A Teaching Evaluation System Based On Visual Recognition Technology

Ying Zhang, Yige Guo, Jingrong Hu, Qiongwen Liang, Jinjian Jiang, Changxing Shao, Yong Wang*

Department of Automation University of Science and Technology of China Hefei, China

* Corresponding author e-mail: yongwang@ustc.edu.cn

Abstract. This paper designs and implements a student-centered teaching evaluation system based on face recognition and pose estimation technology. Our work firstly combines classroom attendance and behavior analysis in an evaluation system. For checking attendance, we select student faces as the identification object, employing a multi-task cascaded convolutional networks (MTCNN) as a face detector and a deep learning network FaceNet to extract face features. Then the head pose information is analyzed using Ensemble of Regression Trees (ERT) algorithm, which is able to detect 68 key feature points of faces. At last, we design and implement the whole system, including designs of functional modules, service software, database and telecommunication of various parts. This system can check attendance and collect student behavior information automatically, enhancing the intelligent level of the learning and teaching system.

1. Introduction

Teaching evaluation is a series of small-scale evaluation of the ongoing learning in a classroom. It focuses on students’ learning process and performance in classes, which helps teachers to obtain students' learning effect and status to improve teaching efficiency. Therefore, the problem of teaching evaluation design is student-centered and the focus of attention is not on the teaching but the learning of students. Our work pays attention on the behaviors of students in classes to evaluate students’ learning effect.

During classes, attendance and the behavior of the students are two most important evaluation indicators of test score. Attendance reflects the enthusiasm of students and learning behaviors show students' learning effects and statuses. However, manually evaluating students’ attendance and learning behavior is inefficient and directly influences the teaching process. With the development of computer vision technology, visual recognition technology has been widely used in various industries to improve the intelligent level. In this paper, we developed a teaching evaluation system based on the visual recognition technology. The overall diagram of this system is illustrated in Fig. 1.

Face recognition, which is based on facial feature information for verification, is a kind of biometric recognition technology. Since face features are unique and hardly
Fig 1. Overall diagram of the proposed teaching evaluation system.

Irreproducible, it can be more accurate for identification than other biometrics. Therefore, we select students’ faces as the identification object to judge the attendance. Besides, head pose is of great importance to understand and analyze human behaviors, which can indicate students’ different behaviors in classes, such as looking at the blackboard, glancing around or bowing down on the mobile phone, etc. Therefore, head pose, as an evaluation factor, is used to analyze the behavior of students. Vision-based head pose estimation (HPE) is now a hot but tough topic in the field of computer vision.

By computer vision technologies and machine learning algorithms, we focus on the recognition analysis module, which is in charge of taking attendance and pose estimation. At first, a webcam installed in a classroom is used to continuously shoot the classroom scene, which is connected to the recognition analysis modules and database to perform data transmission and analysis. In the proposed framework, the recognition analysis modules are divided into two main modules: face recognition and pose recognition. Face recognition algorithm is used to detect and recognize the student faces for attendance. Comparing to the conventional ways, this system is able to continuously record the students’ attendance. It helps to prevent students from late arrival or early departure. Next, pose recognition is applied to estimate students head poses. Finally, the recognition results will be used to evaluate students’ behavior and performance.

In summary, we proposed a teaching evaluation system that integrates with two main analysis modules: face recognition, head pose estimation. The rest of the paper is organized as follows. Section II covers the related face recognition and pose estimation algorithms. Our methodology is discussed in section III with the explanation on face recognition and head pose estimation. Next, section IV introduces the workflow of the whole system and experimental results. Finally, there is a conclusion in Section V.

2. Related work

2.1. Face Recognition
Face recognition technology is mainly used to extract features of faces. Previous “Eigenfaces” [1, 2] approach was a milestone in face recognition. However, the recognition effect relies too much on the accuracy of feature localization algorithm. Subsequently, Brunelli and Poggio [3] found that the template matching approach is better than the feature-based approach through experiments. Their method has a better performance in illumination invariance, but cannot eliminate the influence of facial
expression changes. Belhumeur et al. [4] proposed the Fisherface approach which first employs the Principal Component Analysis (PCA) for dimensionality reduction and then calculates the Euclidean distance between the target features and the dimensionality reduction features for face identification. Another elastic map matching approach extracts the Jet features [5] to obtain the attribute map of the input images. However, these methods are still sensitive to pose, illumination, and other variable conditions. When certain condition changes, the identification results are unsatisfactory. Deep learning [6] has made great achievements in face feature extraction. It can weaken the influence of external factors and improve the reliability of face recognition, so as to promote the practical value of face recognition technology.

Face recognition is an important part in the classroom attendance module, which extracts the feature of each student and saves it for further processing or analysis. In order to improve the accuracy of attendance system, it is necessary to choose a robust face recognition algorithm. FaceNet, a well-worked deep learning network, is thus used as the face feature extractor in this paper.

2.2. Head Pose Estimation

Generally, head pose estimation refers to the process of estimating the head pose parameters. This section presents the most commonly used computational methods for head pose estimation, which can be categorized in appearance template methods, geometric methods and nonlinear regression methods.

Appearance template methods: In 1994, Deymer [7] introduced the idea of template matching into head pose estimation. First, head pose space was quantized into a number of discrete points, and several template images were prepared for each pose. Then the head image to be estimated was compared with the images in the template library, and the head corresponding to the template image with the highest score is matched, which is the estimation result of the test image.

This type of methods has two major advantages: 1) The number of templates can be dynamically adjusted according to actual needs. 2) No training is required, and there is no need to prepare positive samples and negative samples for training. Comparing with the learning-based method, it can avoid the complicated training sample calibration work, the problem of over-fitting and poor generalization performance caused by improper selection of training samples.

All the mentioned template methods have limited head posture expression, which can only obtain finite discrete pose values. Hence from the perspective of system accuracy, the larger the number of templates, the higher the system accuracy. However, the system time complexity is also proportional to the number of templates.

Geometric methods: This head pose estimation method, which is based on the geometric relationship of the face key points, draws on the perception mechanism of human vision on the head posture, and uses the relative position of the face key points to estimate the head posture. The basic flow of this type of methods is to first determine the location of the key points of the face, and then estimate the head position by the relative position of the key points.

Accurately locating the key points of the face is the first key step in such a method. The mapping of the face geometric position to the head pose is the second key step of this method. Among these methods, the more successful ones are the research results of [8] and [9]. They used the key points around the eyes, nose, and mouth (a total of 28 key points) to design a normalized Singular Value Decomposition (NSVD) to estimate the head pose. Some researches use uncertain face key points to estimate head pose [10], which provide a new research idea for this kind of method, but the precision was limited and can only be used for rough estimation.

The overall advantage of model-based methods is that the head pose can be obtained with only a few facial key-points. The drawback is that the accuracy of the key-point detection is closely linked to the accuracy of the head pose estimation.

Nonlinear regression methods: Head pose estimation formulated as a regression problem, which is a mapping problem from image space to pose space. There are different techniques to learn a nonlinear mapping from head images to pose angles. Support vector regression was successfully used
in combination with dimensionality reducing methods such as principal component analysis [11] and localized gradient orientation histograms [12].

This paper selects the geometric methods to estimate the head pose, due to lower time complexity. Support vector regression is used in combination with Ensemble of Regression Trees (ERT) face alignment algorithm [13] which is able to locate 68 face key points.

3. Methodology
In this section, we will explain in details our approaches towards face recognition and head pose estimation.

3.1. Face Recognition
Face recognition consists of three main steps, which are face detection, face feature extraction and face classification. A multi-task cascaded convolutional networks (MTCNN) [14] face detection algorithm is applied to detect faces in a classroom, and FaceNet [15] will be used to extract face features for the detected faces. Then the trained Support Vector Machine (SVM) [16] classifier is used for recognizing faces and outputting the results.

1) MTCNN: The framework of MTCNN adopts a cascaded structure with three deep convolutional networks that predict face and landmark location in a coarse-to-fine manner, as shown in Fig.2.

Stage 1: The fully convolutional network, called Proposal Network (P-Net), will obtain the candidate windows and their bounding box regression vectors. Then the estimated bounding box regression vectors is used to calibrate the candidates. After that, non-maximum suppression (NMS) is employed to merge highly overlapped candidates.

Stage 2: The output results of P-Net are fed to Refine Network (R-Net), which further rejects a large number of false candidates, performs calibration with bounding box regression and NMS candidate merge.

Stage 3: This stage is similar to the second stage, where the network will output face bounding boxes and five facial landmarks’ positions.

This approach balances performance and accuracy, which is suitable for face detection in the classroom environment.

![Fig 2. The architectures of P-Net, R-Net, and O-Net, where “MP” means max pooling and “Conv” means convolution. The step size in convolution and pooling is 1 and 2, respectively.](image)

![Fig 3. Model structure. This network consists of a batch input layer and a deep CNN followed by L2 normalization, which results in the face embedding. This is followed by the triplet loss during trainin.](image)
2) FaceNet: FaceNet, an end-to-end network, can learn a mapping from face images to a finite dimensional Euclidean space where the distance directly corresponds to a measure of face similarity. The model structure of FaceNet is shown in Fig.3.

Benefit from the novel loss function, called Triplet Loss, this network uses a deep convolutional network trained to directly optimize the embedding itself, rather than an intermediate bottleneck layer. This approach is much greater representational efficiency by outputting only a 128-bytes facial feature vector per face.

We employ the pre-trained FaceNet to extract facial features that will be saved for further processing or analysis.

3) SVM Classification: Support Vector Machine (SVM) is a generalized linear classifier in a supervised learning manner. Some linear indivisible problems may be nonlinearly separable, that is, there is a hypersurface of the feature space separating the positive and negative classes. The kernel method is employed to map the nonlinearly separable problem from the original feature space to the higher-dimensional Hilbert space by a nonlinear function. Therefore, the nonlinearly separable problem is transformed into a linear separable problem. The nonlinear function is also called the Kernel function.

This paper employs SVM as a face classifier. In order to train SVM, face features extracted from FaceNet are used for supervised training the SVM classifier. In the attendance module, the trained SVM can recognize the detected student faces. The process of face recognition is shown in Fig.4.

3.2. Head Pose Estimation

Head pose estimation consists of two parts, which are face detection, head pose estimation.

1) Face detection: In computer vision and image processing, the Histogram of Oriented Gradient (HOG) [12] is a feature descriptor for object detection, which is formed by calculating and counting the gradient direction histogram of the local region of the image. The HOG feature is able to describe the local texture of the image, which is insensitive to direction, scale, illumination. In this paper, the HOG feature is combined with a SVM classifier to detect student faces in head pose estimation.

2) Head pose estimation: We employ the ERT face alignment algorithm to detect face feature points (a total of 68). The algorithm uses regression tree method to detect and locate the key points of the face. For the face in a single picture, the regression tree method is used to estimate the coordinates of the face feature points from a sparse subset, so that high-precision face alignment can be achieved. The detected face features are input into the trained SVM classifier to classify the head poses. There are five head poses: "front view", "left view", "right view", "up view" and "down view". Particularly for the students whose head posture is "front view", the human eye closure test is performed to judge whether the student blinks or closes the eyes.

4. Experiment

This section focuses on the implementation of the entire system, including the four parts: face recognition module and head pose estimation module, service software and database.

Now, we want to explain the workflow of the whole system. 1) We need to set up a LAN between a webcam, service software, recognition analysis modules and database. 2) Device management: Configure the control parameters and communication address of the webcam in service software to manage and operate the webcam. 3) Image or video acquisition: The data stream of real-time video is transmitted from the webcam to the service software through TCP, which broadcasts and decodes the video through the preview interface. So we can get real-time images or videos of the classroom online.
for subsequent analysis. 4) Recognition and analysis: the captured images are sent to face recognition module to identify students. Recognized students will be sent to head pose estimation module for the analysis of head poses. The results are transformed to database and service interface.

4.1. Hardware and Software Conditions
The hardware equipment of this system include a webcam with PTZ, a hard disk recorder, a 16-port Ethernet switch, two desktop computers and several network cables. Considering the high resolution requirements of the original images for recognition process, we use DS-2PT7D20IW-DE(20X) vision webcam in this system which is a 2megapixel infrared network HD PTZ dome camera with a hard disk recorder of HIKVISION for video storage and backup. We set up a LAN by an Ethernet switch for communication between various parts of the system. The service software is installed on one of high-profile desktop computers, and another computer is used to create and manage the database.

The software of the system includes implement of attendance module and head pose estimation module, the design of service interface and database.

4.2. Service Software
We design a service software for management and operation of the webcam, face recognition module and head pose estimation module. The overall architecture of the software is shown in Fig.5. The software is composed of four parts: Preview Interface, Recognition Interface, Configuration Interface and Database Interface. The user interface layer is connected to the business layer through the MFC API. The business layer consists of five parts: Adding Device, Preview and Image Capture, PTZ Control, Recognition and Analysis and Device Configuration. The data stream of real-time video is transmitted from the webcam to the service software through TCP, which broadcasts and decodes the video through the preview module. The standard photos of students and extracted feature files are transmitted from the database to the service software in the form of network shared files. This software communicates with the webcam via the TCP connection of the SDK interface. The overall architecture of the software is shown in Fig.5.

![Fig. 5. The overall architecture of service interface, which is also workflow of this system.](image)
Face recognition interface. This interface is set to display recognition results, including recognized student faces and their student IDs. The results will be saved as attendance record in database.

Head pose estimation interface.

4.3. Attendance Module
1) Dataset: With the support of the school's Academic Affairs Office, the experimental system was installed in classrooms: a high-definition webcam with PTZ, a network switch, a computer installed the recognition and analysis software. Because there is only one webcam and the classroom is large, we divide the classroom into 30 areas. In order to collect a face dataset, the webcam takes 90 classroom images at each area in turn. For each picture, we first detect student faces, then crop these images as 160*160 and save locally by MTCNN. Finally, we divide and label these face images according to student IDs. The dataset has a total of 123 categories, 10740 photos, which is used for the training of the SVM face classifier.
2) Test: We test the face recognition module in a large lecture hall and a medium-sized classroom. In the realistic classroom condition, students often bow their heads and cause facial obstruction, which has a great impact on the results of face detection and face recognition. In the large lecture hall, the average detection rate of MTCNN is 60%, and the average accuracy of face recognition is 72%. While, in the medium-sized classroom, the average detection rate of MTCNN is 72%, and the average accuracy of face recognition is 89.4%. The results of face recognition is shown in Fig.8.

4.4. Head Pose Estimation Module
In this module, we employ HOG detector of Dlib machine learning library to detect faces in the captured images. And Dlib ERT algorithm is used to output the key feature points from the detected student faces. To estimate head poses, the key feature points will be input to the trained SVM classifier. The results of head pose estimation are shown in Fig.9.

4.5. Database
We select MySQL as the database management system. The database storages basic information of students, teachers, courses, classrooms, cameras, etc. Also, captured images and recognition results will be sent to the database and saved.

5. Conclusions
This work presents an automated teaching evaluation system with incorporating with vision technologies and machine learning algorithms. In complicated classroom scene, traditional face recognition algorithms are susceptible to illumination, occlusion, and facial expressions et.al. To obtain accurate representation of the faces, we employ deep learning network FaceNet to output features
vector, which is more robust and accurate. Through above experiments in classrooms, FaceNet feature extractor works well in attendance module. For student behavior analysis, we select head pose as estimation objects, which laid the foundation for student behavior analysis. Next, we will focus on improving the accuracy of head pose estimation.

Acknowledgments
The paper is supported by Provincial key teaching research project of Anhui Province, China (2016jyxm1141).
We would like to thank Yige Guo, Jingrong Hu, Qiongwen Liang, Jinjian Jiang for discussions and great insights on the design of the whole system.

References
[1] Turk M, Pentland A. Eigenfaces for recognition [J]. Journal of cognitive neuro science, 1991, 3 (1): 71-86.
[2] Turk M A, Pentland A P. Face recognition using eigenfaces [C]//Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 1991: 586-591.
[3] Haddadnia J, Ahmadi M, Faez K. A hybrid learning RBF neural network for human face recognition with pseudo Zernike moment invariant[C]//Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290). IEEE, 2002, 1: 11-16.
[4] Wiskott L, Fellous J M, Krüger N, et al. Face recognition by elastic bunch graph matching[C]//International Conference on Computer Analysis of Images and Patterns. Springer, Berlin, Heidelberg, 1997: 456-463.
[5] Liu C, Wechsler H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition[J]. IEEE Transactions on Image processing, 2002, 11(4): 467-476.
[6] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition under varying pose."[C]//Proceedings of 23rd Image Understanding Workshop. 1994, 2: 837-842.
[7] Zhu Z, Ji Q. Robust real-time face pose and facial expression recovery[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006, 1: 681-688.
[8] Tong Y, Wang Y, Zhu Z, et al. Robust facial feature tracking under varying face pose and facial expression[J]. Pattern Recognition, 2007, 40 (11): 3195-3208.
[9] Xu T, Wang C, Wang Y, et al. Saliency model based head pose estimation by sparse optical flow[C]//The First Asian Conference on Pattern Recognition. IEEE, 2011: 575-579.
[10] Murphy-Chutorian E, Doshi A, Trivedi M M. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation [C]//2007 IEEE Intelligent Transportation Systems Conference. IEEE, 2007: 709-714.
[11] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. 2005.
[12] Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1867-1874.
[13] Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23 (10): 1499-1503.
[14] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 815-823.
[15] Platt J. Sequential minimal optimization: A fast algorithm for training support vector machines [J]. 1998.