Intelligent visual enhancement system

Xiaogang Yang¹, Mashuang Ma², Weipeng Li² and Xueli Xie²

¹205 Teaching and Research Office, Xian Research Institute of Hi-Tech, Xian, 710025, China
²Graduate School, Xian Research Institute of Hi-Tech, Xian, 710025, China

Corresponding author: mamas008@163.com

Abstract. Vision-based augmented reality is a new kind of visual application technology, which transfers synthetic sensory information into a user's perception of a real environment. It is perceived by human senses. However, the existing visual enhancement equipment has single function, limited processing capacity and poor interaction. To overcome these shortcomings, we designed a visual enhancement system that integrates cloud computing, AR technology and deep learning. First, small, remote and wireless cameras are used to obtain image data, which need to be uploaded to a cloud. Then, the method of deep learning and feature matching is adopted to carry out facial consistency analysis, which improves the robustness of target detection. And stable target tracking is achieved by time sequence state filtering. Finally, the information of image analysis and processing is transmitted back to the AR device, so the prompts of target text and voice are given for intelligent auxiliary decision-making in time.

1. Introduction

From trend of visual enhancement device development, the hardware technology of visual enhancement has developed rapidly in AR and MR fields, and many landmark products have appeared [1-4]. In April 2012, Google announced a Google Project Glass that was "augmented reality". It materialized the concept of visual enhancement system for the first time and achieved a breakthrough from scratch. Although Google glasses define people's cognition of AR glasses, the function of this product is a display which is far from meeting users' demand for data processing. In addition, the visual effect of this product is not good enough to bring a good user experience.

In 2015, Microsoft officially announced the launch of Hololens, which enabled users to interact with digital content. Users can interact with holographic images in the surrounding real environment, which fills the gap in the data processing capacity of AR glasses.

In addition to the civil domain, vision-based augmented reality has developed in the field of military rapidly. In 2016, the Defense Advanced Research Projects Agency (DARPA) developed" urban officer tactical response, identify, and visualization system". The military intelligence took target recognition and security risk indication as the main requirements to write into the design scheme, which made the visual enhancement system having a high demand for computing power. Now, constrained by the level of hardware manufacturing and algorithm optimization, the pentagon's research is still in its infancy.
Figure 1. Development of visual enhancement technology

Through analysis of the existing products on the market, as shown in Figure 1, we know that the researches of visual enhancement technique are mainly concentrated on the intelligent glasses and wearable devices. We design a set of intelligent visual enhancement system, including: cloud computing, deep learning and AR display technology. The high strength computing is assigned to the cloud, intelligent information processing is fulfilled by deep learning, the user terminal only includes the visual display and audio output device. This scheme can greatly reduce the requirement on the computing ability of wearable devices, release the load of users and achieve better effectiveness.

2. Related work

2.1 Face Detection and Recognition
Facial features include grayscale and skin colour [5-7]. Because contour is an important feature of human face, each organ in the face area has a unique gray distribution feature. Therefore, regional feature parameters including eyes, nose and mouth can be selected for face detection based on statistics. Skin colour features do not change with facial expressions. According to the overlapping of "skin colour" and "non-skin colour" areas, and the distribution of "skin colour" areas, the skin colour model is selected to describe skin colour characteristics. Gaussian model, mixed Gaussian model and histogram model are usually adopted as skin colour model.

Face recognition based on LBPH is a histogram model. The basic idea is as follows: first of all, select a pixel as the centre for judging pixel value with surrounding pixels. Then, adopt binary coding, so as to get the whole image of LBP coding images. Next, LBP image can be divided into 16 areas for each region of LBP coding histogram. We get the whole image of LBP histogram coding, through comparing the different face image LBP coding histogram, to achieve the goal of face recognition.

2.2 Target Detection
Target detection is to detect the location of a specific target from the image. From the perspective of the development of convolution neural network in the field of target detection, the best detection effect is the R-CNN model, and the series of improved versions which it produces[8-10]. In the target detection of PASCAL VOC in 2014, the average accuracy of the R-CNN model proposed by Girshick reached 62.4%. Compared with traditional detection methods, R-CNN is nearly 20% higher. Based on this, the researchers continued to improve the R-CNN model, including: R-CNN, SPP-Net, Fast R-CNN, Faster R-CNN, YOLO[11], SSD and so on.

The main idea of YOLO (You Only Look Once) is to use a convolution neural network to complete the target detection. In the process of implementation, YOLO image is divided into a grid, each network has a total of 30 dimensions. Among them, 8 dimensions is a result of the coordinates of target location, 2 dimensions is confidence, there are 20 dimensions correspond to different target categories. Each grid in YOLO is responsible for predicting one target at most. when an target falls into the grid, the grid is responsible for predicting the location and confidence of the target. Since only one feature is extracted from the whole image, YOLO has the feature of fast detection speed. It can detect 45 images per second, and realize real-time detection of different target.
2.3 Multi-target tracking and state filtering
The main task of multi-target tracking is to locate multiple targets simultaneously in the sequence of images, and maintain their respective ID to record the trajectory. When designing the multi-target tracking algorithm, two main problems need to be considered. First, how to measure the similarity of targets between different frames; Second, how to judge whether the target is the same target between different frames. The former is mainly a modelling problem, such as common appearance model, motion model, interaction model, rejection model, occlusion treatment model and prediction model. The latter is primarily a matter of data association.

Because target detection and face recognition have certain position error and classification error, we adopt multi-target time series state filter to improve accuracy of recognition and stability of target detection. Considering that there may be errors in the classification results of targets. Firstly, the multi-target tracking is used to correlate the two frames of the target, and the target position is filtered. Then, update its category and confidence, inhibit the category of volatility target confidence, and the target with low confidence is still in the original category.

3. Our approach
3.1 Hardware Design
The overall framework structure of this hardware platform consists of four modules [12-14], including image acquisition module, cloud computing module, AR display module and communication module. The overall architecture is shown in the Figure 2.

| System Module | camera | Cloud Computing | Image Display | Communication |
|---------------|--------|-----------------|---------------|---------------|
| Hardware Selection | IP Camera | Workstation with GPU | BT-300 | MERCURY Router |

3.1.1 Image Acquisition Module
In order to obtain the remote image information, the IP camera is used as the image acquisition module. Network camera, or IP camera, consists of network coding module and analog camera. Network coding module collected in analog camera will be converted into digital signal, after the chip compression data transmission, through the WAN or LAN connection to the image analysis processing system.
3.1.2 Cloud Computing Module
When carrying out tasks, such as target detection, face recognition and tracking filtering, the system need to combine YOLO with other methods. Because the common processing equipment cannot meet the real-time data processing, it uses GPU graphics processor hardware for acceleration. In the process of development of this system, the cloud computing platform adopted a high-performance workstation, which processes and returns image data in real time. It is shown in Table 2.

| Index | CPU            | GPU     | Basic Frequency | FPS  |
|-------|----------------|---------|-----------------|------|
| Development | E5-1650 v4 | TITAN X | 3.2GHz          | 33.1~33.5 |
| test   | i5-7200U       | GTX 960 | 2.5GHz          | 16.8~17.2 |

3.1.3 Image Display Module
At present, there are many kinds of AR glasses and the way of connection [15]. So the product in effectiveness is also different from each other. BT-300, HP MR and Microsoft meta2 were selected for comparison in all aspects. In order to meet the needs of outdoor environment and special people, BT-300 augmented reality smart glasses were selected as remote display equipment.

3.1.4 Communication module
Under the conditions of the experiment and test, cloud computing module get access to real-time data from the image acquisition module[16-20]. And project the analysed data onto the display device. There are a lot of data exchanges between these two links. MERCURY wireless router was used to build a LAN from the perspective of real-time image transmission, economic cost and system portability. As Shown in Figure 3.

![Figure 3. Overall structure design of equipment connection](image)

3.2 Software design
The whole system framework of face detection and target recognition of the intelligent visual enhancement system is shown in the Figure 4, and each module will be described in detail below.
3.2.1 Image Pre-processing
In order to obtain more effective image information, the system needs a process of calibration. Vision calibration aims to align the results of AR projection with the real world. So vision calibration error is the difference between projection results AR glasses with the real world. The effect of the error factors include: camera field angle, installation position, AR display field of vision range and user’s distance.

The picture shows a process of vision calibration. Figure 5(a) gets the observation result from the first perspective before calibration, and it can be seen that there is a significant error between the observed calibration plate and the AR display calibration plate. In Figure 5(b), the projection of the calibration plate coincides with the actual observation, and the error is basically eliminated.

3.2.2 Face Detection and Recognition
In order to balance the algorithm efficiency and hardware pressure, the CPU is used to perform face detection and recognition. This paper divides face detection and face recognition into two steps. And the local binary mode histogram (LBPH) scheme is used for face recognition.

The calculation of LBP encoding is shown as follows. First, the pixel coordinates of the first neighbourhood of the current coordinates are calculated:

\[
\begin{align*}
    x_k &= -r \sin(2k\pi / n) + x \\
    y_k &= r \cos(2k\pi / n) + y
\end{align*}
\]

(1)

Where, \( r \) represents the neighbourhood radius and \( n \) represents the numbers around the pixel.

In order to obtain a face recognizer with stronger robustness to noise and avoid local interference, the sampling radius of LBP is set to 2. Then, the pixel value of the neighbourhood is calculated using bilinear interpolation, and the first LBP value of the current coordinate is obtained by setting a field near each sampling point:

\[
LBP(x, y) = \sum_{k=0}^{n-1} LBP_k(x, y) 2^k
\]

(2)

The distance between feature vectors is used as the basis for facial similarity, which is defined as:
\[ \text{sim}(v_i, v_j) = \exp \left( -\frac{|v_i - v_j|^2}{2\sigma^2} \right) \]  

(3)

Where \( v_i \) and \( v_j \) are eigenvectors of two different faces, and \( \sigma \) is the standard deviation.

### 3.2.3 Multi-target tracking and state filtering

The purpose of multi-target tracking is designed to correlate potential identical targets between two sequential pictures according to its location information. The standardized association distance for two sequential pictures is defined as:

\[
d_{ij}(b^{(r-1)}_i, b^{(r)}_j) = \frac{w_p |p^{(r-1)}_i - p^{(r)}_j| + w_s |s^{(r-1)}_i - s^{(r)}_j|}{|s^{(r-1)}_i|} \tag{4}
\]

Where, \( b^{(r-1)}_i \) represents the goal of \( i \) moment, \( b^{(r)}_j \) represents the goal of \( j \) moment, \( p^{(r-1)}_i \) and \( p^{(r)}_j \) represent positions of targets, \( s^{(r-1)}_i \) and \( s^{(r)}_j \) represent the size, \( w_p \) and \( w_s \) represent the weight of the difference between position and size.

Update the confidence and target categories according to their consistency. Confidence update is defined as:

\[
f^{(t)}_j \begin{cases} 
0.8 f^{(r-1)}_j + 0.25 \hat{c}^{(t)}_i, & c^{(t)}_i = \hat{c}^{(t)}_i \\
0.86 f^{(r-1)}_j + 0.1 \hat{f}^{(t)}_j, & c^{(t)}_i \neq \hat{c}^{(t)}_i
\end{cases} \tag{5}
\]

Where, \( f \) represents the confidence of the target, \( c \) represents the target category.

Compared with the target detection results of single image, the improvement of target positioning accuracy and classification accuracy is 15% and 17% respectively. The results of two images are compared between two sequential pictures. In Figure 6, it can be seen that the result of target detection after state filter is more stable.

**Figure 6. Comparison of target detection effects**

### 4. Experimental results

This system is mainly used for automatic detection, recognition and tracking of specific targets such as integrated intelligent precise guidance, intelligent monitoring of security system, blind visual AIDS and visual enhancement of human eyes. The experimental results are shown in the Figure 7.
4.1 Vision calibration error
In order to reduce error of vision calibration, we invited five subjects to participate in the test. Each person uses the calibration board at 10 different distances for vision calibration and record the subjective perception of the relative error of the calibration. A total of 50 set of error data are shown in Table 2. It can be seen that the calibration relative error is basically around 0.049, and the error is relatively stable, which can meet the actual user demand well.

| Experimenter | Numbers | Mean Relative Error | Standard Derivation |
|--------------|---------|---------------------|---------------------|
| 1            | 10      | 0.052               | 0.025               |
| 2            | 10      | 0.064               | 0.031               |
| 3            | 10      | 0.041               | 0.024               |
| 4            | 10      | 0.031               | 0.022               |
| 5            | 10      | 0.059               | 0.018               |
| Total        | 50      | 0.049               | 0.027               |

4.2 Software error
In terms of software error, we mainly test the positioning accuracy and classification accuracy of target detection and face recognition, as well as the overall time of software system. The test platform is a high-performance workstation equipped with Intel Xeon E5 CPU and NVIDIA TITAN X GPU. The statistical results are shown in the Table 3.

In Figure 8, It can be seen that there is no significant relationship between the processing time of a single image and the number of targets.

| number of images | Indicator | Mean |
|------------------|-----------|------|
| Target Detection | 1500      | IoU 0.65 |
| Face Recognition | 1500      | AUC 0.69 |
| Time consumption | 1500      | ms 33.5 |
5. Conclusion
In this paper, we designed intelligent visual enhancement system. It integrates cloud computing technology, AR technology and deep learning technology. The real-time detection of multiple targets and the stable tracking of targets through time series state filtering. It can provide high-precision auxiliary decision-making information for user. From the experimental results, the mean accuracy of recognition is 0.65, and the mean value of face recognition is 0.69, the system time delay for a few milliseconds. So it can be applied to many kinds of tasks, such as intelligent precise guidance, intelligent detection of security system, blind visual AIDS and visual enhancement of human eyes.

References
[1] N. J. Crane, S. M. Gillern, K. Tajkarimi, et al. Visual enhancement of laparoscopic partial nephrectomy with 3-charge coupled device camera: assessing intraoperative tissue perfusion and vascular anatomy by visible hemoglobin spectral response. Journal of Urology, (2010), 184(4):1279-1285.
[2] Y. J. Jang, Y. K. Ryu. Development of a head-mounted visual enhancement device for people with low vision. Proceedings of SPIE - The International Society for Optical Engineering, (2004), 5602:148-159.
[3] K. Tajkarimi, N. Crane, M. Alemozaffar, et al. Visual Enhancement of Laparoscopic Partial Nephrectomy with Three Charge-Coupled Device (CCD) Camera. Journal of Urology, (2007), 177(4):207-208.
[4] G. Fan, J. Li, M. Li, et al. Three-Dimensional Physical Model-Assisted Planning and Navigation for Laparoscopic Partial Nephrectomy in Patients with Endophytic Renal Tumors. Scientific Reports, 2018, 8(1):582.
[5] Q. Li, F. Jiang, Z. Huang. Multi-pose facial features localization based on skin color models. CISP, (2015):297-301.
[6] H. Riri, A. Elmoutaouakkil, A. Beni-Hssane, et al. Classification and Recognition of Dental Images Using a Decisional Tree. CGIV, IEEE, (2016):390-393.
[7] S. Gundimada, V. Asari. Face detection technique based on rotation invariant wavelet features. Proceedings. ITCC, IEEE, (2004):157-158 Vol.2.
[8] J. Redmon, S. Divvala, R. Girshick, et al. You Only Look Once: Unified, Real-Time Target Detection. Computer Vision and Pattern Recognition. IEEE, (2016):779-788.
[9] J. R. Uijlings, K. E. Sande, T. Gevers, et al. Selective Search for Target Recognition. International Journal of Computer Vision, (2013), 104(2):154-171.
[10] K. M. He, X. Zhang, S. Ren, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. European Conference on Computer Vision, (2014), 2014:346-361.
[11] J. Redmon, A. Farhadi. YOLO9000: Better, Faster, Stronger. (2017):6517-6525.
[12] H. Schulzrime, A. Rao, R. Lanphier. Real Time Streaming Protocol (RTSP). RFC, (1998).
[13] A. Rao, R. Lanphier, M. Stiemerling, et al. Real Time Streaming Protocol 2.0 (RTSP). (2014).
[14] G. Song, H. Zhou. The design of image acquisition and display system. ICETC, (2010), 5:V5-23-V5-26.
[15] M. Kurashige, K. Ishida, T. Takeokura, et al. Image display module, (2016).
[16] J. S. Parkinson, E. C. Kofoid. Communication modules in bacterial signaling proteins. *Annual Review of Genetics*, (1992), 26(1):71.

[17] F. X. Nathanial, S. Connely, W. Meyer, et al. Software architecture for wireless data and method of operation thereof: US, US7092998.(2006).

[18] Bavandpour F. Globally distributed virtual cache for worldwide real-time data access,(2016).

[19] T. Xiao, M. Xu, J. Xu, et al. Acquisiting text documents opened by notepad from Windows7 RAM image. *Journal of Computational Information Systems*, (2014), 10(16):7117-7124.

[20] E. Zandvoort. The international service for the acquisition of agri-biotech applications' biosafety initiative,(2010), 3(3):1208-1211.