Research on Key Feature Extraction and Position Accurate Tracking Based on Computer Vision Image

Yan Sun
Beijing University of Posts and Telecommunications
10 Xueyuan Rd. Beijing, China
Sunyan2015@bupt.edu.cn

Abstract: At present, the difficulty in video image recognition mainly focuses on video image segmentation and video image recognition. In video image segmentation, the effect of segmentation has affected by changes in lighting. The active vision technology of coded structured light illumination is used, and only one image is used. Under the condition that the camera system parameters are unknown, it is not necessary to calibrate the three-dimensional scene in the real three-dimensional Euclidean space for position. The human-computer interaction based on video image recognition in this paper is an important research topic of current intelligent human-computer interaction. Because vision-based video image recognition used a camera to capture video image information without the need to purchase expensive equipment, it is more natural and convenient to operate and meets the requirements of human-centeredness and freedom. Computer vision-based video image research has important theoretical value and application prospects, so we believe that the research on video image recognition will receive more attention and development.

1. Introduce
Since the new century, with the continuous improvement of the computer level, human daily life, learning and work have become more and more inseparable from the computer. In the use of computers, people have put forward higher requirements on both the performance and intelligence of computers [1]. The need for human development has not achieved by traditional means. The study of harmonious human-computer interaction is an important topic in the field of computer research in recent years, and one of the main contents is video image recognition technology. The active vision technology of coded structured light illumination is used, and only one image is used. Under the condition that the camera system parameters are unknown, it is not necessary to calibrate the three-dimensional scene in the real three-dimensional Euclidean space. The two main problems that must have solved in theory are as follows. First, what kind of encoding method has used to encode the projector, and the surface of the three-dimensional scene space covered by an image can have characterized to make each feature point on the surface of the space. All can uniquely identify to solve the matching point identification problem. At present, there are many methods related to coding, such as the time-space coding method, the black and white binary pattern space coding method, the colour spectrum coding method, the colour pseudo-random sequence narrow inter-code method, and the like. Because the pseudo-random sequence has good window unique characteristics, it can use to uniquely identify each feature point on the surface of the space and solve the matching point identification problem [2]. Compared with the traditional structured light projection system scheme, the method of projecting the pseudo-random coded image by the projector has a low cost, high accuracy of measurement matching, does not require a high-accuracy
scanning mechanism, simple operation, and can change the coding pattern to adapt to different coding ways and other advantages. Therefore, the active vision system adopts a pseudo-random coded structured light illumination active vision technology [3]. The other is how to extract the feature corners on the coding pattern. This is the core of the characterization technology in the visual detection and the basis for matching the reconstructed 3D model. Due to the accurate location of point features and the ability to provide 3D information, many people have done research on point feature of extraction algorithms.

In the past, machine-centric human-computer interactions have conducted, and mouse and keyboard have generally used as input devices. Humans generally communicate through language. In addition, they can communicate through expressions and video images. Therefore, we should have inspired by human communication. We must turn the centre of human-computer interaction into user-centred, let the machine feel the body language of the user, and the interaction between the human and the machine becomes more convenient and natural. This kind of human-computer interaction method that imitates people's living habits will get more extensive market applications. The main advantage of video image recognition is that it is easy to operate and has become the hottest topic in the field of human-computer interaction. The future human-computer communication will become more convenient and natural because of video image recognition technology. According to different input methods, video image recognition can have divided into data glove based video image recognition and computer vision based video image recognition. When performing video image recognition using the data glove, the computer cannot directly recognize the video image shape and its motion trajectory. You must wear a data sensor that looks similar to your glove so that your computer can obtain the data you want [4, 5]. Through research, it has found that although its recognition effect is very good, it is not perfect, and there are some defects. Such as: the user must carry data gloves when using, very inconvenient to operate, and data gloves are expensive, can be used during the experiment, but not suitable for promotion. Computer vision-based video image recognition users can use their own hands to control the corresponding device, without any additional hardware, and the cost is low and can be widely used.

2. The Image enhancement

Image enhancement (ie, image preprocessing) is used to eliminate noise and other irrelevant information and enhance the detectability of useful information in the image, thereby simplifying the information in the image. In this way, the video image segmentation process can be simplified. The computer cannot directly process the acquired video image, but it needs to be converted into a digital image before it can be processed [6]. When the image is regenerated, transmitted and changed, it will be interfered and influenced by the outside world. The image quality of the image will also be affected. Therefore, there is a change. Common methods for image enhancement (preprocessing) are: smoothing and sharpening.

Image translation means that all the points in an image have shifted along the x-axis in the horizontal direction and along the y-axis in the vertical direction. The images before and after the translation are the same. We set point $P_0(x_0, y_0)$ translate in the x and y directions, respectively $\Delta x$, $\Delta y$, so we can get the point $P(x, y)$.

$$
\begin{align*}
x &= x_0 + \Delta x \\
y &= y_0 + \Delta y
\end{align*}
$$

The relationship between the points on the image before and after the transformation and the relationship between the points $P_0(x_0, y_0)$ and $P(x, y)$ can be expressed by the following matrix.
You can calculate the inverse of the above equation by finding the inverse of the transformation matrix:

\[
\begin{bmatrix}
0 & 1 & \Delta x \\
1 & 0 & \Delta y \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
x_0 \\
y_0 \\
1
\end{bmatrix}
= \begin{bmatrix}
x \\
y \\
1
\end{bmatrix}
\]

(2)

If a point is not in the original image, its pixel value is set to 255 or 0. After the translation, each pixel can find its corresponding point in the original image. In order to avoid the content of the original image being removed from the display area, the new size of the translated image needs to be calculated according to the actual situation [7]. If you increase the width $|\Delta x|$ of the newly generated image and increase its height $|\Delta y|$, you can display the removed part of the image.

For the feature extraction part to perform later, the zoom invariance of the video image needs to have taken into consideration. Therefore, the scaling operation needs to perform in this paper. While scaling, we must ensure that the entire image has uniformly scaled to ensure the image quality. The commonly used method for scaling is linear interpolation. That is, a linear function is used to make pixels in the original image correspond to the points in the new image.

At the time $f_x = f_y$, the scaling of the image on the x-axis and y-axis is the same. This scaling is called full scaling. Now, the scaling of the image will cause the position of the pixels between the original pictures to change and become distorted. Suppose the points $P(x_0, y_0)$ in the original image are scaled and the corresponding point in the new image is $P(x, y)$, and the correspondence between them is shown in Figure 1.

![Figure 1. The Image of Pixel zoom](image)

Figure 1. The Image of Pixel zoom

The relationship between two points before $P(x_0, y_0)$ and $P(x, y)$ after scaling can be represented by the following matrix.
The inverse operation is the following formula.

\[
\begin{bmatrix}
x
\end{bmatrix} = \begin{bmatrix}
f_x & 0 & 0 \\
0 & f_y & 0 \\
0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
x_0 \\
y_0 \\
1
\end{bmatrix}
\]

(5)

The inverse operation is the following formula.

\[
\begin{bmatrix}
x_0 \\
y_0 \\
1
\end{bmatrix} = \begin{bmatrix}
\frac{1}{f_x} & 0 & 0 \\
0 & \frac{1}{f_y} & 0 \\
0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
x \\
y \\
1
\end{bmatrix}
\]

(6)

\[
\begin{align*}
x_0 &= \frac{x}{f_x} \\
y_0 &= \frac{y}{f_y}
\end{align*}
\]

(7)

When a new image has generated by scaling, the pixels in the pixels may not be able to find corresponding points in the original image [8]. In order to alleviate this situation, interpolation must have performed.

3. Research on Key Feature Extraction and Position Accurate Tracking Based on Computer Vision Image

3.1 Colour feature extraction

We know that colour is the most direct visual feature of our human beings to perceive images or graphics. The characteristics of colour have described by the image and the surface colour of the image region. It has integrity and stability. The extraction methods of colour features are colour histograms, color sets, etc. To understand it from the point of view of probability, the frequency of occurrence of grayscale can have regarded as the probability of its appearance. Such a histogram corresponds to the probability density function PDF, and the probability distribution function is the sum of the accumulated histograms, that is, the integral of the probability density function [9]. Finally, all requirements have fulfilled by writing source code in a C++ language environment. Figure 2 shows the principle of video image recognition.

![Figure 2. Principle of video image recognition](image)

3.2 Feature Extraction
The shape feature is one of the characteristics of the object's essence, and it is also the visual perception feature of human's intuitive perception of one-dimensional contour. However, the formula is different from the features of color and texture. It needs the premise of image segmentation. Therefore, compared with the underlying features, the extraction and description of shape features is relatively difficult. Under normal circumstances, two types of methods respectively represent the shape features, one is the outline feature and the other is the area feature. The area feature of the image has related to the entire shape area, and the outline feature of the image is mainly focused on the external boundary of the object. Figure 3 shows the video image positioning principle.

![Figure 3. The Video image positioning principle](image)

The projection \((u_l, v_l)\) of the spatial point \((X, Y, Z)\) on the left camera image plane can be expressed as follows.

\[
\begin{bmatrix}
  u_l \\
  v_l \\
  1
\end{bmatrix}
= k_1
\begin{bmatrix}
  m_{11} & m_{12} & m_{13} & m_{14} \\
  m_{21} & m_{22} & m_{23} & m_{24} \\
  m_{31} & m_{32} & m_{33} & m_{34}
\end{bmatrix}
\begin{bmatrix}
  X \\
  Y \\
  Z \\
  1
\end{bmatrix}
\] (8)

The projection \((u_r, v_r)\) of the spatial point \((X, Y, Z)\) on the right camera image plane can be expressed as follows.

\[
\begin{bmatrix}
  u_r \\
  v_r \\
  1
\end{bmatrix}
= k_r
\begin{bmatrix}
  m'_{11} & m'_{12} & m'_{13} & m'_{14} \\
  m'_{21} & m'_{22} & m'_{23} & m'_{24} \\
  m'_{31} & m'_{32} & m'_{33} & m'_{34}
\end{bmatrix}
\begin{bmatrix}
  X \\
  Y \\
  Z \\
  1
\end{bmatrix}
\] (9)

![Figure 4. The Filtering method and implementation](image)

Figure 4 gives the filtering method and implementation. Boundary feature method is to obtain the image shape parameters through the description of the boundary features. The Hough transform detection parallel line method and the boundary direction of histogram method are classical methods. The Hough transform is a method that uses the global characteristics of the image to connect the edge...
pixels to form the closed boundary of the area. The basic idea is the duality of the point-line; the boundary direction of histogram method firstly obtains the edge of the image by the differential image, and then makes as for the histogram of the edge size and direction, the usual method is to construct an image grayscale gradient direction matrix.

### 3.3 Video image recognition and positioning results

The security requirements of power grids in different regions are different. Some power grid economic dispatches can meet the requirements by considering the safety constraints of normal operating conditions, but some power grids also need to take into account the economic dispatch of line security constraints. Therefore, the economics of the grid under different security levels cannot have generalized [10]. Therefore, it is necessary to perform security level assessment on the operation status of each optimal solution and classify the optimal solutions belonging to the same security level. This facilitates decision makers to be more economical at the same level of safety. Finally, the optimal compromise solution has obtained from these solutions according to the multi-attribute decision method or the network loss rate. This solution takes into account the requirements of the system on security and economy. The overall optimization process has shown in Figure 5.

![Figure 5. The video image recognition and positioning results](image)

### 4. Conclusions and Future Work

This paper focuses on basic video image processing and fast positioning respectively, proposes operation platform applications for feature extraction and other technical operations, and based on feature extraction, polar coordinate features, motion features, and structural features come together. Moreover, these features have the ability to express in different directions. Applying them together can greatly increase the expressiveness of the behaviour. In addition, the use of fused features allows for a very large degree of differentiation between very similar behaviours and avoids misidentification. The use of fusion features makes the experimental results more discriminating and robust, and its processing efficiency can reach more than 90% or even higher.

### References

[1] Dell A I, Bender J A, Branson K, et al. Automated image-based tracking and its application in ecology[J]. Trends in ecology & evolution, 2014, 29(7): 417-428.

[2] Yuan C, Sun X, Lv R. Fingerprint liveness detection based on multi-scale LPQ and PCA[J]. China Communications, 2016, 13(7): 60-65.

[3] Huang A S, Bachrach A, Henry P, et al. Visual odometry and mapping for autonomous flight using an RGB-D camera[M]/Robotics Research. Springer, Cham, 2017: 235-252.

[4] Zhang K, Liu Q, Wu Y, et al. Robust visual tracking via convolutional networks without training[J]. IEEE Transactions on Image Processing, 2016, 25(4): 1779-1792.

[5] Liang R Z, Shi L, Wang H, et al. Optimizing top precision performance measure of content-based image retrieval by learning similarity function[C]/Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE, 2016: 2954-2958.
[6] Ma C, Huang J B, Yang X, et al. Hierarchical convolutional features for visual tracking[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 3074-3082.

[7] Rautaray S S, Agrawal A. Vision based hand video image recognition for human computer interaction: a survey[J]. Artificial Intelligence Review, 2015, 43(1): 1-54.

[8] Yang J, Park M W, Vela P A, et al. Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future[J]. Advanced Engineering Informatics, 2015, 29(2): 211-224.

[9] Forster C, Pizzoli M, Scaramuzza D. SVO: Fast semi-direct monocular visual odometry[C]//Robotics and Automation (ICRA), 2014 IEEE International Conference on. IEEE, 2014: 15-22.

[10] Koch C, Georgieva K, Kasireddy V, et al. A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure[J]. Advanced Engineering Informatics, 2015, 29(2): 196-210.