Statistical Research Based on Machine Vision: A Review

Xin Xiong¹, Li Huang², Yangwei Cheng³, a

¹Infrastructure and Logistics Management, Wuhan University of Science and Technology, Wuhan, Hubei, 430081, China
²College of Computer Science and Technology, Wuhan University of Science and Technology, Hubei, Wuhan 430081, China
³Key Laboratory of Metallurgical Equipment and Control Technology of Ministry of Education, Wuhan University of Science and Technology, Hubei, Wuhan 430081, China

a970619543@qq.com

Abstract. Machine vision technology has been widely studied and applied in the development of science and technology, and has achieved remarkable achievements in real-time monitoring. Aiming at the upper limit of construction workers' personnel, there are safety hazards in the unclosed elevators. The development of the machine vision number statistics is discussed in detail, mainly from the status quo of the population statistics research, the direct method of the population statistics and the application of the indirect method. According to different algorithm research, the problems existing in the current research direction are analyzed, and the number of people in the future elevators is forecasted.

1. Introduction

Today, with the rapid development of China's economy, the construction of infrastructure is constantly strengthening, and high-rise buildings are becoming more and more. Building elevators are an indispensable tool for vertical transportation during the construction of these high-rise buildings. The safety of construction workers on building lifts is closely linked to the number of passengers on the building's lift. Machine vision technology is a process that uses a camera and a computer to replace the human eye to identify, track, and measure the target. It can be described in detail as converting the captured target into an image signal, and processing it through a dedicated processing unit. It can be described as converting the ingested target into an image signal and processing it through a special processing unit, and converting the image signal into a digital signal by using image pixel distribution and image information such as brightness and color, and performing various operations on the digital signal. In this way, the characteristics of the target are extracted, and then the features are judged, and the entire process of the device action at the scene is controlled according to the result of the judgment [1].

Under normal circumstances, the number of passengers on a building lift is limited, but sometimes the elevator becomes crowded for convenience and the risk factor becomes high. The traditional overload detection is by weighing, because the diversity of people's weight will lead to too many people but the weight does not reach the upper limit of weighing, which is a safety hazard for construction workers in the unclosed elevator space. In response to the above problems, the machine
vision technology is applied to the building elevator, which enables real-time monitoring and ensures the safety of personnel.

2. Direct method

In many public places, such as elevators in large markets, urban rail transit stations, and train stations, real-time supervision will be used to obtain information on the flow of personnel. This information can alleviate market and traffic congestion. Earlier, the United States applied traffic statistics technology to intelligent transportation systems, and Japan used the passenger counting technology in urban traffic management systems long ago.

At present, there are not many methods for counting the number of people directly in the car of a building elevator. However, there are many related studies in the elevator car[1] which is very similar to the building elevator car. Most of the techniques for counting the number of passengers in an elevator car are infrared detection[2] and computer vision[3]. In reference[2], the principle and structure of the automatic passenger counting method using infrared sensors are introduced in detail, and a dynamic time warping algorithm is proposed for signal recognition. In view of the disadvantages of the traditional Dynamic Time Warping (DTW) algorithm, such as large buffer space and strong dependence on endpoint detection, the author proposes an endpoint detection method with good adaptability.

Since the position of the camera is placed on the elevator, most of the human head features are detected. The human body recognition technology of head features is generally divided into the following two types[3]: face detection and head shape feature detection. Face detection mainly focuses on the detection and analysis of the frontal face area, generally considering the face part. The shape, color distribution information, and skin color, etc., therefore, the face detection image can only extract enough facial information from the close-range frontal face image. The head detection extracts the required information from the head position and size of the owner of the input image, and tracks, identifies, and analyzes related human targets based on the information. The goal is to search for and position the head of the human body from the sequence of images. In the dynamic detection and recognition of human beings, the human face is easily blocked by external influences, and the information of the human head is relatively complete. Therefore, in multi-angle, large-distance image shooting, head feature extraction and detection methods are usually used. Literature[4] proposed a method for identifying the number of people combined with the color and shape features of the human head. The method reduces the influence of shadow and illumination changes by image binarization, and estimates the number of people in the current scene according to the shape feature detection, tracking the human head, and analyzing the motion trajectory of the target.

The real-time supervision of the personnel in the elevator, in other words, is a real-time tracking and counting process for a moving target. The number of people can be divided into direct and indirect methods[5]. The direct method is also called the method based on target detection. It separates the individuals in the video and trains the classifier to directly count through feature extraction[6-7]. The indirect method is to measure the extracted features by machine learning or statistical analysis to achieve the purpose of counting[8].

3. Direct method

By recognizing the shape of the human body and its position for personnel counting, as long as the individual is accurately segmented from the image[9], the counting is easy to achieve and is not affected by image distortion and number of people. However, in the case of personnel congestion, the results of this method are not trustworthy. The direct method is applicable to the partial detection of an individual, for example, only detecting the head, the shoulder, etc., but the drawback of the partial detection is that the result obtained when the occlusion is severe is unreliable. Detection-based methods can be divided into two categories: model-based methods and trajectory cluster-based methods.
3.1. Model Method
The model and the human body shape can be used to achieve recognition more easily, and the recognition effect is also very good. The model method can be divided into head-like and overall detection methods. In the overall detection method: Rittscher proposed a video segmentation population segmentation system based on the expectation maximization formula[7]. The literature[7] selects the shape and position of the object as a parameter of the likelihood function to segment the pedestrian. Zheng Xiangxiang and others proposed a method for tracking and counting people based on head detection. The method uses the Adaboost algorithm to perform head detection, refines the head detection into multiple sub-detection processes from the head, and performs real-time monitoring through the feature histogram. This approach mitigates the interference of static errors on the experiment[10]. In head-like detection: Lin et al. proposed a crowd detection method using wavelet templates and machine vision[11]. Literature[11] uses the Haar wavelet transform feature to extract the head-like contour, and then classifies it by SVM to determine whether it is a human head. When the pedestrian density is large, the recognition rate is improved by the perspective transformation technique. If the image is too occluded or the outline of the head is not clear, it will be difficult to meet its real-time performance. In [12], a method based on double ellipse model to detect human head as a statistical basis for the number of people is proposed. After obtaining the contour features of moving objects, the head ellipse can be detected. This method improves the detection accuracy and reduces the misjudgement. However, if the picture is blurred, the target cannot be determined accurately. And Zhao Jun[13] and others have done a good job in this respect, and the combination of mathematical morphology and HSI color space effectively removes the image from the middle. The interference and noise in the head region are obtained by the edge detection, and finally the head contour information is used to determine and identify the target.

3.2. Track Cluster Method
This method can detect the movement and position of the human body by tracking a specific point on the person. Due to the noise of RGB color images, the difficulty of segmentation, and the incomplete structural information, the RGB-D camera using the new generation of sensing technology can simultaneously acquire RGB image sequences and depth image sequences, and also acquire bones. After the information of the points, the skeleton points are grouped, the feature extraction and fusion are performed by Gaussian weighted PHOG method, and finally classified by the sparse representation classifier. This method also achieves a high recognition rate in some databases, but the drawback is that the time is more complex. The trajectory cluster method is reliable and low when the population density is small and the occlusion is serious.

4. Indirect method
The indirect method usually separates local and global features from the foreground image of the crowd. This method is often more efficient than the direct method, but it is more difficult to perform direction counting. Since the features of its detection are easier relative to detecting a single individual, many features in the foreground such as foreground area[15], texture features[16], edge direction histograms[17], edge count values[18] are utilized. The regression function is used to count and estimate population density, as shown in Figure 1.
Such methods are sensitive to occlusion and viewing angles, and have problems in complex scenes where edge-based features are highly inaccurate in the case of complex backgrounds and uneven human clothing textures; in crowded situations, it is difficult to segment the foreground image from the background. Extracting a large number of features can be very time consuming, especially the extraction of edge features. Therefore, some researchers have proposed to use local features to overcome these problems while reducing the required training data. Hussain et al. proposed an automatic pixel-based population density estimation system [19]. Firstly, the foreground image is obtained by removing the background information from the reference image, the edge texture feature is extracted by edge detection, and then the extracted foreground block is corrected by zooming the perspective distortion, and is used as an input of the BP neural network to estimate the number of people. The literature uses supervised training methods to classify people into five categories, from low to high. Because the number of pixels in different regions of the image is very different, this method mainly corrects the distortion caused by the camera's view angle, the more complex technology of key point cluster is proposed in [20]. This method considers many factors that affect the relationship between the number of feature points and the number of people. The most important thing is to solve the influence of camera vision and crowd density. The number of SURF key points is calculated by graph-based clustering algorithm. Then the distance of each cluster is obtained by using inverse perspective mapping transformation. Finally, the number, distance and density of moving interest points are calculated. Equal information is used as input of support vector machine to estimate the number of people. But this method will affect the experimental results because it can not filter the static corners, and [21] solves this problem very well. First, extract the corners of the video frame; secondly, in order to eliminate the influence of the background corners on the statistical results, the algorithm uses optical flow method to estimate the motion vector of the corners, so as to filter the static corners; finally, the number of people is regressed by the first-order dynamic linear model. Compared with [20], the results of this method are more accurate and convincing.

5. Future prospects
At present, the theoretical research and practical application of the number of people based on machine vision have achieved remarkable results, but there are still many problems to be studied and solved in the future:

1) Due to the diversity of the detection object, the complexity of feature extraction, and the variability of the background, the description of the object is insufficient, the reliability of feature extraction is low, and the segmentation of the image is difficult, which brings the detection and classification of the target. Difficulties, the recognition rate has yet to be improved.

2) In the online inspection of machine vision, due to the large amount of data, high dimensionality of feature space, redundant information, and various problems in reality, the algorithmic ability to extract target information from a large amount of data is insufficient and inefficient.
3) Due to factors such as the built environment, lighting and noise, weak frequencies are difficult to distinguish from noise. In the future research, an algorithm is needed to make the detection system stable, reliable, and robust, and can adapt to illumination changes, noise, and external environmental interference.

With the development of the Internet of Things, intelligent video surveillance systems have become one of the hot spots, making the population statistics algorithm a hot spot. Realizing a fast, high-precision, real-time recognition system is one of the important research directions.

References
[1] Ying J and Li Q 2014 Computer J. Engineering and Applications 50 165-168.
[2] Wang Z W, Zeng Y J, Tian T and Yang N 2017 J. South Agricultural Machinery 48 94.
[3] Cheng S R, Chen H L and Hu R C 2017 J. Modern Computer 35-40.
[4] Wang Q and Feng Y 2010 J.COMPUTER MEASUREMENT & CONTROL 18 101-105.
[5] Conte D, Foggia P and Percannella G 2010 J. IEEE Computer Society 225-232.
[6] Tao Z, Ram N, Bo W 2007 J. IEEE Transactions on Pattern Analysis & Machine Intelligence 30 1198-1211.
[7] Rittscher J, Tu P.H and Krahnstoever N 2005 J. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 486-493.
[8] Zhang A Z and Li M 2012 J. Optical Engineering 51 047204-047204-13.
[9] Hou Y L and Pang G K H 2011 J. IEEE Transactions on Systems Man and Cybernetics 41 24-33.
[10] Zheng X X and Zhang Y 2011 J. COMPUTER APPLICATIONS AND SOFTWARE 28 44-46.
[11] Lin S F, Chen J Y and Chao H X 2001 J. IEEE Transactions on Systems Man & Cybernetics Part A Systems & Humans 31 645-654.
[12] Zhang J F, Mei X and Xu S S 2012 J. Computer Science 36 499-502.
[13] Zhao J W, Hou Q T, Li J P and Peng B 2013 J. Journal of Shandong University (Engineering Science) 43 6-10
[14] Mao X, Wang L and Li J J 2018 J. Computer Science 45 22-27.
[15] Ryan D, Denman S and Fookes 2009 J. Digital Image Computing: Techniques and Applications 81-88.
[16] Li X Y 2017 J.China Management Informationization 177-180.
[17] Xu Y X and Chen F 2017 J. China Management Informationization 22 852-862.
[18] Shi W S, Zhang X Z, Wang Y F and Zhang Q Y 2019 J. Computer Research and Development 56 69-89.
[19] Hussain N, Yatim H S M, Hussain N L 2011 J. Safety Science 49 824-833.
[20] Conte D, Foggia P and Percannella G 2013 J.Machine Vision & Applications 24 1029-1042.
[21] Cao Z T, Li X H and Zhou J L 2017 J. Journal of Computer Applications 37 141-143.