A VISION-BASED INTELLIGENT SYSTEM FOR EYE CLINIC AUTOMATION

Lim Chang Siang, Li Zhenghao Kelvin, Zhong Xiaohui, Zheng Xiaolan

Institute of Systems Science, National University of Singapore, Singapore 119615

ABSTRACT

Vision loss poses a significant health risk, particularly in aging populations. It is closely linked to reduced quality of life, increased dependence, and worsened health outcomes. Timely detection and treatment can help delay the progression of vision loss. However, regular eye checks often face accessibility challenges due to the scarcity of trained specialists. To address this issue, our project aims to enhance the accessibility of routine eye health assessments by developing a vision intelligent system to enable automating visual acuity checks. By leveraging technology, we aspire to improve the availability and convenience of such assessments, ultimately promoting proactive eye care.

***Index Terms*—** machine vision, medical, automation, ophthalmology,

1. INTRODUCTION

Sight is often considered the sense that is most important and valued [1]. However, accurate measurement of a person's vision is difficult. This is not only because it is subjective, but also that vision itself has many facets. One of the universally used measure of vision is the visual acuity test.

Visual acuity is a core component of the eye examination. It provides the tester an objective measurement of the subject's vision. This is usually done via a reading off a visual acuity chart, while covering one eye at a time [2]. In this project, we aim to design a vision system to be able to detect a patient's pose during visual acuity examination. In particular, whether or not patients are wearing glasses, and whether the occlude is correctly positioned and used.

Accurate measurement of a person's vision is difficult. This is not only because it is subjective, but also that vision itself has many facets. One of the universally used measure of vision is the visual acuity test [2].

The visual acuity test consists of a display showing varying sizes of letters or characters, called optotypes (Figure 1), usually arranged in a vertical fashion in decreasing order of size. Patients are instructed to read, with each eye singly, till the smallest letter that they can see. For patients who require glasses to read, we would instruct the patients to read with their glasses on.

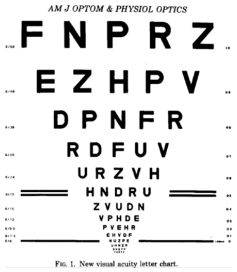


Fig. 1. Example of optotype used during visual acuity testing.

To ensure that the patients are using one eye in turn, they are instructed to hold an occlude (Figure 2) to cover the fellow eye. Some versions of the occluder have an additional pinhole segment to reduce the effect of refractive error.

Fig. 2. Occluder with pinhole.

2. LITERATURE REVIEW

Visual acuity test is based on a rule-based scorecard to provide an objective information about the person's vision health. The rule-based nature of the task makes it an attractive candidate for automation.

2.1. Related research works

According to the World Health Organization (WHO), there are over 2.2 billion people worldwide with varying degrees of vision impairment. It is estimated that at least half of these cases can be prevented through early screening and detection [3]. One effective examination method is the visual acuity check, traditionally performed in optician or specialist clinics

by trained professionals. However, this approach limits scalability, especially considering the increasing needs of an aging population. To address this emerging demand, several groups have explored the feasibility of digitalizing the Visual Acuity (VA) check, allowing easy access via smartphones.

EyeChartPro, for example, has developed an iPad application that enables patients to self-check their vision at home [4]. Although this study found that the iPad display method is as accurate as the traditional lightbox method for individuals with 20/20 vision, it showed greater variation for those with visual impairments. In another study, researchers developed the "V@Home" application, which allows patients to autonomously perform VA checks at home using their iPhones. The application includes audio analysis using Mel-frequency cepstral coefficient (MFCC) to analyze speech and provide automatic test results. Users receive instructions from the application to conduct the test. Evaluation of this application demonstrated high reliability and accuracy compared to routine care [5].

These works collectively demonstrate the potential for scaling the VA test through digitalization. However, it is worth noting that these studies were conducted in controlled environments, and concerns have been raised about ensuring patients follow the test guidelines and do not attempt to manipulate the results [4]. To the best of our knowledge, there is currently no digital VA application with patient monitoring capabilities to automatically observe patient actions, provide feedback, and offer suggestions for correction in real-world implementations.

2.2. Related commercial solutions

In the realm of self-check systems for visual acuity (VA), there exists at least one commercially available product. In 2008, a start-up called SoloHealth introduced a product called EyeSite, which is a standalone booth enabling individuals to assess their vision health. The pilot trial of this product involved testing 15,000 individuals over a span of 5 weeks. Notably, 25% of participants had never undergone an eye exam before, indicating the potential accessibility of self-test kiosks, particularly for individuals facing significant barriers to visiting a doctor's clinic [6].

3. SYSTEM REQUIREMENTS

In this study, we have selected Tan Tock Seng Hospital (TTSH) as our case study, specifically focusing on the Ophthalmology Specialist Outpatient Clinic (SOC) located in Singapore. To gain valuable insights, we conducted interviews with clinical experts in the field. Through these interviews, we were able to identify four key requirements that aim to address the challenges faced by the clinic and alleviate their pain points (Figure 3).

- Ability to confirm the person's identity through their identification document.
- Ability to confirm that the examination room has only one person present.
- Ability to determine if the person is wearing glasses or not.
- Ability to determine which eyes (left or right) is occluded using the occuluder tool.

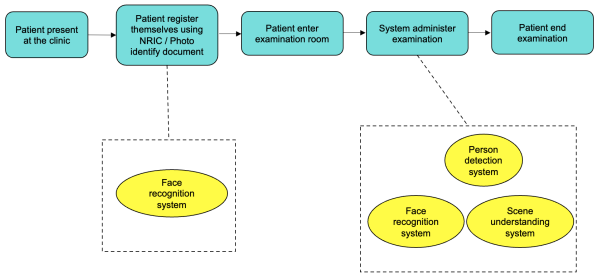


Fig. 3. Overview of system requirement

Based on the requirements, we proposed to develop 3 vision intelligent systems, namely,

- Face Recognition System: Confirm that the right patient is receiving care
- Person Detection: To ensure that there is only one person present for the examination
- Scene Understanding System: To provide context and interpretation on the patient actions

4. PROPOSED APPROACH

4.1. Person Recognition System

The person recognition system is used for 2 purposes:

- Patient Registration: The person recognition system is designed to register patients into the identity database by utilizing information from their National Identity Registration Card (NRIC). The system will effectively identify the portrait photo, detect the person's name, extract the facial features from the portrait, and securely store them within the database.
- Patient Re-identification: The system will accurately identify the person present in the scene, crop their face, extract the facial features, and utilize a gallery method to identify the closest matched identity for re-identification purposes.

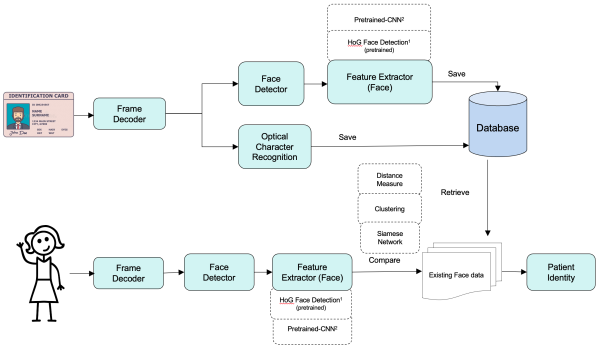


Fig. 4. Approach for Person Recognition System

In order to extract facial features, we employed a combination of techniques, including the use of Histogram of Gradient (HoG) and various Convolutional Neural Network (CNN) feature extractors. To compare two face images, we utilized multiple methods, such as cosine and Euclidean distance measures, as well as a Siamese network, to determine the most effective approach.

To extract the name information from an image, we utilized an established Optical Character Recognition solution (Tesseract) to extract the text. Additionally, we employed a rule-based approach to accurately determine the name string.

4.2. Person Detection

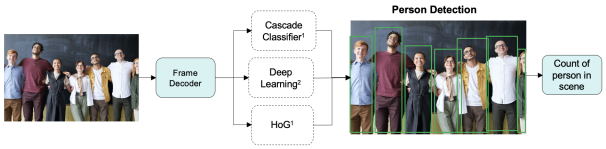


Fig. 5. Approach for Person Detection System

The use case for person detection is to count the number of human person present in the scene and report the number accordingly for downstream usage. For our solution, we intend to experiment both machine learning and deep learning approaches, including using Cascade Classifier, Histogram of Gradient (HoG) and existing Deep Learning model. Training of a custom cacasde classifier model require huge amount of data and computational power, we intend to first explore the feasibility of existing models for our purpose.

4.3. Scene Understanding System

The purpose of scene understanding system is to able to identify the scene with person into the following categories:

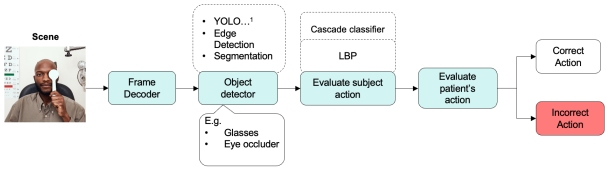


Fig. 6. Approach for Scene Understanding System

- Wearing Glasses
- Not Wearing Glasses
- Occulude Left Eye
- Ouuclde Left Eye with Pinhole on Right Eye
- Occulude Right Eye
- Occulide Right Eye with Pinhole on Left Eye

The identification of eye occlusion actions can be accomplished by detecting the presence of an "occluder" tool within the scene. To achieve this, we will explore a range of image processing techniques such as edge detection and segmentation, complemented by the implementation of modern object detection techniques like You Only Look Once (YOLO) [7]. These approaches will enable us to detect the object and ascertain the specific type of action occurring in the scene.

5. DATASET

To develop the person recognition system and person detection system, we utilized the Labelled Face in the Wild dataset, a publicly available resource [8]. Employing the scikit-learn package, we configured the parameters to exclusively include classes with a minimum of 100 available images. Consequently, we acquired a total of 1140 facial images representing 5 distinct person classes.

For the scene understanding system, we encountered a challenge as there were no readily available curated open-source datasets specifically designed for individuals wearing eye occluders. Therefore, in order to proceed with this particular project module, we took the initiative to create our own dataset. To ensure ethical compliance and patient privacy, the dataset collection process was conducted at the Tan Tock Seng Hospital Eye Clinic, with the assistance of volunteers, specifically administrative personnel. This approach allowed us to gather the necessary data while upholding the highest ethical standards.

All images were taken in a brightly lit indoor environment against a plain backdrop. Participants all had to wear masks as part of hospital policy. This is also important as the system is intended to be deployed in a clinical setting and it's expected that the end-users will be wearing masks. We also

collected images of participants who wore spectacles and participants without spectacles. This will also be representative of the population for which this system is intended for.

All images were captured using an Apple iPhone 11 camera (Cupertino, California, United States). The images were taken in portrait mode, at a resolution of 3024 x 4032 pixels, in RGB mode. Files were saved as .jpeg format in the database.

A total of 40 images were collected from 10 participants in 4 poses, this translates to 4 classes, each with 10 instances. These were taken consecutively and on the same day. The 4 poses are (1) right eye occluded without pinhole (2) left eye occluded without pinhole (3) right eye occluded with pinhole (4) left eye occluded with pinhole.

Fig. 7. Samples of person with occuluder images with human annotation using Roboflow [9]

Using images collected from volunteers, we employed an unsupervised segmentation tool called the Segment Anything Model (SAM) to generate segmentation masks for various objects present in the scene. From a dataset comprising 40 images, we successfully created a total of 1153 masks. Among these, 1092 masks corresponded to the "others" category, while 15 masks were classified as "left," 18 as "left_pinhole," 13 as "right," and 14 as "right_pinhole" (Figure 8). Since the dataset exhibited significant class imbalance, we performed oversampling on the underrepresented classes. Consequently, we obtained a final dataset comprising 5460 masks, with 1092 images allocated to each of the 5 classes (Figure 9).

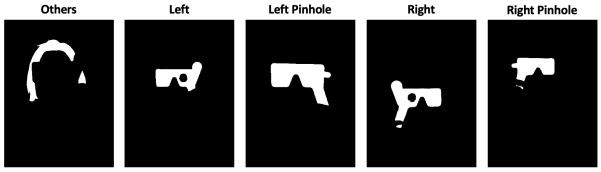
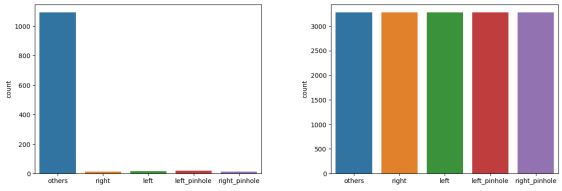


Fig. 8. Masks generated using Segment Anything Model



(a) (b)

Fig. 9. (a) Original Dataset, (b) Balanced using resampling approach.

6. EXPERIMENTAL RESULTS

6.1. Person Recognition System

In our person recognition system, we employed various feature extraction techniques for human face images. These techniques encompassed Histogram of Gradient (HoG), ResNet50, VGG19, VGGFace, and a customized CNN model trained using a Siamese network (Figure 10). To train our custom CNN model, we utilized 3380 pairs of images and evaluated its performance on 850 pairs of images. To ensure efficiency, we implemented an early stopping rule, terminating training if there was no improvement in accuracy after 10 epochs.

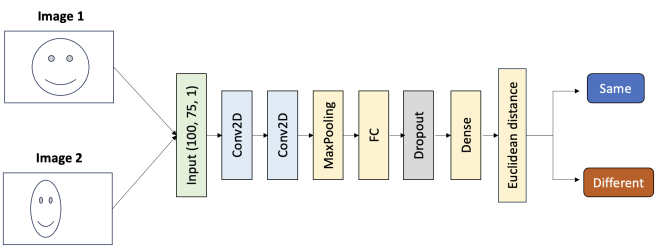


Fig. 10. Siamese model architecture

Table 1. Performance comparison for Person Recognition System.

Approach	Distance Measure	Accuracy
HoG	Cosine Similiarity	0.4647
HoG	Euclidean Distance	0.4647
ResNet50	Euclidean Distance	0.5017
VGG19	Euclidean Distance	0.3160
VGGFace	Euclidean Distance	0.6690
Custom CNN	Euclidean Distance	0.1101

For ResNet50 and VGG19, we utilized pre-trained weights from the ImageNet dataset [10] to perform feature extraction. VGGFace, renowned for facial recognition, was trained on human face images using the VGG19 architecture [11]. Our training process involved employing the Labeled Face in the Wild (LFW) dataset [8], and we evaluated the model's accuracy in correctly identifying image pairs.

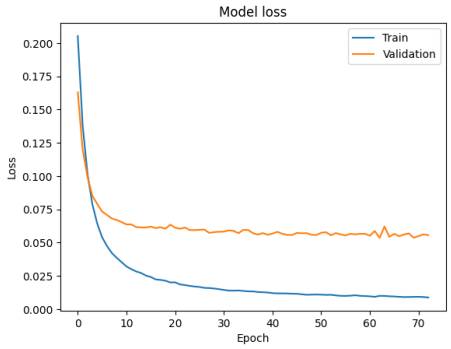


Fig. 11. Training loss with Siamese Network after 70 epoch

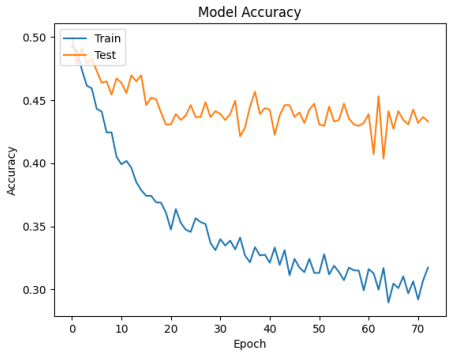


Fig. 12. Accuracy with Siamese Network after 70 epoch

Table 1 summarises the result of our experiment. Using HoG feature vector, we found no difference in using either Cosine similarity or Euclidean distance measure. As such, we decide to stick to only using Euclidean distance measure. In our evaluation, we found that VGGFace has the best accuracy performance of 67%, while our custom CNN model has the lowest score of 11% (Table 1).

6.2. Face Detection System

In our solution, face detection plays a crucial role in the following tasks:

1. Identifying individuals and confirming their identity against registered information.
2. Counting the number of people present in the scene, ensuring only one person is detected.

There are several face detection methods available, including:

- CNN
- Eigenfaces
- Fisherfaces
- Haar Cascade

Among these methods, we selected Haar Cascade due to its ease of deployment and fast detection capabilities.

The Haar Cascade algorithm is implemented through the `CascadeClassifier` class provided by the OpenCV library. This algorithm is based on the Viola-Jones algorithm, which is widely recognized for its efficiency in object detection, including face detection.

The method described above has certain limitations and challenges. To assess its reliability, several experiments were conducted, including variations in face angles and movement.

It was observed that this model is capable of detecting faces within a certain range of angles. However, when the face moves rapidly, the bounding box may not accurately display the face's actual location. In real-world scenarios, patients are typically instructed to minimize fast movements and face the camera directly, mitigating this issue.

Another factor tested was the presence of occlusion on the face. The model performed well in detecting faces with occlusion covering a substantial portion of the face. However, if the coverage is less than half, the accuracy of detection may be compromised.

Furthermore, the model demonstrated effectiveness under different lighting conditions, including both weak and bright light environments. Nevertheless, in practical scenarios, it is advisable to ensure sufficient brightness for optimal performance.

These experiments shed light on the strengths and limitations of the face detection method, providing insights for its application in real-world scenarios.

The `CascadeClassifier` utilizes a cascade of weak classifiers, which are simple and computationally efficient

classifiers. These weak classifiers are combined to form a strong classifier capable of effectively detecting the target object.

Fig. 13. Different angles of the face and movement

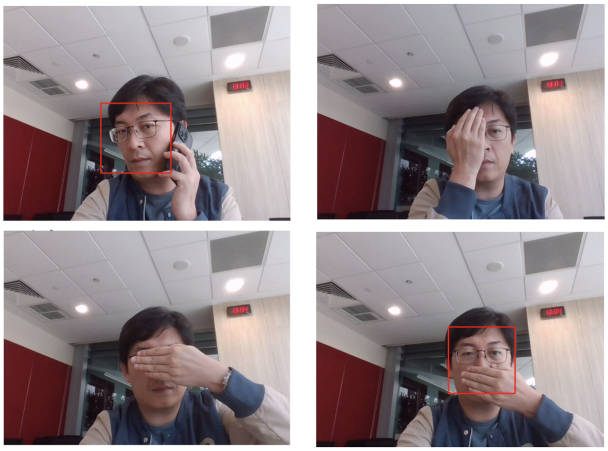


Fig. 14. 2) Occlusion of face

6.3. Scene Understanding System

In our scene understanding system, the glasses detection function plays a crucial role in verifying the patient's status before conducting the acuity test. In this project, we have explored hand-crafted techniques to detect the presence of glasses.

The traditional approach we employed focuses on examining the presence of a glasses bridge across the nose area. By determining the presence or absence of a bridge, we can deduce whether glasses are worn by the patient.

To begin, we locate the nose area of individuals using face landmark detection. This technique involves identifying and

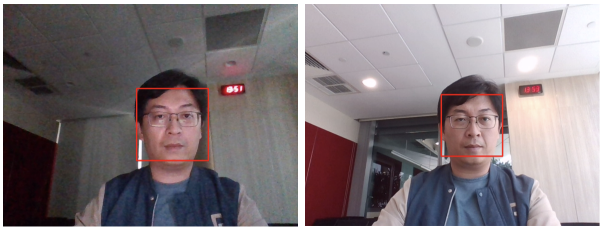


Fig. 15. 3) Different lighting condition

tracking key points on a person's face. Over time, the advancement of this technique has expanded the number of key points from a mere five to several hundred. In our project, we utilized the widely-used 68 key points face landmarks. To do this, we obtained a pre-existing .dat landmark file containing the information for these 68 points.

Within the 68-point landmark, the nose area is represented by points 28 to 31. We utilized the Dlib face detection model to acquire all 68 facial landmarks. Subsequently, we applied Gaussian blur and Canny edge detection to generate an edge profile for the area of interest. By examining the edge detection results, we can identify the profile of a glasses bridge, if present.

The final step involves determining the presence of glasses. We draw a vertical line across the nose area and analyze if any white-colored pixels are detected. The presence of white color signifies the existence of a glasses bridge, thus confirming the presence of glasses on the nose. Conversely, the absence of white color indicates the absence of glasses.

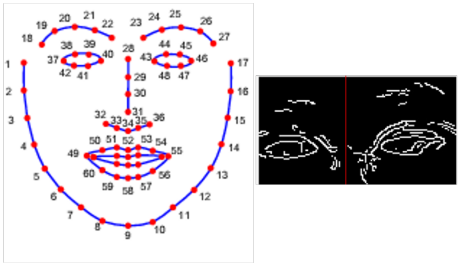


Fig. 16. Using Edge Detection to Detect Glasses

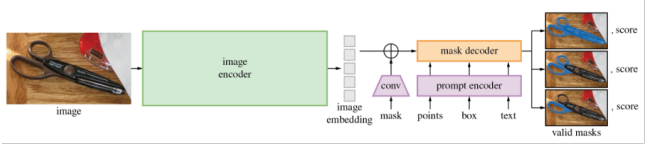


Fig. 17. Segment Anything Model

In order to determine if a person is holding an occluder

over their eyes, object detection was required to identify the presence of the occluder. Due to the relatively small size of our dataset, we opted for transfer learning instead of training the model from scratch.

We selected YOLOv5 as the convolutional neural network (CNN) backbone. YOLOv5 belongs to the You Only Look Once (YOLO) [7] family of computer vision models and was released by Ultralytics in 2020. YOLOv5 offers four main versions: small (s), medium (m), large (l), and extra-large (x), each providing progressively higher accuracy rates but requiring varying amounts of training time [12].

To begin, we utilized an online image annotator called Roboflow [9], to perform image annotation. The images were annotated with labels such as *RE_occluder*, *LE_occluder*, *RE_occluder_PH*, and *LE_occluder_PH* (Figure 7). Once the annotation process was completed, the labels were downloaded in YOLO's labeling format.

The images were resized to a standard size of 640 x 640 pixels, which is the default size for YOLOv5. Subsequently, the dataset was divided into training and testing sets. Our initial training involved a new classifier specifically for the four classes, while keeping the remaining layers frozen. The training parameters were set as follows: batch size of 10, 30 epochs, and the yolov5s6 model was utilized. We chose yolov5s6 because it strikes a good balance between speed and accuracy. Notably, the yolov5s6 model includes an additional output layer designed for detecting larger objects. These models benefit greatly from training at higher resolutions, resulting in improved detection outcomes.

However, in our specific case, we encountered suboptimal results. Despite our best efforts, fine-tuning the model proved to be challenging and ultimately unsuccessful. During the testing phase, the YOLOv5 object detector failed to accurately detect the occluder.

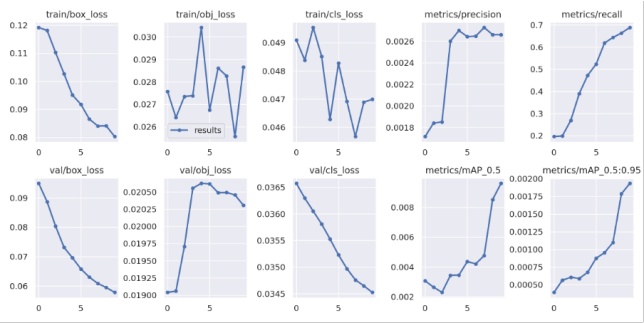


Fig. 18. YOLOv5 evaluation metrics

Next, we tried using the Segment Anything Model (SAM). Segment Anything Model, introduced by Meta in 2023, is a powerful tool for image segmentation. It can effectively segment images into masks without requiring additional training and exhibits impressive zero-shot performance across various

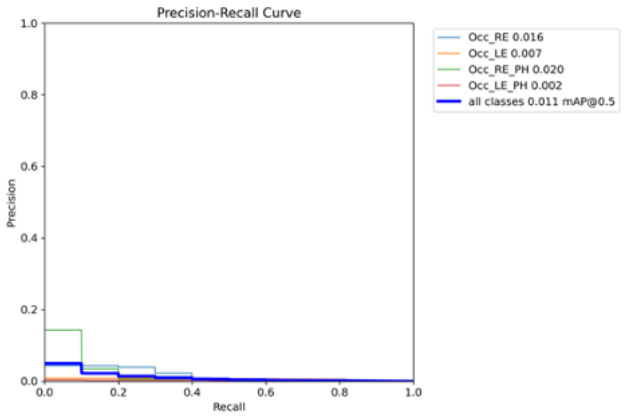


Fig. 19. YOLOv5 Precision-Recall Curve

segmentation tasks. Trained on the extensive SA-1B mask dataset, the model consists of a featurization transformer block and a decoder head that supports prompts. The model's architecture can be visualized using the diagram below.

We opted to leverage this model due to its outstanding segmentation capabilities. The usage of the model is straightforward. Following the installation of necessary dependencies and downloading the model's Git files, which include the weights, the SAM model is loaded into memory. We specifically chose the smallest ViT-B encoder to prioritize faster inference speed. To automatically generate masks, the `SAMAutomaticMaskGenerator` function is employed. This function returns a list of masks, with each mask represented as a dictionary containing various information pertaining to the mask.

Our objective is to create a dataset consisting of occluder and non-occluder masks. This dataset will be utilized for a subsequent SVM classifier to determine the orientations of occluders. Once the masks are generated, we iterate through each mask and save them accordingly.

While this method proves effective, we encountered a limitation when applying this approach to videos. It became evident that SAM's processing speed is insufficient for real-time video detection. The generation of masks for each frame takes several seconds, resulting in a delay in displaying the classification results in the video.

Nevertheless, we used the segmentation masks generated by the SAM model and trained a multi-class SVM classifier. Through confusion matrix evaluation, we observed that the classifier excelled in effectively distinguishing between different classes, yielding an impressive accuracy rate of 99.9% (Figure 20). This outcome showcases the efficacy and reliability of our approach in achieving highly accurate and precise classification results.

To address the limitation of slow processing speed, we opted to utilize a thresholding method for generating segmen-

tation masks in real-time processing. By incorporating the thresholding approach with SVM, we were able to achieve a processing speed of approximately 40 to 50 frames per second, facilitating real-time video analysis. However, we observed that the accuracy did not match our expectations in the final implementation.

This discrepancy in accuracy can be attributed to the inherent noise and reduced precision of the segmentation masks created through the thresholding method, as compared to the masks generated using SAM. Unfortunately, due to time and resource constraints, it was not feasible for us to create a new set of masks for training purposes. As a result, we focused on fine-tuning the hyperparameters for the threshold to optimize both accuracy and prediction stability.

Though we faced challenges in achieving the desired accuracy, we made the most of the available resources and implemented a solution that provided real-time video analysis.

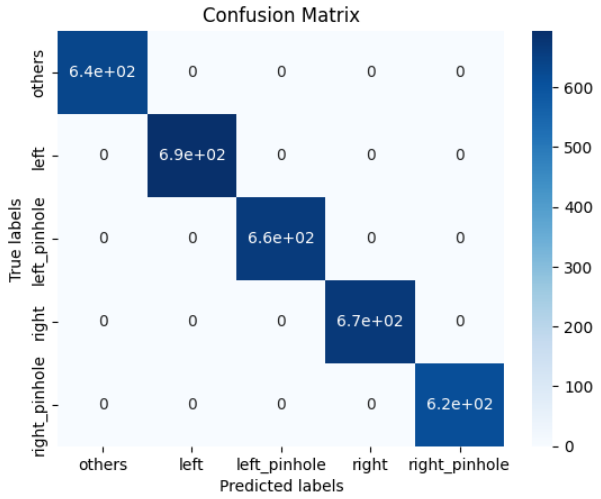


Fig. 20. Confusion Matrix of SVM classifier for action recongnition

6.4. Discussions and limitations

In this project, our primary objective was to develop a robust and comprehensive vision intelligent system capable of automating patient monitoring in the eye clinic. Our system encompasses multiple functionalities, including the ability to detect and extract identity information from an image of an identification card, detect and analyze portrait images, and perform face recognition for patient identification.

Throughout our development process, we conducted comprehensive evaluations of various approaches for feature extraction. Amongst the different feature extraction models, VGGFace emerged as the top performer, demonstrating superior performance in our evaluations.

However, during deployment, we encountered a significant limitation with our model. It became evident that the portraits extracted from the identity cards often differed greatly from the real-world appearance of the individuals. Most facial recognition systems rely on images captured directly from a person's face, incorporating various angles to gather an ample amount of facial features for robust reidentification. Without a suitable data augmentation technique, relying solely on a single portrait image for effective person reidentification can be an exceptionally challenging task.

The discrepancy between the identity card portraits and real-world appearances poses a significant obstacle to accurate and reliable recognition.

To ensure that our system is not only accurate but also capable of real-time analysis, we made careful considerations in selecting computation-efficient methods. For instance, we leveraged the well-established Haar-Cascade classifier for efficient person detection. Additionally, we employed the canny edge detection technique to effectively identify the presence or absence of glasses.

To facilitate scene understanding and action detection, we adopted a single-frame analysis approach, utilizing image segmentation masks classified by a support vector machine (SVM) classifier. Initially, our intention was to employ the Segment Anything Model (SAM) for real-time image segmentation. However, during testing, we encountered significant delays in generating masks for a single frame. Despite our attempts to optimize performance by reducing image size or cropping, these efforts proved ineffective. Generating masks for each frame required a minimum of 10 seconds, falling short of the real-time analysis requirement of 20 to 30 frames per second. Consequently, we had to pivot to an alternative approach utilizing a simple thresholding method for mask creation.

While we achieved reasonably accurate predictions through hyperparameter adjustments, masks created via thresholding were inadequate for distinguishing between classes with subtle differences (e.g., "left" versus "left_pinhole"). This limitation arises from the fact that masks generated using the thresholding method contain significant noise and lack the precision offered by the SAM model.

Despite the challenges faced, we made the necessary adjustments to ensure some level of accuracy in our predictions. However, it is important to acknowledge that the precision and distinction provided by the SAM model's masks were not fully replicated using the threshold approach.

7. CONCLUSIONS AND FUTURE WORK

In summary, we have developed a vision intelligent software solution capable of performing facial recognition, person detection, and recognition of patient actions. However, the current state of our solution is limited to a very specific and controlled environment.

To enhance its performance, future efforts should focus on the redesign of the patient registration workflow, enabling the collection of more facial features for a more robust person re-identification process.

Additionally, the current approach of using a single frame combined with image segmentation for scene understanding is suboptimal. To address this, future work should concentrate on creating a comprehensive training dataset and utilizing computational efficient segmentation techniques suitable for real-time analysis. Moreover, it would be beneficial to explore other real-time video analysis techniques, such as multi-frame analysis, to further improve the system's capabilities.

8. AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows:

- **Lim Chang Siang**, conceptualized, literature review, system design, draft manuscript, experiment, system implementation and integration, performance evaluation and analysis, person recognition system, scene understanding system.
- **Li Zhenghao, Kelvin**, conceptualized, literature review, product requirement specification, draft manuscript, data collection, domain expertise, scene understanding system, performance evaluation.
- **Zhong Xiaohui**, conceptualized, requirement analysis, product requirement specification, system design, draft manuscript, experiment, person detection system.
- **Zheng Xiaolan**, system design, draft manuscript, literature review, experiment, implementation, performance evaluation and analysis, scene understanding system.

All authors reviewed the results and approved the final version of the manuscript.

The authors confirms sole responsibility for the following: solution conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

9. REFERENCES

[1] Jamie Enoch, Leanne McDonald, Lee Jones, Pete R. Jones, and David P. Crabb, "Evaluating whether sight is the most valued sense," *JAMA Ophthalmology*, vol. 137, no. 11, pp. 1317–1320, November 2019.

[2] Paulus T. V. M. de Jong, "A history of visual acuity testing and optotypes," *Eye*, 2022.

[3] World Health Organization, "Blindness and visual impairment," 2021.

[4] Zhao-Tian Zhang, Shao-Chong Zhang, Xiong-Gao Huang, and Ling-Yi Liang, "A pilot trial of the ipad tablet computer as a portable device for visual acuity testing," *Journal of Telemedicine and Telecare*, vol. 19, no. 1, 2023.

[5] Xiaotong Han, Jane Scheetz, Stuart Keel, Chimei Liao, Chi Liu, Yu Jiang, Andreas Müller, Wei Meng, and Mingguang He, "Development and validation of a smartphone-based visual acuity test (vision at home)," *Translational Vision Science and Technology*, vol. 8, no. 4, August 2019.

[6] US Pharmacist, "Solohealth launches interactive vision test kiosk," 2021.

[7] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[8] Gary B. Huang, Marwan Mattar, Honglak Lee, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep., University of Massachusetts, Amherst, 2008.

[9] Roboflow, "Roboflow," Accessed 2021.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," pp. 248–255, 2009.

[11] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," *British Machine Vision Conference*, 2015.

[12] Ultralytics, "Yolov5: A state-of-the-art object detection system," https://github.com/ultralytics/yolov5, 2020, Accessed: May 15, 2023.