Emotion Recognition Algorithm Based on Panorama-plane Mapping Dataset and VGG16 in Prison Monitoring System

Zeping Zhang, Jing He and Zhiwei Zhang

School of Information Science and Engineering, Nanjing University Jinling College, Nanjing, China, 210089

* zhanghiwei@nju.edu.cn

Abstract. This paper proposes a digital processing scheme for efficient panoramic-plane image mapping for general-purpose panoramic cameras pictures, and builds an efficient emotion recognition algorithm based on VGG16 for real-time monitoring in prison environment. Panoramic-plane mapping algorithm is used to reduce the distortion of the input images and analyse the collected facial data to predict the sentiment of the prisoners. Compared with traditional monitoring methods, this algorithm is able to automatically sense the emotional changes of prisoners, which might improve the situation that traditional prison monitoring requires a large amount of human and material resources. We have proved that the method shows a good result on reducing distortion, besides our method runs with a small computation load so it is optimized to support real-time processing through experiments. The model trained in the experiments has reached a considerable accuracy up-to 79.75% on emotion classification at the same time. It is hopeful that the research could utilize in the future development of unattended prison monitoring.

1. Introduction

With the rapid development of computer vision, video surveillance systems have become one of an essential part of modern security tools in prison management. From analogue signal video monitoring that began in the 1970’s to digital video technology that emerged from the 1990’s, and intelligent networking video surveillance that has risen for years, many innovations in the field of surveillance keep continuous evolution. However, traditional video surveillance can only provide simple monitoring functions without providing any warning information based on the behaviour characteristics or sematic contents parsed out of prisoners. Actually, it is particularly important for a monitoring system to observe the facial expressions of prisoners, which might be a significant precursor to various harmful or dangerous events. In order to deal with the abnormal situation of prisoners as soon as possible, prison staffs have to pay attention the screen or loop the video continuously for a long time, which is easy to cause the observers fatigue and negligence. It becomes a necessary requirement for an intelligent monitoring system that it can interpret personnel's emotion from his/her facial expression and record them automatically. When a certain regular expression sequence appears, an alarm can be sent out in time to remind the prison staffs.

In recent years, innovative work of peers in the video surveillance field has been continuously proposed. Utilizing the IoT sensors to identify the posture and behaviour of prisoners automatically, escaping actions or any improper behaviours of prisons will be alerted and the sensors generated data are uploaded [1]. RGB-D camera technology is used to prevent prisoner suicide attempts [2]. Real-time
tracking of prisoners through RFID tag enhance the safety of prisoners, prison staff and the public [3]. Mark R-CNN-related technology is applied to analyse and recognize the prisoner video semantics and instance segmentation [4]. Wide-angle panorama camera has been widely adopted in prison environment to take as wide perspective view field as possible in the limited detention space to meet the supervision of every prisoner. But the fisheye lens in the camera introduce serious distortion problems. How to apply a high-accuracy image classification algorithm to distorted real-time video frames is an encountered technique issue in the application.

This article describes a panorama camera-based emotion recognition algorithm that can be used in intelligent surveillance systems. It uses the Cascade Classifier from OpenCV library to capture the personnel face areas contained in the panoramic image of the prison, and then apply the VGG training to them to obtain the recognized emotions. In order to eliminate the distortion effect lead by the fisheye lens in camera, the panoramic image is flattened by the plane mapping process before recognition model training. A weight model was established by VGG16 training on a dataset containing four different basic facial expressions collected. The gathered images of the original faces in the dataset are all Asian, making it more suitable for Asian face feature recognition. The overall processing reduces the calculation of the entire workflow by optimizing the digital image processing algorithm and makes the entire system available for real-time monitoring.

2. Panorama image to flat plane mapping process
In order to eliminate lens distortion and improve the visual effect, an image distortion correction algorithms based on dual longitudes model [1] is applied and a fast algorithm is implemented to improve the process efficiency and effect in which a large number of matrix operations are taken place of the conditional loops in actual engineering. The process contains three steps, namely effective area calibration, effective area correction and interpolation correction.

2.1. Calibration Effective Area
Because edge detection is susceptible to noise in the image, the input RGB image is converted from color space to gray space and binarized before the Canny edge detection algorithm [6] is applied. To utilize the Canny edge detection, Sobel kernel is used to smooth the image in both horizontal and vertical directions to obtain the first order derivative of the image. Let the first derivative of the image in the horizontal and vertical direction be $G_x$ and $G_y$, respectively, the edge gradient magnitude and direction of each pixel can be found as follows,

$$
\begin{align}
\text{Edge Gradient}(G) &= \sqrt{G_x^2 + G_y^2} \\
\text{Angle}(\theta) &= \tan^{-1}\left(\frac{G_y}{G_x}\right)
\end{align}
$$

The useless pixels in the image which can not construct the edges have to be removed. That means the pixel with gradient exceed the adjacent maximum edge value are not wanted. As shown in Figure 1, for a specified gradient direction, points B and C are used to check whether point A constitutes a local maximum. If not, A is set to 0.

![Figure 1. Check for useless pixel](image1)

Figure 2. The radius after circle center correct
Finally, the intensity gradients of all edges are evaluated. $\text{minVal}$ and $\text{maxVal}$ are defined as thresholds. If the intensity gradient of an edge exceeds $\text{maxVal}$, an edge is determined, and edges with value below $\text{minVal}$ are discarded. The edge intensity between $\text{minVal}$ and $\text{maxVal}$ are treated as follows: if it is connected to a pixel that is determined as an edge, it is an edge; if it is connected to an edge pixel that is not an edge, discarded. Among the obtained effective edges, the $x$ values of the leftmost and rightmost pixels as well as the $y$ values of the top and bottom pixels are defined as ranges of the effective area. Because panoramic camera sensors usually do not use 1:1 scale, that means if you want to get the radius of a circular black edge, you need to find the maximum value of the left-right, or top-bottom lengths as the radius of the black edge (Figure 2). In addition, since the center of the circle of the panoramic camera is not always in the center of the image, there may be a possible offset. This method also helps to correct this bias.

2.2. Effective area correction
Define the spherical point coordinates $(x, y, z)$, the target image point coordinates $(i, j)$, and the relationship between the spherical point coordinates and the target image point coordinates can be expressed as:

$$\begin{align*}
x &= \frac{R}{\sqrt{\tan^2(\pi - \frac{\pi}{2R} \times j) + 1 + \tan^2(\pi - \frac{\pi}{2R} \times j) / \tan^2(\pi - \frac{\pi}{2R} \times i)}} \\
y &= \frac{R}{\sqrt{\tan^2(\pi - \frac{\pi}{2R} \times i) + 1 + \tan^2(\pi - \frac{\pi}{2R} \times i) / \tan^2(\pi - \frac{\pi}{2R} \times j)}} \\
z &= \frac{R}{\sqrt{1 + 1 / \tan^2(\pi - \frac{\pi}{2R} \times j) + 1 / \tan^2(\pi - \frac{\pi}{2R} \times i)}}
\end{align*}$$

(2)

This paper uses the orthogonal projection algorithm to establish the mapping relationship between spherical coordinate points and fisheye image pixels, (Figure 3 and Figure 4). $p_1, p_2$, and $p_3$ are the projection points of $p$ on the xoz, yoz, and xoy planes, respectively. $\phi$, $\theta$ and $\omega$ are the positive angles between $op_1$ and the $x$-axis, and $op_2$ and $y$-axis, $op_3$ and the $z$-axis, respectively. The correspondence between the angles $\phi$, $\theta$ and the coordinate values of the $p$ point $(x, y, z)$ can be established:

$$\begin{align*}
x^2 + y^2 + z^2 &= R^2 (z \geq 0) \\
\tan(\phi) &= \frac{z}{x} (0 \leq \phi \leq \pi) \\
\tan(\theta) &= \frac{z}{y} (0 \leq \theta \leq \pi)
\end{align*}$$

(3)

The spherical point $p$ is mapped to $\mathbb{p}''(u, v)$ point on the fisheye image, and the corresponding relationship between them can be described as follows:

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

(4)

where $(x_0, y_0)$ is the center of the fisheye image, and then,
\[
\begin{cases}
  u = \frac{R}{\sqrt{\tan^2 \phi + 1 + \tan^2 \phi / \tan^2 \theta}} + x_0 \\
  v = \frac{R}{\sqrt{\tan^2 \theta + 1 + \tan^2 \theta / \tan^2 \phi}} + y_0
\end{cases}
\] (5)

Therefore, the mapping relationship between the target image points and the image points collected by the fisheye camera can be obtained.

\[\text{Figure 3. Spherical coordinates} \quad \text{Figure 4. Fisheye coordinates} \quad \text{Figure 5. Bilinear interpolation}\]

2.3. Bilinear Interpolation Algorithm

The accuracy of the bilinear interpolation algorithm [7] is higher than that of the nearest neighbor interpolation algorithm [8], and faster than the 3-time convolution interpolation algorithm [9]. When the pixel value of the point \(p\) is unknown, the gray values of the four nearest neighbor pixels around \(p\) are used as weighted to determine the pixel value at point \(p\) (Figure 5). Let \(A = (x_1, y_2), B = (x_2, y_2), C = (x_1, y_1), D = (x_2, y_1)\) are the neighbor point of \(p\), pixel value of \(R_1(x, y_2), R_2(x, y_1)\) are calculated first and then value of \(p\) as follow,

\[
\begin{aligned}
  f(R_1) &\approx \frac{x-x_1}{x_2-x_1} f(C) + \frac{x-x_2}{x_2-x_1} f(D) \quad \text{where } R_2 = (x, y_1) \\
  f(R_2) &\approx \frac{x-x_1}{x_2-x_1} f(A) + \frac{x-x_2}{x_2-x_1} f(B) \quad \text{where } R_3 = (x, y_2) \\
  f(P) &\approx \frac{y-y_1}{y_2-y_1} f(R_1) + \frac{y-y_2}{y_2-y_1} f(R_2)
\end{aligned}
\] (6)

The final mapped and tailored image is shown in Figure 6. The overall flat mapping effect of the panorama image is satisfied, but the edges outside the black frame is severely stretched which has little effect in the subsequent neural network recognition and are abandoned. The image inside the black box are taken as the input of the following emotion classification procedure.

\[\text{Figure 6. Mapped and tailed panorama image}\]
3. Utilizing VGG16 to classify emotion

VGG is a convolutional neural network model proposed by K. Simonyan and A. Zisserman of Oxford University [10], which could achieve 92.7% test accuracy (top-5) on ImageNet. We trained our sentiment classification model with VGG16 and ReLU as the activation function. The neural network contains a total of 138 million parameters, and finally achieved an accuracy of 79.75%. The network architecture of VGG16 is shown in the following Figure 7:

![Figure 7: Classic VGG16 model configuration](image)

The data set used to train the model is derived from autonomous collection, more than 10,000 images, including four typical categories: Sad, Neutral, Surprise and Happy. An example of a dataset is shown in Figure 8:

![Figure 8: Dataset examples](image)

4. Experiments

To verify the effectiveness of the algorithm proposed above, experiments are carried out from three perspectives, the effect and speed of panorama-plane mapping processing of the original camera images as well as the recognition accuracy of VGG model. Finally, a satisfied effect is obtained, and it is verified that it has a stable work load when working for a long time. The experiment platform is a Linux based environment on Intel Core i7 7700K, 16GB memory and NVIDIA GTX 1080 hardware.

4.1. Digital Image Processing Effect and Speed

In the experiments, four sets of photos are randomly captured from the 1-hour continuous and stable record to show the actual effect of the algorithm (Figure 9). The first row corresponds to the original image collected by the camera, and the second row corresponds to the digital image processed in this article. The obtained image frame rate of the video stream is stable at about 12fps when our experimental platform runs stably (Figure 10). It can be considered that the load is relatively stable in the range of 10 and 14 most of the time. The single frame processing speed is about 0.083 seconds, proven the efficiency of digital image processing.
4.2. Actual Effect of VGG16 Model and Accuracy
The training was performed on the self-collected expression dataset, and the trained VGG16 model achieved an accuracy of 79.75%. The confusion matrix is shown in Figure 11, and a sample of the recognition effect from our model is shown in Figure 12 and Figure 13.
5. Summary
In this paper, an improved real-time visual correction method based on a panoramic fisheye camera is implemented, and an emotion monitoring model based on VGG16 is deployed on it. It is proved that the distortion correction method not only has a good mapping effect, but also has the characteristics of high speed and stability, which guarantees the image processing frame rate and could be used in real-time monitoring in prisons. The proposed method improves the current situation of traditional prison monitoring that requires a large amount of human and material resources to be invested. It is hopeful that this research might promote the development of unattended prisons in the future.

References
[1] Suvarna Nandyal, Soumya Mugali, 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), July, 2019
[2] W. Bouachir and R. Noumeir, "Automated video surveillance for preventing suicide attempts," 7th International Conference on Imaging for Crime Detection and Prevention (ICDP 2016), Madrid, 2016, pp. 1-6.
[3] M. S. Selamat and B. Y. Majlis, "Considering RFID Inmate Tagging Application to Enhance Prison Management," 2006 IEEE International Conference on Semiconductor Electronics, Kuala Lumpur, 2006, pp. 253-257.
[4] Y. Liu, Y. Yang, A. Shi, P. Jigang and L. Haowei, "Intelligent monitoring of indoor surveillance video based on deep learning," 2019 21st International Conference on Advanced Communication Technology (ICACT), PyeongChang Kwangwoon Do, Korea (South), 2019, pp. 648-653.
[5] Wei Lisheng, Zhou Shengwen, Zhang Pinggai, Sun Yizhou. Double longitude model based correction method for fish-eye image distortion [J]. Chinese Journal of Scientific Instrument, 2015, 36(02): 377-385.
[6] J. Canny, "A Computational Approach to Edge Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-8, no. 6, pp. 679-698, Nov. 1986.
[7] Jaderberg, Max & Simonyan, Karen & Zisserman, Andrew & Kavukcuoglu, Koray. (2015). Spatial Transformer Networks. Advances in Neural Information Processing Systems 28 (NIPS 2015).
[8] H. Jégou, M. Douze and C. Schmid, "Product Quantization for Nearest Neighbor Search," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 1, pp. 117-128, Jan. 2011.
[9] R. Keys, "Cubic convolution interpolation for digital image processing," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, no. 6, pp. 1153-1160, December 1981.
[10] Simonyan, Karen & Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.