entire local global. The actual application effects of the two matching algorithms are shown in Figure 3:

**Figure 3. Trend chart of the application of DTW and MT matching algorithms**




The feature-based stereo matching algorithm refers to the disparity obtained by matching some features in the image. The features in an image usually involve points, lines, and geometric regions. Among them, the point feature is the most important feature. When using feature point matching, the feature point needs to be extracted before matching, and the extracted feature point must be clearly distinguishable from other pixels, so as to ensure the accuracy of parallax. This matching method is simple to implement, fast in matching speed, and has good real-time performance (Hu, 2021; Liu, 2020).

This system uses the DTW algorithm to locate and describe the human body characteristics. The system's positioning is compared with the actual human body characteristics of the human body, and then the system will make corrections to give the correct self-contained algorithm.

$$\mathrm{T}=\left\{\mathrm{T}\left(1\right),\cdots,\mathrm{T}\left(\mathrm{m}\right),\cdots,\mathrm{T}\left(\mathrm{M}\right)\right\} \tag{9}$$

$$\mathrm{R}=\left\{\mathrm{R}\left(1\right),\cdots,\mathrm{R}\left(\mathrm{n}\right),\cdots,\mathrm{R}\left(\mathrm{N}\right)\right\} \tag{10}$$

It can be expressed as follows:

$$\mathrm{T}\left(\mathrm{M}\right)=\left\{\theta_{\mathrm{t1}},\theta_{\mathrm{t2}},\cdots,\theta_{\mathrm{tx}}\right\} \tag{11}$$

$$\mathrm{R}\left(\mathrm{N}\right)=\left\{\theta_{\mathrm{r1}},\theta_{\mathrm{r2}},\cdots,\theta_{\mathrm{rx}}\right\} \tag{12}$$

In the feature sequence, the algorithm of the system is as follows:

$$\mathrm{P}=(\mathrm{k}_{\mathrm{i}},\mathrm{l}_{\mathrm{i}}),\mathrm{i}=1,\mathrm{j}=1,\cdots,\mathrm{M} \tag{13}$$

$$k_{\mathrm{l}}=l_{\mathrm{l}}=1 \tag{14}$$

$$k_{M}=l_{M}=M,k_{i}<k_{i+1},l_{i}<l_{i+1} \tag{15}$$

In the above formula, k is the depth base point density of the video sequence, and l is the time length sequence of the video. From this, it can be deduced that the distance difference between the human body feature and the system algorithm is shown in the following formula:

$$\mathrm{d}\left[\mathrm{T}\left(\mathrm{l}_{\mathrm{j}}\right),\mathrm{R}\left(\mathrm{k}_{\mathrm{i}}\right)\right]=\sum_{\mathrm{n}=1}^{\mathrm{x}}\mathrm{k}_{\mathrm{n}}\theta_{\mathrm{rn},\mathrm{l}_{\mathrm{j}}},\theta_{\mathrm{tn},\mathrm{k}_{\mathrm{j}}} \tag{16}$$

$$\mathrm{D}\left[\mathrm{T},\mathrm{R}\right]=\sum_{\mathrm{k}_{\mathrm{i}},\mathrm{j}_{\mathrm{i}}\epsilon\mathrm{P}}\mathrm{d}\left[\mathrm{T}\left(\mathrm{l}_{\mathrm{j}}\right),\mathrm{R}\left(\mathrm{K}_{\mathrm{i}}\right)\right] \tag{17}$$

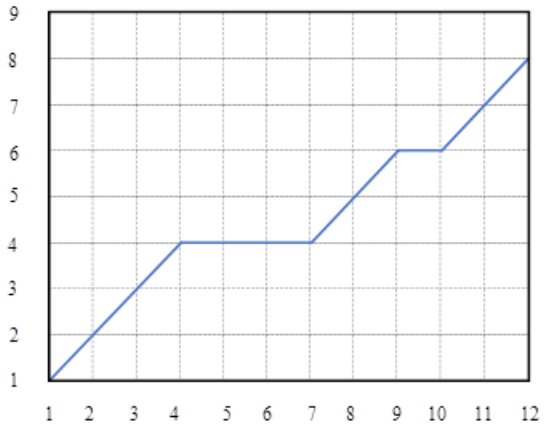
$$\mathrm{DTW}\left[\mathrm{T},\mathrm{R}\right]=\mathrm{minD}\left[\mathrm{T},\mathrm{R}\right] \tag{18}$$

$$\underset{ES}{\rightarrow}=\left(S_{X}-e_{x},s_{y}-e_{y}\right) \tag{19}$$

$$\underset{EH}{\rightarrow}=\left(h_{X}-e_{x},h_{y}-e_{y}\right) \tag{20}$$

Where h is the height of the human body in the machine video, and e is the plane position of the human body in the video sequence. When the human body makes a series of motion cycles, the simplified diagram of the optimal path of the feedback algorithm of the system is shown in Figure 4:

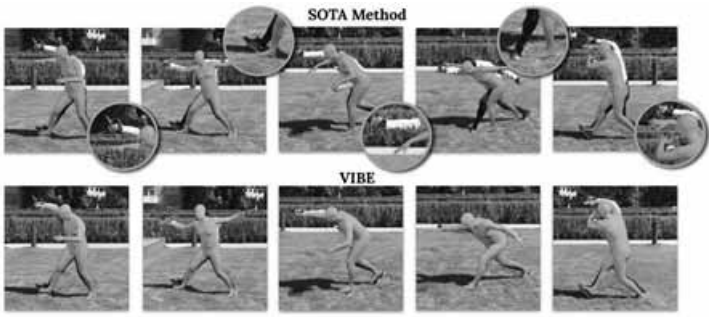
**Figure 4. DTW shortest path diagram**




## 2.3 Human Body Movement Data Collection

Among them, an algorithm is used to accurately determine the seven multi-pole posture features of the human body's characteristic posture. This is the 17-dimensional multiple human posture feature vectors of the human body image satellite launched by China in 2012. Therefore, the computational complexity and storage capacity can be effectively reduced, and the system can achieve a good real-time recognition effect (Kaitlyn, 2020; Xu, 2020). The 3D convolutional neural network applies trainable filters and local neighborhood pooling operations to the original input to obtain a hierarchical and gradually complex feature representation. There are practices that show that it can achieve very good results if appropriate regularization items are used for training. One thing that CNN is also popular is that it will be invariant to poses, lighting, and complex backgrounds. At present, the public video database technology of human facial action video recognition data in the video center can be roughly divided into three types of technologies and some commonly used video databases are shown in Figure 5:

**Figure 5. Common data sets for video human behavior recognition**




However, since the matching effect mostly depends on the selection of feature points, the requirements for feature point selection are very high, and the effect is not very ideal in some areas with slow gray changes or weak texture areas (Ranjan, 2020; Souza, 2020). Since the number of feature points is small, the disparity map obtained is a sparse disparity map, so after this kind of

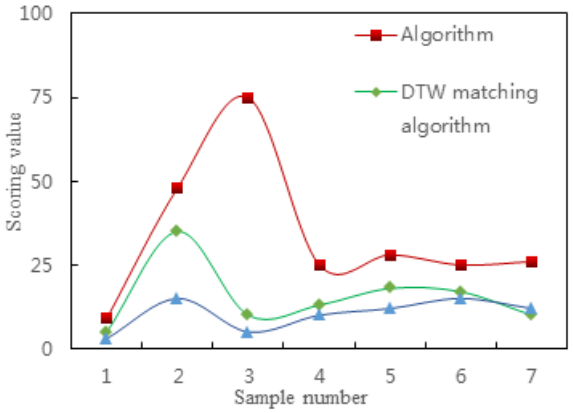
matching, an interpolation algorithm is often needed to obtain a denser disparity map. Interpolation method, also known as "interpolation method", is to use the function f (x) to insert the function value of several points in a certain interval to make an appropriate specific function. Take known values at these points, and use the value of this particular function as an approximation of the function f (x) at other points in the interval and this method is called interpolation. If this particular function is a polynomial, it is called an interpolation polynomial. The positional deviation of the pixels imaged by two cameras in the same scene of the disparity map. Because the two binocular cameras are usually placed horizontally, the positional deviation is generally reflected in the horizontal direction. For example, the X point in the scene is the x coordinate on the left camera, and the image on the right camera is the (x+d) coordinate, and d is the value of the x coordinate point in the parallax map. As shown in Table 2 for database data comparison:

**Table 2. Comparison of database efficiency of different detection methods**

| Video length | Number of screens | Bit rate | DTW algorithm | Video coding algorithm | Algorithm |
|---|---|---|---|---|---|
| 1 | 24 | 15550 | 2.14 | 1.17 | 0.89 |
| 2 | 24 | 15550 | 1.89 | 1.25 | 1.27 |

A comparison of the motion segmentation method based on video coding and the efficiency of the motion detection algorithm in this article in the KTH database. It can be seen that the efficiency of the motion detection algorithm is not only affected by the length of the video, but the higher the bit rate, the longer the execution time. In addition, the efficiency of the algorithm in this paper is more advantageous overall, and it is more suitable for real-time detection of actions. At the same time, for the motion at the edge of the image, the algorithm in this paper also has a good detection effect. The comparison data between the algorithm in this paper and other algorithms is shown in Figure 6:

**Figure 6. Comparison of data collection between the algorithm in this article and other algorithms**




In summary, in view of the advantages and disadvantages of the above three algorithms, in order to achieve real-time target detection effects, it is of great significance to study the background subtraction method and the inter-frame difference method. The inter-frame difference method is

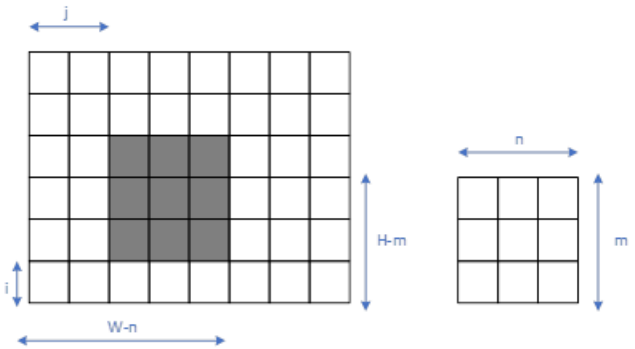
usually used to detect moving targets that change regularly in the video. The change of the target will cause the change of the corresponding pixel in the image. Therefore, the gray scale is different at the same position between adjacent frames, and the pixel difference of adjacent frames is the target. A reliable background model is the key to background subtraction.

## 3. EXPERIMENT AND SIMULATION

### 3.1 Image Matching Tracking

The basic operation process of image matching tracking is summarized as follows: First, it is necessary to initially determine how to directly express a pending target in image matching (Broadly speaking, it is not necessarily a pending target with a clear characteristic meaning in the ordinary technical sense), that is, how to reflect the candidate points of the basic characteristics of the target to be determined. The second is how to solve the problem of the corresponding nature of the target, and the process of solving the target can generally be considered as starting from a candidate point target on a pending image. Target comparison and matching between images is the comparison of targets between two or more images in a group, from which different shooting objects in the group can be found. Then taking out the partial magnification according to the location for feature extraction to capture the information face of the image. The specific model block is shown in Figure 7:

**Figure 7. Schematic diagram of image matching**




The above method of completely and quickly searching the global template optimal matching allows you to quickly search and get a global optimal template matching, but it requires a lot of calculations. Because global template matching requires optimal matching of nodes at (w-m+1)×(h-n+1) positions that have no reference value, and most of them are nodes that are not optimally matched in actual situations. It needs to do some useless work, the calculation amount and speed are slow, so it can not fully meet the requirements of real-time data detection (Matilla, 2020; Hashmi, 2020). To get a fast and accurate matching method, first analyzing the key elements of image matching. The matching algorithms in different application contexts are composed of the following three elements:

(1) Image feature combined space. It is generally composed of a combination of multiple image features that participate in the matching at the same time. The selected image feature space can improve the performance of the matching system, reduce the image search speed space, and reduce the uncertainty of the image noise, etc., which have a direct impact on the algorithm of the

matching system. During the matching algorithm, you can choose to use global image features, local image features, or a combination of the two.

(2) Similarity measurement. Classical mean square similarity error measurement functions include the maximum mean square cross-correlation measurement function (MR), the minimum maximum mean square level error measurement function (mse), and the minimum mean square average absolute value error and real value measurement function (MAD), etc. (Cui, 2020; Kolykhalova, 2020).

(3) Optimal search of various transform space parameters. This search method strategy is to quickly find the optimal estimation of various transformation space parameters such as image translation, rotation, etc. in a search image space using various suitable search calculation methods. Then making the image similarity between the two images or the image after rapid transformation the greatest. This search method strategy mainly includes endless annealing search, hierarchical annealing search, simulated annealing search algorithm, powell method and direction search acceleration method, dynamic space planning method, genetic algorithm and dynamic neural network. The comparison of the algorithm detection results is shown in Table 3:

**Table 3. Comparison of detection results of various algorithms**

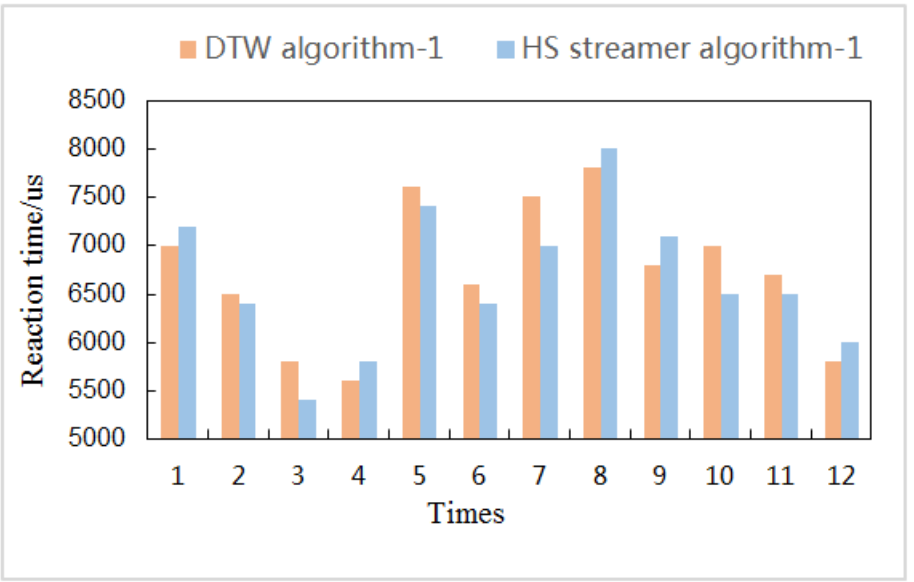
| method | Recall rate | False positive rate | Overall accuracy |
|---|---|---|---|
| HS optical flow | 0.64 | 0.03 | 0.56 |
| Mahmoudi optical flow | 0.58 | 0.05 | 0.66 |
| Gaussian mixture model | 0.61 | 0.012 | 0.71 |
| Haines background modeling | 0.68 | 0.013 | 0.8 |
| Movement energy flow | 0.76 | 0.008 | 0.78 |

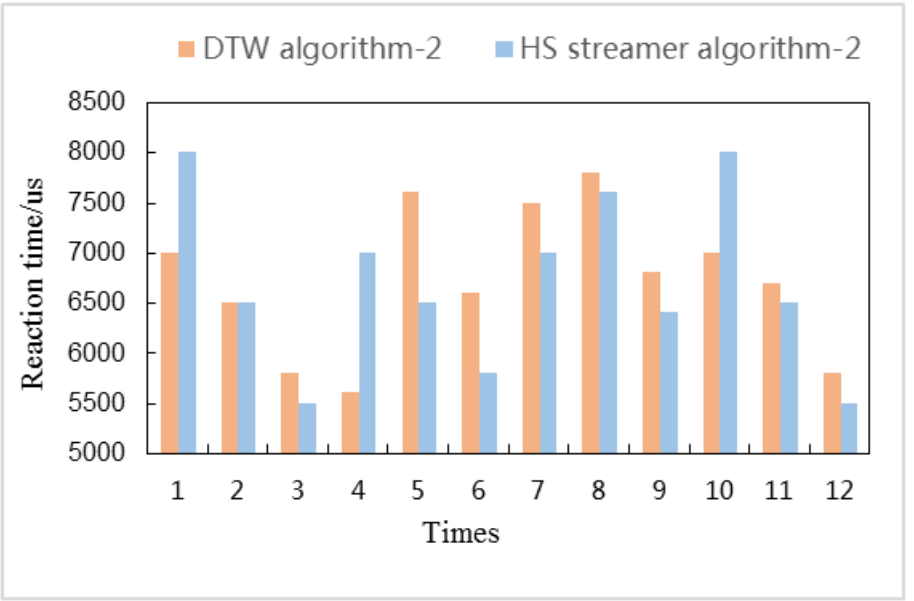
## 3.2 Human Motion Capture Experiment

Motion energy flow explores the law of image changes caused by motion, so it can be applied to action recognition. This paper uses the extracted motion energy flow as the action feature to further verify the efficiency of the motion energy flow in human body action recognition. In the KTH database, according to the default settings recommended by the original author, a data set of 16 individuals is selected as the training sample, and the remaining data set is used as the test sample. The k-means clustering algorithm is used to directly cluster the extracted motion energy flow features into 4000 categories. K-Means algorithm is a simple iterative clustering algorithm that uses distance as a similarity index to find K classes in a given data set. And the center of each class is obtained based on the mean value of all the values in the class, and the center of each class is described by the cluster center. And then using the classic bag-of-words model to decode to complete the motion characterization; when classifying the motion, the classic support vector machine architecture is used for classification (Wu, 2020; Yoo, 2020).

It can be seen from the experimental results that the movement energy flow can better realize the movement recognition with large difference, and it also has a high discrimination rate for the movement with large difference in speed, such as running and walking. Compared with the classic HS optical flow method, the motion energy flow is more accurate in the information representation of human actions, which indirectly proves that its overall recognition effect is better than similar algorithms. The sensitive response time of the algorithm to actions is shown in Figure 8:

**Figure 8. Response time of DTW algorithm and HS streamer algorithm to human motion capture**




In practical applications, the latter two metrics are used more frequently because of the smaller amount of calculation. The latter two metrics can be slightly improved to further improve the matching speed: from the calculation formulas of DTW and HS, it can be seen that the calculation process of the metric is actually the accumulation process of the gray value difference of each pixel. Since MSE and MAD are squaring operations and absolute value operations, respectively, the accumulation of differences is gradually increasing. The smaller the accumulated error, the more similar the matching template and the corresponding sub-image. Conversely, if the accumulated error is large and exceeds

a certain threshold, it means that the matching template does not match the corresponding sub-image, and it is meaningless to continue the calculation. Therefore, in the process of calculating the mismatch metric, if DTW and HS are greater than the preset threshold, the calculation is stopped, so that the matching time can be shortened.

It can be seen that the movement energy flow has certain difficulty in recognizing actions with little difference in movement speed, and the recognition rate of actions with big difference in speed and movement type is higher. Comparing the average recognition rate of motion energy flow with the current popular algorithms, it can be seen that its accuracy is relatively high, and energy flow is the concept of energy conduction. The difference in the form of energy production in the organization, the difference in the energy production capacity of the organization, and the energy exchange between humans and the environment can all lead to the flow of human energy. The energy flow is the fundamental condition to ensure the vitality of the human body. The specific results are shown in Table 4:

**Table 4. Average recognition rate of motion energy flow and current popular algorithms**

| Method | Average recognition rate |
|---|---|
| Goudelis etc. | 94.12% |
| Laptev etc. | 89.25% |
| SIFT + bag of words model + support vector machine | 92.14% |
| HS optical flow + bag of words model + support vector machine | 95.26% |
| Sports energy flow + bag of words model + support vector machine | 95.44% |

The skeleton is extracted from the binary image of target detection, and the main idea of skeleton extraction is: set the refinement conditions in advance, traverse the pixels in the input binarized image in turn, and determine whether each point meets the set conditions. If it is satisfied, then mark, after traversing the whole image, all the mark points are finally deleted, and the first pixel stripping is completed. Then the above process is repeated continuously until the whole image no longer changes, and the target skeleton image is obtained at this time. The algorithm maintains the topological structure of the target area, while reducing the complexity of subsequent processing.

### 3.3 Human Motion Capture System Design

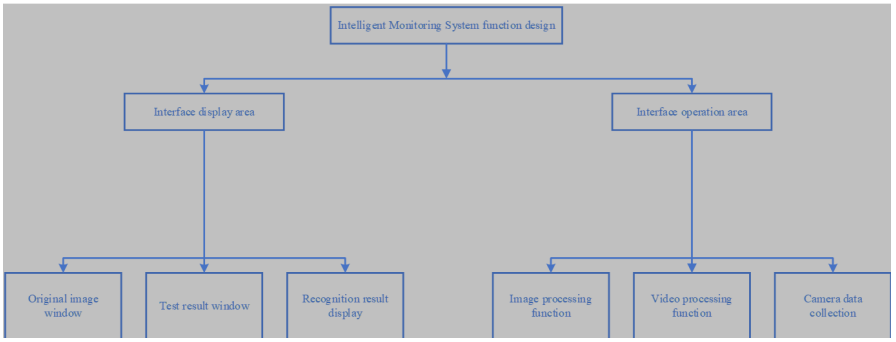
The algorithm first extracts and normalizes each satellite human body posture feature map target that can display the fast motion of the human body in the video by using G and GMM respectively, and then extracts a multi-pole posture feature for each feature target. Before the foreground detection, the background is trained, and a Gaussian mixture model is used to simulate each background in the image, and the number of Gaussian mixtures for each background can be adaptive. Then in the test phase, GMM matching is performed on the new pixel. If the pixel value can match one of the Gaussians, it is considered as the background, otherwise it is considered as the foreground. Since the GMM model is constantly being updated and learned throughout the process, it has certain robustness to dynamic backgrounds. And a statistical model of satellite human body posture feature target posture map analysis is established. Finally, by using g and svm to extract a training sample respectively, the first statistical model classification is performed on the human pose feature target pose map in each frame; secondly, the seven algorithms are based on the human body image similarity between two adjacent frames in a video, and the automatic similarity adaptation is used to accurately distinguish and then allocate a video for accurate judgment to determine the first statistical model classification in the time and angle period. After that, the weight of the statistical classification result

is carried out for the second time of the judgment result; finally, seven algorithms are used to count a secondary statistical classification used in the discrimination time period, and after the judgment result is judged, the secondary statistical classification is performed on it and the judgment result is classified three times.

This article uses a PC with Windows 7 system, 6GB memory, dual-core Intel Core i5 M 430@2.27GHz as the programming environment, and uses OpenCV2.4.11 to design and implement the algorithm. Because the code portability based on OpenCV is good, and in order to improve the practicality of this algorithm, this paper uses MFC and DirectShow streaming media processing library to jointly complete the design of the human body gesture recognition system.

According to demand analysis, the functional modules of the system are shown in Figure 9:

**Figure 9. Classification of system function modules**




The system needs to meet the following requirements:

(1) The system has a good human interaction interface, and the area allocation is reasonable, which is convenient for users to operate;
(2) With image reading and edge detection functions;
(3) With functions of video reading, human moving target detection and gesture recognition;
(4) With camera data collection function.

According to system design requirements, it is divided into display area and operation area. Among them, the display area includes the original image or video display window, the moving target detection result display window, and the real-time gesture recognition result window, which can respectively realize static and dynamic display. The operation area includes an image processing operation area, a video processing operation area and a video acquisition area. The image processing operation area can realize image reading and edge detection; the video processing operation area can realize video reading, moving target detection and gesture recognition; the video collection area can realize the video collection of the computer camera or external video collection equipment. Each dynamic operation can realize the timely pause and resume function, which is convenient for the user to view the target and posture at a certain point in time.

Test each function of the system, and the results are as follows:

(1) Image processing function. Taking the classic image lena as an example, the system can clearly display the original image and the result of its edge detection.

(2) Video processing, detection and recognition functions. At this time, it is the pause processing screen of the lena_walk1.avi video in the Weizmann database. The system displays the video to be processed, the result of moving target detection, and the result of gesture recognition respectively. The advantage of the algorithm in this paper is that it has low computational complexity and can be used in real-time systems. It can be seen that the result output of moving target detection and gesture recognition is completely consistent with the original video, which meets the real-time requirements of video processing.

## 4. EXPERIMENTAL RESULTS AND ANALYSIS

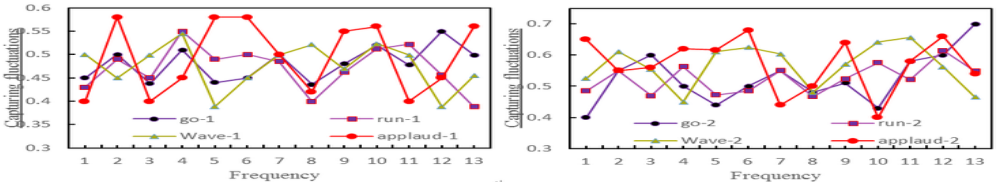
### 4.1 Experimental Data Preprocessing

The image size (pixel resolution) collected by the camera in this article is 320×240. For the detection of moving human targets, the experiment first initializes the parameters of the camera and collects background images in the video range where the moving target will appear. Since the background changes with the light, in order to obtain a relatively stable background, the program will continuously collect multiple frames of background and average the method to obtain the initial background image. Then, the three-frame difference method is used to detect whether a moving target enters the video range. If a foreground target enters, the start-up procedure, the target detection procedure described above, is started to detect. That is, the target binary image obtained after processing by the three-frame difference method, filtering and mathematical morphology. The edge detection map can extract a clearer edge contour, and the noise removal effect is good. It can be seen from the experimental results that the method described in this paper can detect the clear outline of the moving target well, and can judge whether there is a moving target entering the video range. If a moving human target enters the cheek range, start the moving target tracking program.

From the action recognition results, it can be concluded that within a certain application range, the action recognition method of this application software has the following characteristics:

(1) The motion recognition method based on motion energy flow has better recognition effect in application scenes with simple background, and better recognition effect for obvious motions such as bending over. For similar motions such as waving and applauding, the recognition effect is poor, and the vibration amplitude of the motion recognition effect is relatively small.
(2) Action recognition based on gradient transformation feature algorithm also has a good effect in application scenarios with simple backgrounds. However, the recognition effect of walking motions is poor, and the recognition effect of waving hands is not ideal. Its motion recognition effect has a certain degree of randomness and volatility.
(3) Action recognition based on image potential energy difference template has relatively poor results in application scenarios with simple backgrounds. There is a certain probability of misrecognition for walking, running, waving, applauding and bending over, and its action recognition effect is also volatile. The specific fluctuation data is shown in Figure 10:

**Figure 10. The system captures the fluctuation value of various actions**



Through the combination of gray channel and frame difference channel, the accuracy of the network's classification of human actions is improved. In this paper, in the specific network structure design of the dual-channel 3D convolutional neural network model, many experiments are carried out on the network structure and parameter settings, and the influence of the network structure on the network performance is analyzed, and a better network structure on the KTH data set is obtained and parameters. While ensuring the classification effect, the number of parameters is greatly reduced, the training efficiency is improved, and the classification accuracy is improved. One is that the three postures have similar or identical posture forms during the movement change process, which causes misjudgment; the other is that noise affects the incomplete detection of moving human targets, resulting in unrepresentative extracted features and deviations in recognition results .

### 4.2 Results and Discussion

This chapter conducts simulation experiments on the feature extraction and performance analysis of the above-mentioned types of prediction and training video models on a UCF-11 data set. The UCF-11 data set only contains 11 user behavior feature categories. At the same time, similar to the UCF-101 data set, it is also a video from daily life collected from YouTube, but the scale is small. All final models can use the connection layer vector located in the penultimate layer as a feature vector after dimensionality reduction of the model extracted by the final model. The performance of several types of CNN pre-training models is compared with the experimental results.It can be seen from the experimental results that on the UCF-11 data set, the use of pre-training models for feature extraction can achieve better classification results. Among them, the InceptionResNetV2 model has the highest accuracy rate, which has a very deep network structure, but the addition of the residual module makes the computational complexity of the network model not increase drastically as the number of network layers deepens. Deeper video network feature model analysis can be used to better analyze, extract and derive network features in various video network image processing, which makes the feature generalization analysis capability of video network stronger.

In this paper, by using the dual-channel 3D convolutional human neural network technology, it can achieve 92.5% of the accuracy of human body capture features at the same time, which is better than the method of using multiple traditional feature extraction techniques given at the same time in the related literature. A full-automatic human action feature recognition system model that can recognize a variety of human action behavior features based on the extracted human features. In the collection of UCF-11 data, the reason for the poor performance of traditional features is that the UCF-101 data set has many action categories and the video shooting environment is complex. Feature extraction methods are difficult to extract features that can accurately describe the type of motion. And if using this kind of deep machine learning, it can automatically extract many general, abstract and middle-level data features from a large amount of feature data, and achieve high feature recognition rate and accuracy.

## 5. CONCLUSION

Based on the results of data analysis, this paper compares three more classic indoor motion model target pose detection algorithms: the traditional frame gap scoring method, the background motion subtraction, addition, and division method, and the optical convection method. In the end, the Gaussian background motion subtracting addition and division technique based on traditional mixed Gaussian background motion modeling is selected as the main indoor target motion detection algorithm in this article. Aiming at the problems of the traditional new GMM detection algorithm, and the motion model target frame, it is easy to cause the motion smear tracking phenomenon and the poor detection performance of the traditional algorithm during the detection. A new GMM detection algorithm that can combine the Ronski Boolean function and the traditional frame gap scoring method is proposed. The improved algorithm of this technology increases the conditions for updating the attitude parameter

values of the motion model. The attitude judgment is performed on the correlation changes of different spatial attitudes between attitude pixels at the same coordinate node in two adjacent motion frames. It uses the determinant in the Ronsky matrix as the basis for the calculation of posture discrimination. In this way, it determines whether the posture pixels have changed or not, and solves the problem that the skill training and technical learning of each posture pixel based on the traditional algorithm GMM is limited to only one time domain. The main indoor posture models are established and classified according to the common indoor motion posture pixels. The main posture models include fast walking, jumping and fast running. A total of 3716 valid samples of each pose are selected from 12890 sample images in the standard video library, and a multi-class classifier based on SVM is trained by extracting the feature vectors of the valid samples. The improved algorithm in this paper can achieve the expected design goals, the detection effect of moving targets is better, the accuracy of human body gesture recognition is higher, and the experimental effect based on the standard video library is better, which can reach about 92.5%. However, there are still some limitations and shortcomings, the main disadvantage is that the classification of human motion capture is not comprehensive enough, because this article only focuses on walking, running, waving and applauding and there is still room for improvement to solve the complex situation in the actual video capture.

## FUNDING AGENCY

Open Access Funding for this article has been covered by the authors of this manuscript.

## REFERENCES

Ashhar, K., Noor-A-Rahim, M., Khyam, M. O., & Soh, C. B. (2019). A Narrowband Ultrasonic Ranging Method for Multiple Moving Sensor Nodes. *IEEE Sensors Journal*, *19*(15), 6289–6297. doi:10.1109/JSEN.2019.2909580

Bernd & Eckenfels. (2018). Vision-sensoren vereinfachen die einrichtung von pick-and-place-anwendungen: direkte kommunikation mit dem roboter. *Elektro-Automation: Elektrotechnik + Elektronik Inder Industrie, 71*(6), 70-71.

Chang, K. C., & Seow, Y. M. (2019). Protective Measures and Security Policy Non-Compliance Intention: It Vision Conflict as a Moderator. *Journal of Organizational and End User Computing, 31*(1), 1-21.

Choi, Sungjoon, & Songhwai. (2016). Vision-Based Coordinated Localization for Mobile Sensor Networks. *IEEE Transactions on Automation Science and Engineering, 13*(2), 611-620.

Cui, Q., Sun, H., Kong, Y., Zhang, X., & Li, Y. (2021). Efficient Human Motion Prediction using Temporal Convolutional Generative Adversarial Network. *Information Sciences*, *2020*(545), 427–447. doi:10.1016/j.ins.2020.08.123

Ding, S., Qu, S., Xi, Y., & Wan, S. (2019). Stimulus-driven and concept-driven analysis for image caption generation. *Neurocomputing*.

Harfmann, B. (2016). Ensuring safety through sensor solutions. *Beverage Industry*, *107*(7), 42–43.

Hashmi, M. A., Riaz, Q., & Zeeshan, M. (2020). Motion Reveal Emotions: Identifying Emotions From Human Walk Using Chest Mounted Smartphone. *IEEE Sensors Journal*.

Heydari, M. J., & Ghidary, S. S. (2019). 3D Motion Reconstruction from 2D Motion Data Using Multimodal Conditional Deep Belief Network. *IEEE Access*.

Holt, C., Seacrist, T., & Douglas, E. (2019). The effect of vehicle countermeasures and age on human volunteer kinematics during evasive swerving events. *Traffic Injury Prevention*, *21*(1), 1–7. PMID:31750733

Hu, Cao, & Yang. (2021). Performance Evaluation of Optical Motion Capture Sensors for Assembly Motion Capturing. *IEEE Access*.

Ammann. (2020). Human motion component and envelope characterization via wireless wearable sensors. *BMC Biomedical Engineering*, *2*(1), 1–15. PMID:32903362

Jiménez Bascones, Graña, & Lopez-Guede. (2019). Robust labeling of human motion markers in the presence of occlusions. *Neurocomputing, 353*(11), 96-105.

Kolykhalova, K., Gnecco, G., & Sanguineti, M. (2020). Automated Analysis of the Origin of Movement: An Approach Based on Cooperative Games on Graphs. *IEEE Transactions on Human-Machine Systems,* 1-11.

Li, Y., Xiao, J., Xie, D., Shao, J., & Wang, J. (2019). Adversarial learning for viewpoints invariant 3D human pose estimation. *Journal of Visual Communication and Image Representation*, *58*(JAN), 374–379. doi:10.1016/j.jvcir.2018.11.021

Liu, S. Q., Zhang, J. C., & Li, G. Z. (2020). A Wearable Flow-MIMU Device for Monitoring Human Dynamic Motion. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Ma, H., Yan, W., & Yang, Z. (2019). Real-time Foot-Ground Contact Detection for Inertial Motion Capture based on an Adaptive Weighted Naive Bayes Model. *IEEE Access*.

Matilla, R. S., Chatzilygeroudis, K., & Modas, A. (2020). Benchmark for Human-to-Robot Handovers of Unseen Containers With Unknown Filling. *IEEE Robotics and Automation Letters*, *5*(2), 1–1.

Poornima, J., Vishnupriyan, J., & Vijayadhasan, G. K. (2020). Voice Assisted Smart Vision Stick for Visually Impaired. *International Journal of Control and Automation*, *13*(2), 512–519.

Ranjan, A., Hoffmann, D. T., Tzionas, D., Tang, S., Romero, J., & Black, M. J. (2020). Learning Multi-human Optical Flow. *International Journal of Computer Vision*, *128*(4), 873–890. doi:10.1007/s11263-019-01279-w

Shanthi, K. G. (2020). Smart Vision using Machine learning for Blind. *International Journal of Advanced Science and Technology*, *29*(5), 12458–12463.

Shimada, S., Golyanik, V., Xu, W., Pérez, P., & Theobalt, C. (2021). Neural monocular 3D human motion capture with physical awareness. *ACM Transactions on Graphics*, *40*(4), 1–15. doi:10.1145/3450626.3459825

Souza, C., Gaidon, A., & Cabon, Y. (2020). Generating Human Action Videos by Coupling 3D Game Engines and Probabilistic Graphical Models. *International Journal of Computer Vision*, *128*(5), 1505–1536. doi:10.1007/s11263-019-01222-z

Vision Based Localization for Multiple Mobile Robots Using Low-cost Vision Sensor. (2016). *International Journal of Handheld Computing Research, 7*(1), 12-25.

Wan, S., Qi, L., Xu, X., Tong, C., & Gu, Z. (2019). Deep Learning Models for Real-time Human Activity Recognition with Smartphones. *Mobile Networks and Applications*, 1–13.

Wei, H., & Kehtarnavaz, N. (2019). *Semi-supervised faster rcnn-based person detection and load classification for far field video surveillance*. Academic Press.

Wu, L., Alqasemi, R., & Dubey, R. (2020). Development of Smartphone-Based Human-Robot Interfaces for Individuals with Disabilities. *IEEE Robotics and Automation Letters*.

Xu, Z., Chang, W., & Zhu, Y. (2020). Building High-fidelity Human Body Models from User-generated Data. *IEEE Transactions on Multimedia*.

Xuan, T. N., Ngo, T. D., & Le, T. H. A. (2019). Spatial-temporal 3D Human Pose Reconstruction Framework. *Journal of Information Processing Systems*, *15*(2), 399–409.

Yoo, C. H., Ji, S. W., & Shin, Y. G. (2020). Fast and Accurate 3D Hand Pose Estimation via Recurrent Neural Network for Capturing Hand Articulations. *IEEE Access*.

Zidek, K., Vasek, V., Pitel, J., & Hosovsky, A. (2018). Auxiliary device for accurate measurement by the smartvision system. *Mm Science Journal*, *2018*(1), 2136–2139. doi:10.17973/MMSJ.2018_03_201722

*Fang Yu, associate professor, master, graduated from Northeast Normal University in 2005. Now she works in Jilin Sport University. Her research direction is physical education. E-mail: yufang1559@163.com.*