IMPROVING COARTICULATION PERFORMANCE OF 3D AVATAR AND GAZE ESTIMATION USING RGB WEBCAM

BY

KUANGXIAO GU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Adviser:

Professor Thomas S. Huang

Abstract

This thesis explores two applications of computer vision in psychology-related studies: enhanced patient portal messages using 3D avatar and gaze estimation using a single RGB camera. The first application aims to help patients, especially those with poor health and low medical literacy, to understand messages delivered by patient portal systems by enhancing the messages with a 3D avatar. The avatar is built from real human face images and can deliver both semantic and emotive information, the latter of which is expected to help the patients to get a better, gist-level understanding of the portal messages. The second application aims to estimate eye gaze direction with an RGB camera. Preliminary results show the potential of the proposed method, although rigorous quantitative evaluation still needs to be done. While the proposed method cannot achieve the resolution and accuracy of commercial eye trackers, it is able to greatly reduce the cost since only one RGB camera is required.

Table of Contents

Chapter 1: Introduction

Computer vision can be dated back to the Summer Vision Project in the 1960s. At that time, the task of computer vision was oversimplified to building a visual system that could segment different parts of objects in an image. This task is very simple for human beings since we can always give accurate and clear boundaries of different objects. However, it turns out to be much harder for computers. Most people at that time thought computer vision was about processing the lights and colors in an image and using that information to perform different tasks. However, the truth is much more complicated, involving the basic definition of the concept “object”. Suppose we want to segment a table from the background in an image. This seemingly simple problem is complicated by the fact that tables might differ from each other in color, size and style. Even the same table may have different shapes in two images. In order to successfully find the boundary, a computer needs a set of rules that can take into account all those variations. In this example, a table is simple yet very hard to define due to the large variation between different kinds of tables and the way they appear in an image. All those factors need to be taken into account, which complicates the problem.

One way to solve this problem is to define a set of rules that take as many discriminative factors as possible into account. This kind of approach is called rule based, or heuristic, since the rules are inspired by existing concepts. Sometimes this approach can be very effective, especially when the concept we are considering is very simple. Unfortunately, most concepts have too many variations and it is very hard to come up with a set of rules to cover all those varieties.

Another approach is to let the computer learn the concept by itself: this is known as machine learning. For example, we want a computer to detect a car. First we need to train the computer by providing many labeled images of different cars. Sometimes the representation of a car is extracted from an image, sometimes the whole image (every pixel) is used as a representation. Once training is done, the computer has established some rules and features to detect a car and therefore is able to perform the task.

Despite the complicated nature of computer vision, many fields have successfully applied this technology in practical use. In industry, computer vision technologies are widely used for automation and inspection of products. For example, the auto manufacturer Rover used CV to inspect the outlines of its vehicles and match them with the CAD model, which can achieve 0.1 mm accuracy. Hewlett-Packard integrates CV into its digital printers for defect detection, color calibration and alignment. In medicine, computer vision technologies are used to detect and localize tumors in medical images. Computer vision can also be found in everyday life. Some vehicles are equipped with active safety devices, which utilize CV to help drivers stay in the middle of the lane or activate the emergency brake to avoid collisions. Many smart cameras have face detection, and some can even label each face with the corresponding name. As we can see, computer vision is now widely used and it helps to make our lives much easier.

In this thesis, two applications of computer vision in psychology-related studies are discussed. The first one is emotive facial expression synthesis during speech with application in personal health record systems (PHRs), which are defined as follows:

"An electronic application through which individuals can access, manage and share their health information, and that of others for whom they are authorized, in a private, secure, and confidential environment." [1]

PHRs have been listed as a top priority by the U.S. Secretary of Health and Human Services, the National Coordinator for Health Information Technology and the Administrator of the Centers for Medicare and Medicaid Services. Traditional PHRs feature patient portals where patients can manage their medical records and access medical information. Over 70 percent of consumers wish to use portal systems to remotely access medical test results, according to an article from the Institute of Medicine (IOM) (as cited by Peter Kuhn [2]). In addition, physicians can provide feedback and recommendations to their patients through portals in a timely fashion. Tang et al. [1] have shown that many consumers have high satisfaction levels with earlier versions of PHRs. One important reason is that this medium enables patients to do self-care based on the information provided by the system, and it serves as a bridge between patients and their healthcare providers.

Despite the fact that many patients have a positive attitude toward such systems, most PHRs are underutilized. The most important reason is that many patients lack the technical knowledge to understand the messages and results delivered by the portal system. According to the IOM [3], nearly half of American adults have difficulty understanding and using written health information. This greatly reduces the usability of PHRs and puts a heavy load on hospitals since patients tend to return to their hospital to seek advice.

The proposed 3D avatar portal is one promising solution to this dilemma as such an approach is expected to increase the usability of the portal system. The 3D avatar is basically a facial expression and speech synthesizer, which takes plain text and emotion markers as input. The output of the system is a “talking head” with facial expressions and lip movements matching the text and specified emotions. Along with an emotive speech synthesizer, this system simulates a real human talking, which is able to deliver both semantic and emotive information. This system can be applied to PHRs to enhance portal messages, which simulates physician-patient face-to-face communication and brings the gist-level information to the patients to help them get a better understanding of their medical results.

The second application is eye gaze direction estimation. The estimation of eye gaze direction has always been an interesting problem in computer vision and it plays an important role in psychological studies. Although there are many commercially available eye trackers, most of them are very expensive due to the exclusive hardware being used to guarantee performance. Despite the high accuracy and resolution of those commercial eye trackers, many researches in psychology cannot fully utilize their advantage. For example, some psychology experiments only need to know which of the nine pictures shown on a screen a test subject is looking at. In such cases, the commercial eye tracker will be overkill because only a rough gaze direction estimation will be sufficient to decide which picture the test subject is looking at. The goal of this application is to use current existing hardware commonly found in a typical office, namely an RGB webcam and a computer, to achieve rough gaze estimation. This will enable psychology researchers to use simple setups to conduct experiments and therefore greatly reduces costs.

The remainder of this thesis is organized as follows. In Chapter 2, related works in applications of avatars in medical fields and gaze estimation are reviewed. In Chapters 3 and 4, the technical details of the 3D avatar system and the gaze estimation system are discussed, respectively. In Chapter 5, experimental results of those two applications are discussed. Chapter 6 summarizes the work done in this thesis and provides suggestions for future work. Chapter 7 gives the conclusion.

Chapter 2: Related Works

Many researches have been done in both applications discussed in this article. For the application of avatars in medical systems, Huckvale et al. [4] used an emotive avatar to help patients with schizophrenia during therapy. The therapy was basically a conversation between the patient and an avatar, where the voice of the avatar was that of a therapist in another room. At the beginning of the conversation, the attitude of the avatar was set to be abusive. As the therapy went on, the patients were encouraged to stand against their avatars. At the same time, the avatar's attitude became less abusive and more helpful and supportive. The result shows that the patients felt better at the end of the communication.

Lisetti et al. [5] used a 3D emotive avatar to help patients become aware of their unhealthy lifestyle and provide suggestions along with guidance. The authors used an emotive 3D avatar and speech synthesizer to deliver messages to the patients. During the interaction with the avatar, the system monitored the emotion of the patient, and appropriate adjustments of the avatar's emotion were made by analyzing images of the patient's face with a face recognition engine. The evaluation result shows that 75% of the human testers felt interacting with the avatar was as comfortable as or more comfortable than interacting with a human.

Another application was studied by Bickmore et al. [6]. In their work, a 2D avatar was built, the main purpose of which was to teach patients about their post-discharge and to provide self-care instructions. The motivation was that around 20% of patients discharged from hospitals in U.S. end up returning to their hospitals, mainly due to insufficient medical literacy to understand self-care instructions. Despite the simple appearance of the 2D avatar, the result was promising. The most important reason pointed out by the patients was that unlike human beings, the 2D avatar could repeat the medical information as many times as the patients required. Furthermore, the patients were able to get more detailed information from the avatar if necessary. Those two advantages made the 2D avatar system a very effective tool for both instructing patients with low medical literacy and freeing the physicians from the burden of that instruction.

For the gaze direction estimation, various methods have been proposed and studied. Generally speaking there are two kinds of approaches: feature-based and appearance-based.

For the feature-based approach, Li et al. [7] estimated gaze directions by the shape of the pupil limbus. The algorithm started by finding the pupil limbus which corresponds to the points with the strongest intensity gradient along radial directions going outwards from the assumed pupil center, where the assumed pupil center is chosen as the pupil center from the previous frame. After finding several candidate points on the pupil limbus, the RANSAC algorithm followed by ellipse fitting was employed to find the ellipse parameters, which were finally used to estimate the gaze direction. This method is very effective when using an IR head-mounted camera, which is able to provide high resolution images with great contrast.

Another approach was studied by Wang et al. [8]. In their approach, an eye model parameterized by iris radius was built first. Meanwhile, the upper and lower eyelids were modeled and fitted onto two parabolas. Then the iris boundary between those two eyelids was extracted by applying a set of morphological filters followed by edge detection. In the end, an ellipse was fitted onto the iris boundary and the final gaze direction was computed from the ellipse parameters.

In most commercial eye trackers, however, a method known as cornea reflection is most widely used. The idea is to illuminate the eye region with an IR illuminator. Then due to the reflective property of the cornea, a bright dot can be observed and easily localized in the eye region. After that, this white dot is selected as a reference point and the relative position between the iris center and this point can be used to estimate the gaze direction. This method requires more dedicated hardware to work, which makes it more expensive. But the estimation accuracy and resolution are very high, usually less than 0.5 degrees.

Unlike the feature-based approach, the appearance-based approach does not explicitly extract detailed features from the eye image. Instead, the whole eye image is used “as a whole part” to estimate the gaze direction. For example, Cadavid et al. [9]

used an appearance-based method to classify baby gaze direction as looking at the camera or looking away from the camera. In their approach, the face region was tracked using the Active Appearance Model (AAM). Then this face region was normalized and the eye region was extracted from it. Due to the large pixel number of the eye region, dimensionality reduction using Laplacian eigenmap was performed on the eye region to generate a reduced feature vector. Finally, this feature vector was used to train a SVM classifier. Their approach was tested on eight subjects and the reported average accuracy was 91%.

Chapter 3: Audio-Visual Emotive 3D Avatar

3.1 Platform Description

A 3D avatar is a graphical representation of a person. The avatar can change facial expressions and make lip movements during speech, which makes it look like a real person talking. The content of speech, along with the emotion expressed by the avatar, will be specified by the user. For example, if the physician wants the avatar to deliver a good test result, he needs to provide the test result as plain text and specify the emotion to be expressed by the avatar, which is happy in this case. After that, the avatar will read out the test result with a happy facial expression in a happy voice.

The construction of the avatar is divided into two steps: Generating 3D avatar from 2D image and animating the 3D avatar.

The first step was completed by Tang et al. [10]. Given a 2D frontal image of a person's face, the algorithm first localizes several keypoints on the face image. Then a generic 3D model is deformed to match the keypoints of the 2D image, where the points in between those keypoints are interpolated. Finally, texture mapping is performed on the 3D model, which gives us a realistic 3D avatar. This process is demonstrated in Figure 1.

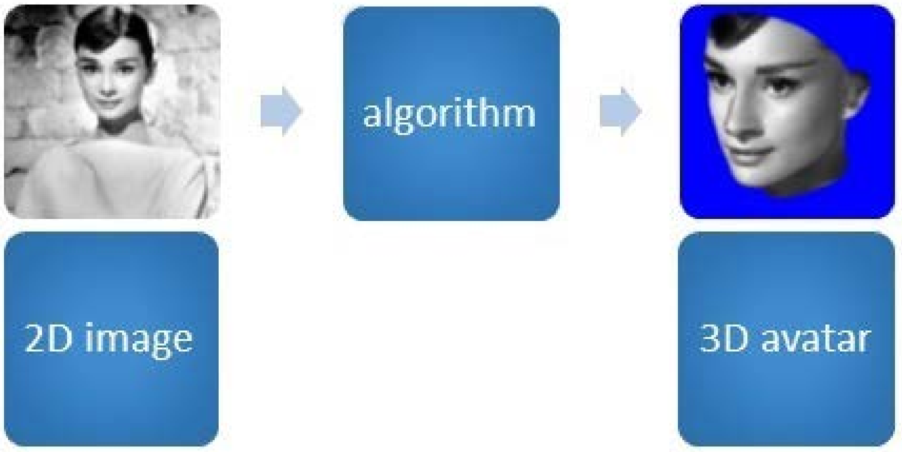


Figure 1: Avatar construction from 2D face image.

After the 3D avatar has been constructed, the next step is to animate it. A program which takes the avatar from the previous section and animates it using plain text with emotion markers has also been implemented previously by Tang et al. [10]. Specifically, the avatar will speak the text in different emotions specified by the emotion markers. The lip movement and expression of the avatar will also match the given text and emotion markers so that the avatar will actually look like a real person talking.

For the text-driven portion, a text-to-speech synthesizer is utilized to convert the plain text into phonemes that will be used to drive the avatar's lip movement. These phonemes will also be used to drive a speech synthesizer to produce the audio. The emotional markers will be used to change the avatar's expression during the speech.

Finally, the synthesized audio and video are synchronized by aligning keyframes and interpolation. Specifically, each keyframe corresponds to one phoneme (or viseme in video synthesis) and the gap between two adjacent keyframes is determined by the duration of that phoneme. For video synthesis, time interpolation on facial landmarks between two keyframes is used to generate the animation. For audio synthesis, a unit selection approach is used, which simply concatenates different phoneme segments to generate the audio sequence. Figure 2 demonstrates the synchronization mechanism.

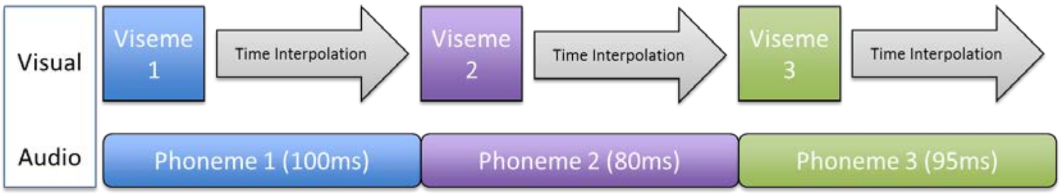


Figure 2: The synchronization of video and audio sequence in the avatar system.

3.2 Visual Performance Improvement

One big challenge for a 3D avatar is the co-articulation problem: how to combine the lip movements, which are controlled by both speech and the expression, so that the final result looks natural. The current avatar program performs well at neutral expression. However, when other expressions are combined with speech, the overall appearance looks unnatural. The reason is that the current method combines the lip movement and expression linearly, whereas the rules for realistic lip movement under expression are often much more complicated.

In the original program developed by Tang et al. [10], there are six templates corresponding to six basic emotions (angry, afraid, disgusted, happy, sad, and surprised). The appearance of the avatar is changed according to those templates. Meanwhile, another set of templates corresponding to different visemes is used to change the avatar's lip shape. The best method for combining these two templates in order to achieve natural-looking facial movements and speech characteristics remains unclear, however. The current program, which simply combines the templates linearly, does not perform well, especially for emotions which have a greater influence on the lip shape.

Therefore a new template-based method is proposed here to generate lip movement in combination with different expressions. Specifically, instead of defining templates for lip movement and for expression separately, we propose to store a template for each viseme under each expression. By doing this, we no longer need to worry about how to combine the lip movements resulting from expression and speech. Instead, we simply use the corresponding template based on the current viseme and expression. Figure 3 illustrates the difference between these two approaches.

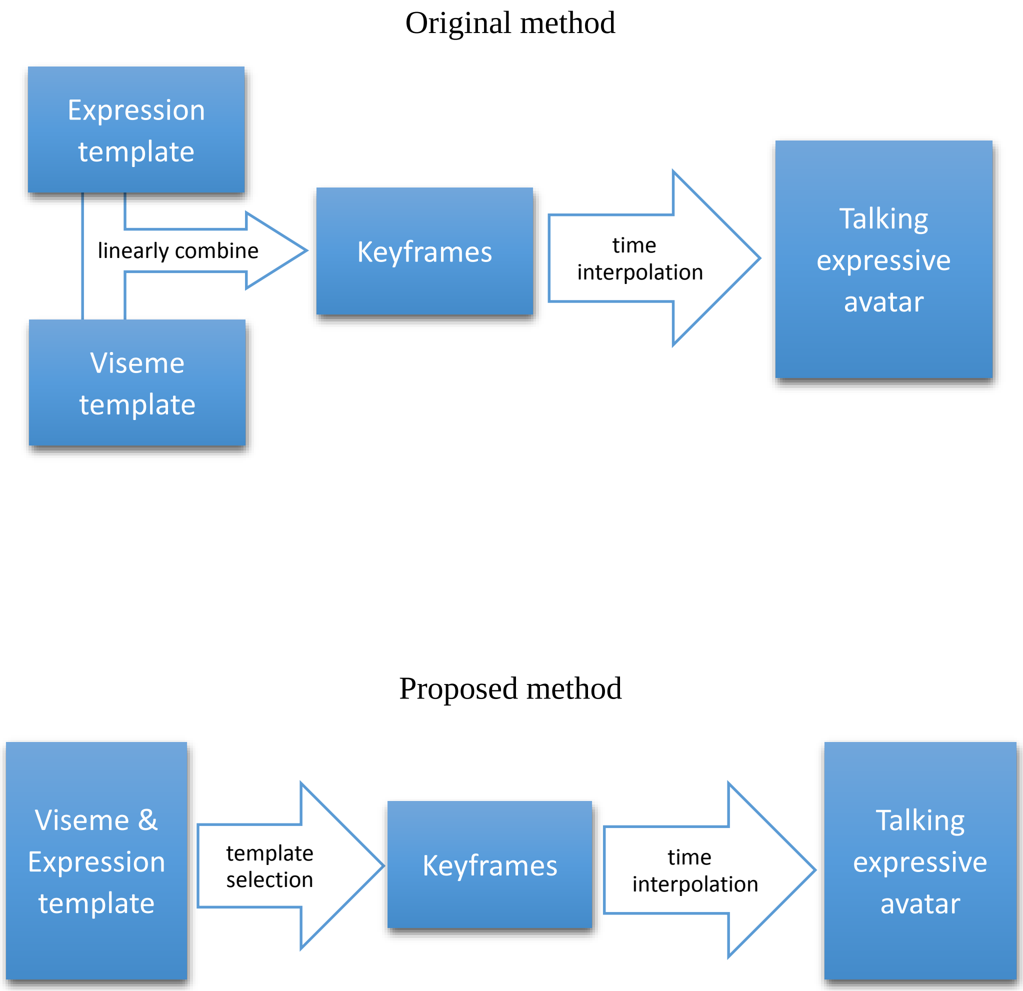


Figure 3: The proposed method bypasses the difficulty of combining lip movement due to speech and expression, which is typically represented by a complex model. Instead, we select the template which describes the lip shape for each viseme under each expression to construct keyframes and then interpolate in the time domain to generate the final talking avatar video.

Chapter 4: Eye Gaze Direction Estimation

4.1 Eye Model and Template

The proposed method is based on a simple eye model, which assumes a sphere shaped eyeball and circle shaped iris. This eye model has three parameters: iris radius, eyeball radius and the distance between iris plane and eyeball center. Figure 4 demonstrates the side view of this model.

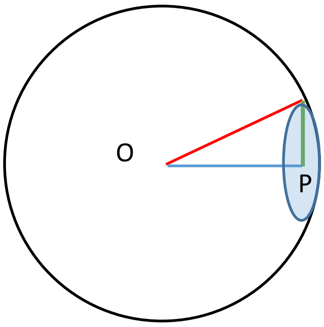


Figure 4: Side view of a human eye. Point O is the eyeball center and P is the iris center. The light blue area represents the iris plane. The green line represents the radius of the iris. The red line represents the radius of the eyeball, which is set to 2.1 times the radius of the iris. The length of the blue line equals the distance between eyeball center O and the iris plane.

The iris radius is measured from the image taken by the webcam. Specifically, an eye detector is applied first to get an eye image at the first frame. Then a circle detector is used to find the circles in the eye image. This requires the user to look into the camera at the first frame. Multiple circles may be detected and the one with the largest radius is assumed as the iris boundary. This assumption works since other false alarm circles are mostly small, which are from the image noise. So picking the largest one will give the iris boundary. After the iris radius has been decided, the eye model can be constructed. Specifically, the eyeball radius is set to be the iris radius multiplied by an empirically determined constant, which is 2.1, since most human eyes share similar size and structure. The iris-plane to eyeball center distance is computed from the geometric relation.

Once we get the eyeball model, the appearance template can be constructed. This step is completed by rotating the eyeball model by a certain angle in 3D and finding the corresponding appearance of the template after rotation. First, a center template which represents the eye looking at the center is defined. It is simply a pattern which can be locked on to the iris boundary in the eye image. Two kinds of templates are tested here:

- The first template looks like a 2D ring-shaped Laplacian of Gaussian (LoG). In fact, the template is constructed by computing a 1D LoG and then expanding it to a symmetric ring in 2D, such that the profile of a radial slice in 2D equals the 1D LoG. When convolved with an edge map of an eye image, the resulting value of convolution will be large when the white ring in this template is aligned to the edges of the iris. The edge map of the eye image can be computed using oriented filters or simply compute the 2D gradient of the eye image.

  The center template and some rotated templates are shown in Figure 5.

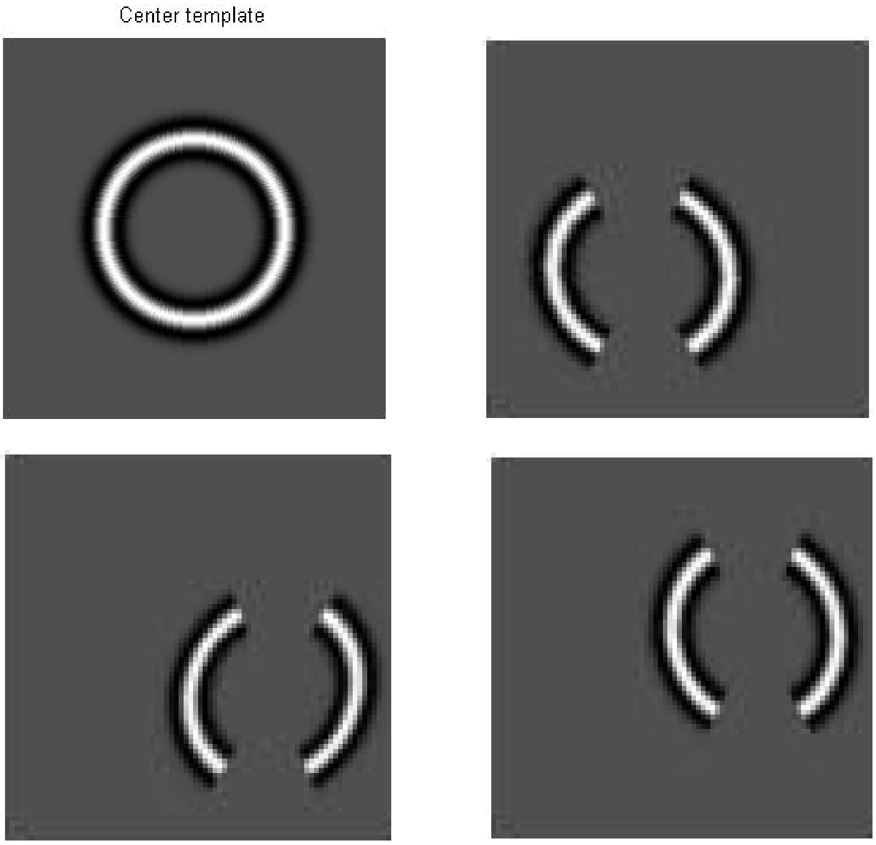


Figure 5: The center template (top left) and rotated templates. Note that the top and bottom part of the rotated templates are removed to avoid the eyelids.

- The second template is a symmetric ring-shaped derivative of Gaussian with the inner part having negative value and the outer part positive value. Since the iris region is very often darker than the sclera, this template will be able to lock onto the iris boundary using the original eye image.

  The center template and some rotated templates are shown in Figure 6.

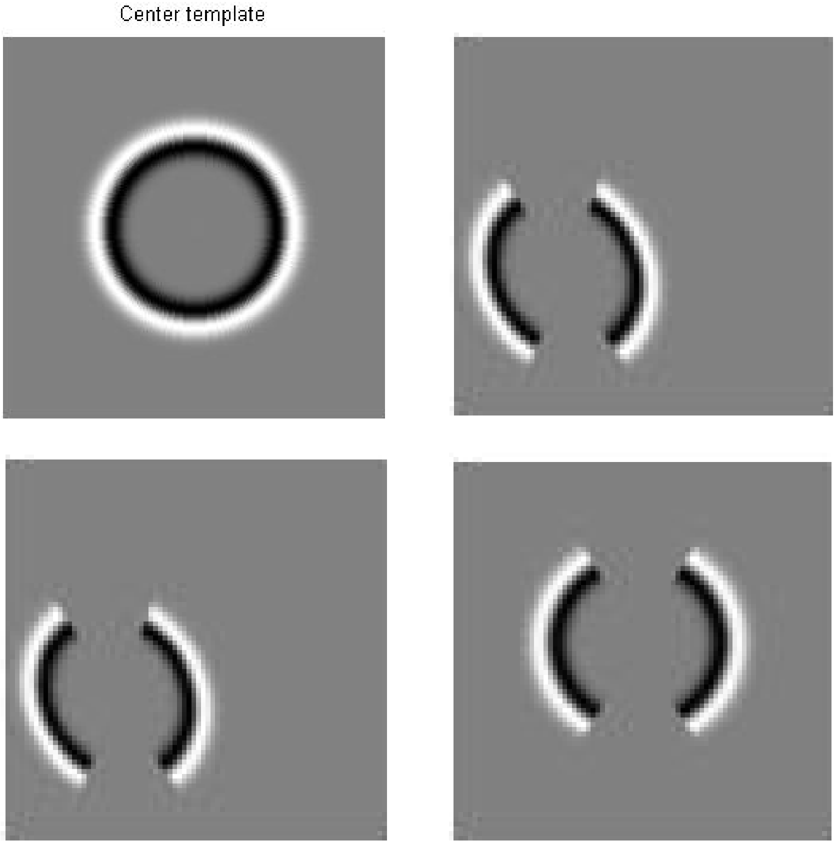


Figure 6: The center template (top left) and rotated templates. Note that the top and bottom part of the rotated templates are removed to avoid the eyelids.

4.2 Tracking the Eye Region

The proposed method does not extract the iris boundary. Instead, it tries to find the template that best matches the raw eye region image. Since the templates are created with the eyeball located at the center of each template, they can only be applied to eye images with the eyeball at the image center. As a result, each eye region image must be aligned and stabilized.

The alignment is achieved by tracking a facial landmark and using it as a reference point to get the eye region image, assuming the relative position between the eye and the landmark remains constant. The facial landmark used here is the nose. Assuming the user keeps his head steady, the shape of the nose will not change and its position relative to the eye region will be fixed. The tracking method is similar to Lucas-Kanade tracking, except that only one feature point (centered at the nose image) with a relatively large neighbor window (20-by-20 pixels) is used. The original Lucas-Kanade tracking with multiple feature points failed to achieve adequate stabilization, which leads to a jumpy eye region image and a noisy estimation result. To overcome the drifting problem, a correlation between the current nose image and the original nose image is computed at each frame. If the correlation falls below 97%, a local search will be performed to reset the tracking.

Once the nose can be precisely localized in each frame, the eye region image can be precisely localized as well because the relative positions of the nose and eyes are fixed. The user will be asked to specify the nose position in the first frame, where the relation between the nose center and the iris center (given by the circle detector) can be computed and used later.

The overall result is a stabilized eye region image, which can be used for template matching.

4.3 Gaze Estimation

For comparison, two methods for gaze estimation are implemented here. The first one is the proposed template matching method, where the corresponding template should be able to lock onto the iris boundary. The second method is to extract the iris boundary points and fit an ellipse onto them. The center of the ellipse, which corresponds to the iris center, can be used for gaze estimation.

4.3.1 Template Matching

The final step for the proposed method is template matching. Each precomputed template will be pixelwise multiplied with the eye image and the score will be the summation. The template giving the highest score will be selected and its parameters used as the estimated gaze direction, which are angles for horizontal and vertical rotation of the eyeball.

To improve the speed, only the templates with angles around the previous estimated angles are used in the current estimation step, if the two consecutive eye images do not change too much. For example, suppose the previous estimated angles are 10 degrees horizontal and 6 degrees vertical, then the current estimation will only consider 5 to 15 degrees horizontal and 1 to 11 degrees vertical. This local search obviates multiplication on all templates and thus speeds up the program. In the case of fast eyeball movement, the difference between the previous eye image and the current eye image is calculated, and should this difference be larger than a threshold, all templates will be searched to determine the best estimated parameters.

4.3.2 Ellipse Fitting

To perform ellipse fitting, at least five points are required since an arbitrary ellipse has five degrees of freedom. To get iris boundary points, we first convolve the center template with the eye image and find the point with strongest response. This point gives us an approximate location of the iris center. The reason is that even when the eyeball is rotated, the iris still looks like a circle, whose center can be located by the center template with some error. Then starting from this point and going outwards, we sample the eye image pixels along the radial direction in different angles. To find the iris boundary, simply find the pixel with the largest gradient value along this line. The sampled pixels and the resulting iris boundary points are shown in Figure 7.

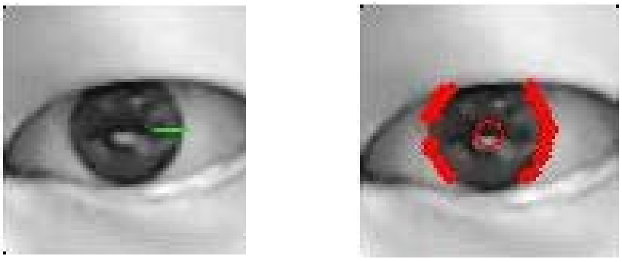


Figure 7: The left image shows the sampled pixels along one radial direction (green line). The right image shows the detected iris boundary points (red dots).

After the iris boundary has been located, an ellipse can be fitted onto it. The general formula for an ellipse used here is

$$\mathrm{Ax}^2 + 2\mathrm{Bxy} + \mathrm{Cy}^2 + 2\mathrm{Dx} + 2\mathrm{Fy} + \mathrm{G} = 0$$

and the ellipse center can be computed as

$$\mathrm{x}_{\mathrm{center}} = \frac{CD - BF}{B^2 - AC}$$

$$\mathrm{y}_{\mathrm{center}} = \frac{AF - BD}{B^2 - AC}$$

To find the parameters A,B,C,D,F,G, direct least square fitting with RANSAC is used. The final estimated ellipse center is a weighted average of the ellipse center s in each iteration in RANSAC, where iteration with fewer outlier has higher weight. The estimated iris centers are shown as a red circle in Figure 8.

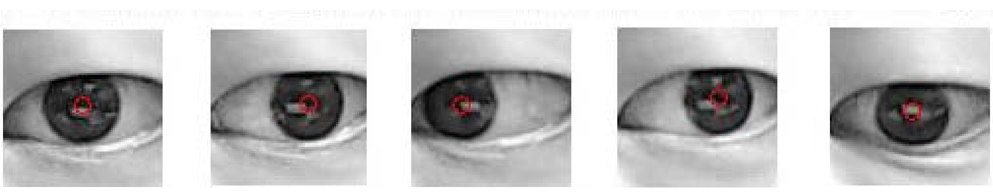


Figure 8: The red circle represents the estimated iris center.

One noticeable advantage of using ellipse fitting is that it does not require eye images to be aligned in each frame and a translation is allowed. As a result, the drift problem caused by the tracking error can actually be eliminated. The idea is as follows:

Define the following notations:

$(\mathrm{x}_0\ y_0) = initial\ nose\ position$

$(\mathrm{x}_\mathrm{t}\ y_t) = nose\ position\ change\ due\ to\ head\ translation$

$(\mathrm{x}_\mathrm{d}\ y_d) = nose\ position\ drift\ due\ to\ tracking\ error$

(nose position with respect to global coordinate in the current frame)

$(\mathrm{u}_0\ v_0) = initial\ iris\ center\ position$

$(\mathrm{u}_\mathrm{t}\ v_t) = iris\ center\ position\ change\ due\ to\ head\ translation$

$(\mathrm{u}_\mathrm{d}\ v_d) = iris\ center\ position\ drift\ due\ to\ tracking\ error$

$(\mathrm{u}_\mathrm{g}\ v_g) = iris\ center\ position\ change\ due\ to\ gaze\ change$

(iris center position with respect to eye region image, as described in section 4)

The observed nose position at the current frame is $(\mathrm{x}_0 + \mathrm{x}_\mathrm{t} + \mathrm{x}_\mathrm{d}, \mathrm{y}_0 + \mathrm{y}_\mathrm{t} + \mathrm{y}_\mathrm{d})$ and the observed iris center position is $(\mathrm{u}_0 + \mathrm{u}_\mathrm{t} + \mathrm{u}_\mathrm{g} + \mathrm{u}_\mathrm{d}, \mathrm{v}_0 + \mathrm{v}_\mathrm{t} + \mathrm{v}_\mathrm{g} + \mathrm{v}_\mathrm{d})$. Notice that if the nose position drifts to the right by 5 pixels, for example, then the iris center position will drift to the left by 5 pixels because we are assuming the relative position between the nose and the eye is fixed (namely no head movement). This implies $(\mathrm{u}_\mathrm{d}\ v_d) = -(\mathrm{x}_\mathrm{d}\ y_d)$ and we can cancel out this drift effect by adding up the observed nose and iris center positions as

$$(\mathrm{x}_0 + \mathrm{u}_0) + (\mathrm{x}_\mathrm{t} + \mathrm{u}_\mathrm{t}) + \mathrm{u}_\mathrm{g}$$

$$(\mathrm{y}_0 + \mathrm{v}_0) + (\mathrm{y}_\mathrm{t} + \mathrm{v}_\mathrm{t}) + \mathrm{v}_\mathrm{g}$$

where the terms in the first parenthesis are the initial positions of the nose and the eye in first frame; the second parenthesis terms are from the head movement, which is assumed to be 0 in this project; the third parenthesis terms are from gaze change, which can be computed as current observed position minus initial position.

Chapter 5: Experiment Result

5.1 3D Avatar

The original avatar program combines the lip movement due to speech and expression linearly without taking into account the correlation between them, which results in a bad visual performance. The proposed method defines the lip shape for each phoneme under each expression separately, where those lip shapes are currently manually tuned. The outcome is visually more natural. The images in Figure 9 show the comparison.

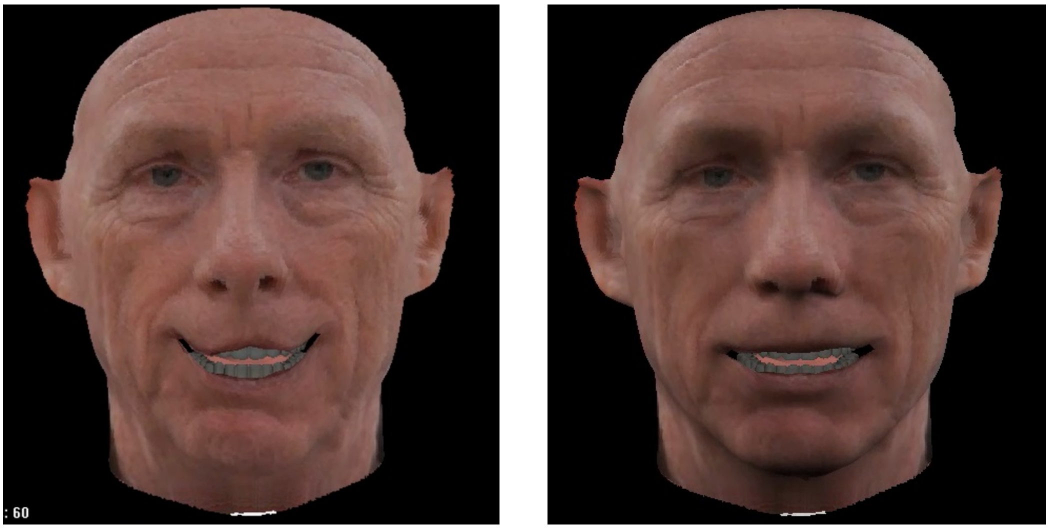


Figure 9: The left picture is the original result. We can see that the lip shape, especially the lip corner, looks unnatural. The right picture shows the improved result, where the lip shape looks more natural. Both pictures show the person speaking phoneme ‘e’ in a happy emotion.

Currently three emotions are included in the improved model: neutral, happy and sad. Other emotions, like angry, surprised and afraid, are not included since they are hardly used during patient-physician communication.

To further improve the naturalness, the shoulder and hair have been added to the avatar. When the head rotates during speech, the shoulder will rotate accordingly to increase naturalness. Figure 10 shows some final avatars from the current system with background added.

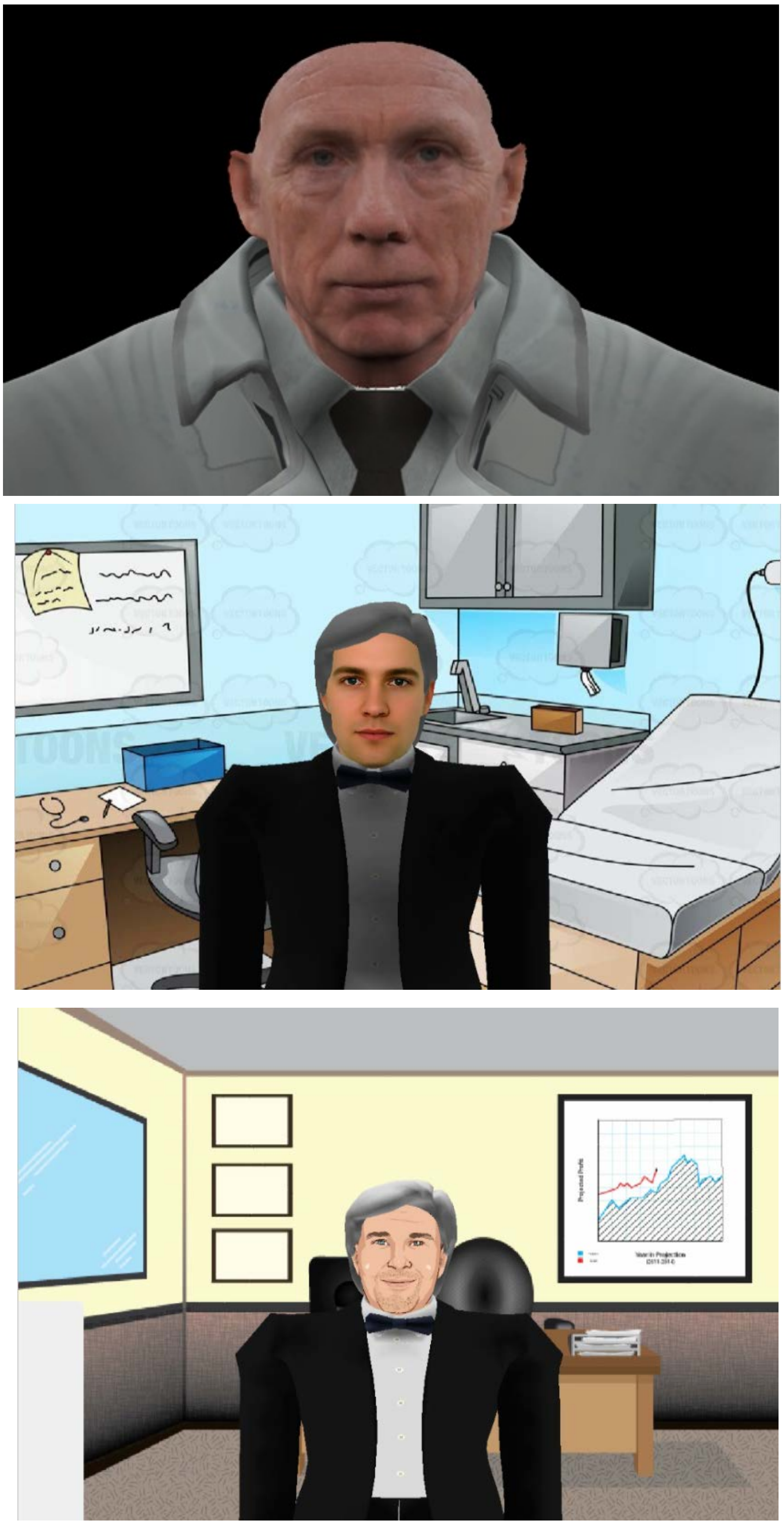


Figure 10: Three different avatars constructed from different face images. The avatar on top is built from a real face image. The other avatars are built from cartoon face images with hair added.

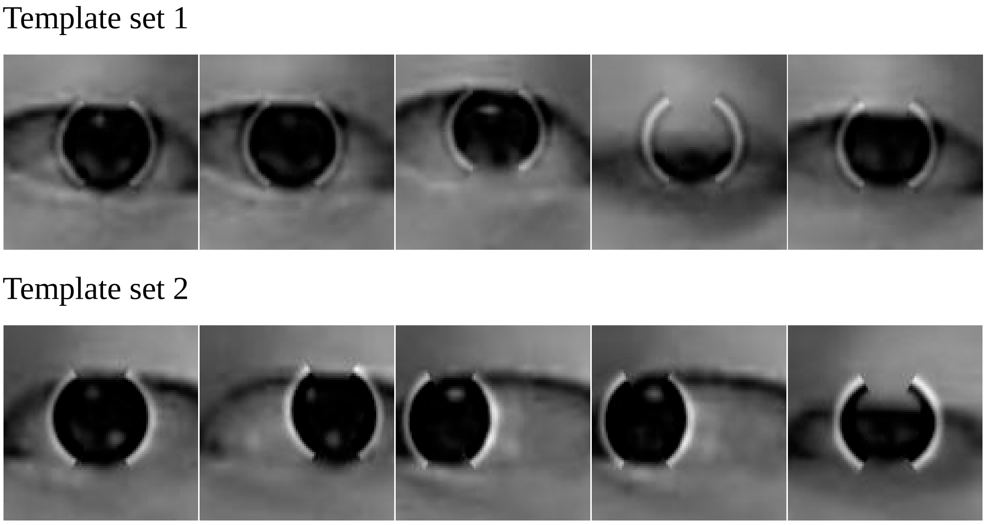
5.2 Eye Gaze Direction Estimation

Since no calibration process is included in this project, it will be hard to evaluate how well the proposed method estimates which point on the screen the user is looking at. Instead, the evaluation will mainly focus on the consistency of the estimated gaze direction within each trial and the ability of the algorithm to follow the eye movement.

Two sets of images are used during testing. The first set of images are captured using Logitech C920 webcam with a resolution of 1920-by-1080. The test subject was about 40cm in front of the camera. The size of the eye region is around 85-by-120 pixels. The second set of images are captured using a low resolution camera with eye region of 29-by-43. During the test, the test subject was asked to try not to rotate or move his head.

- Low resolution image

  During the test of the low resolution image, the images are enlarged by a factor of 3; otherwise, the iris will be too small and the estimation turns out to be dominated by quantization error. Some templates overlaid on the corresponding eye images are show in Figure 11. Notice that the templates are not very sensitive to occlusion caused by eyelid or eyelash.



Template set 1

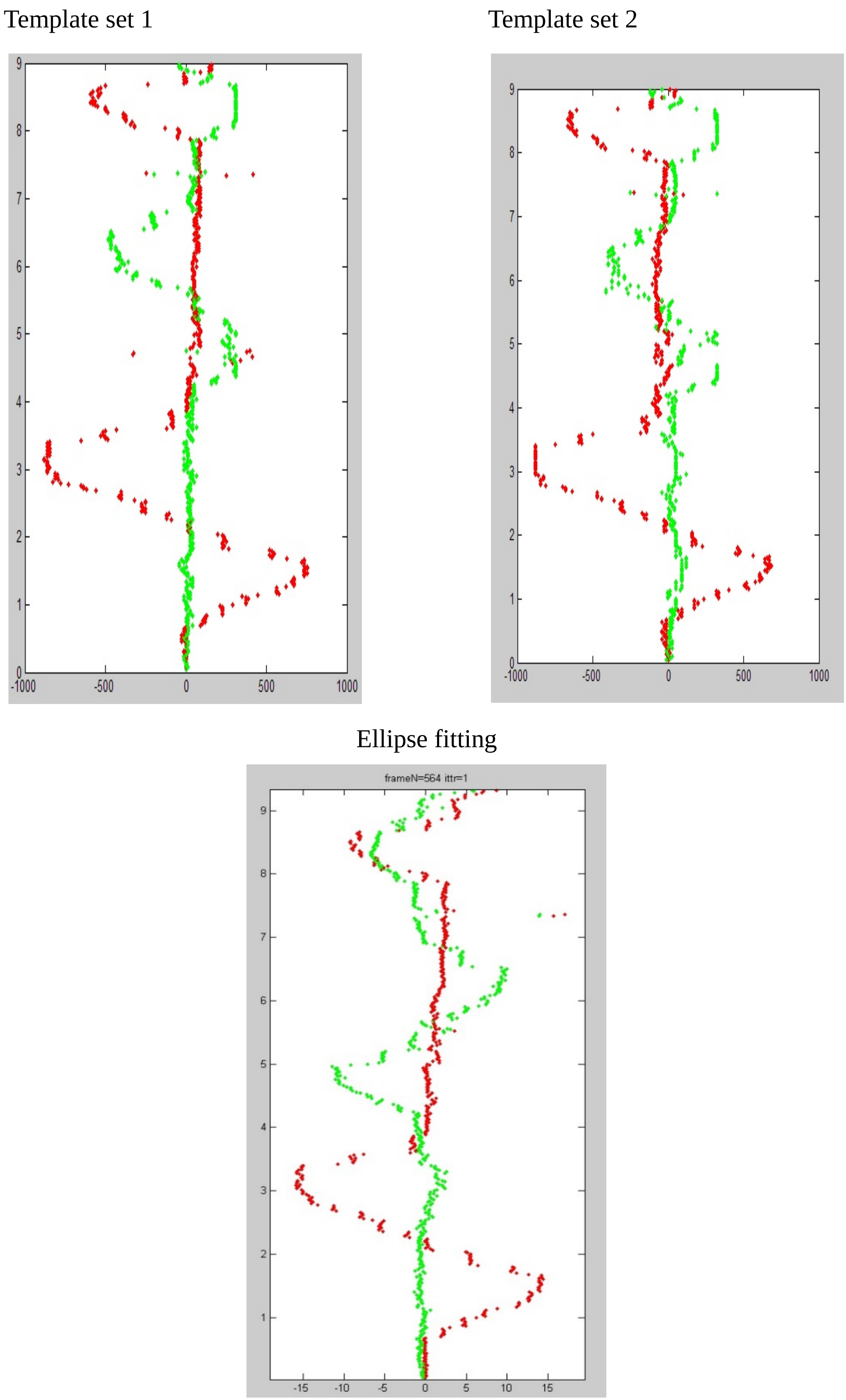
Template set 2

Figure 11: Templates overlaid on eye images. The top row shows the first template set. The bottom row shows the second template set.

In terms of consistency within this trial, Figure 12 shows the horizontal angle (red dots) and vertical angle (green dots) along the time axis (going from bottom to top). The horizontal axis has been enlarged by multiplying by a constant for demonstration purpose.

During the test, the test subject first varied the horizontal gaze angle while keeping the vertical angle constant and then varied the vertical angle while keeping the horizontal angle constant. Finally, the test subject looked diagonally to the upper left corner. As we can see, the program was able to follow the test subject's gaze direction. The horizontal angle appears to be much more stable than the vertical angle, largely due to the fact that there are many more vertical iris boundaries, which contribute to horizontal angle estimation, than horizontal edges (mostly occluded by eyelid). In addition, it turns out that template set 1 gave better results in terms of vertical angle since it was less noisy than template set 2.

The ellipse fitting result is less accurate for large eye rotation. The reason is that when the eyeball rotates away from the center, part of the iris boundary will become blurred. As a result, the iris center estimate will become inaccurate for large rotations. The ellipse fitting approach result is also included in Figure 12 for comparison.


Template set 1

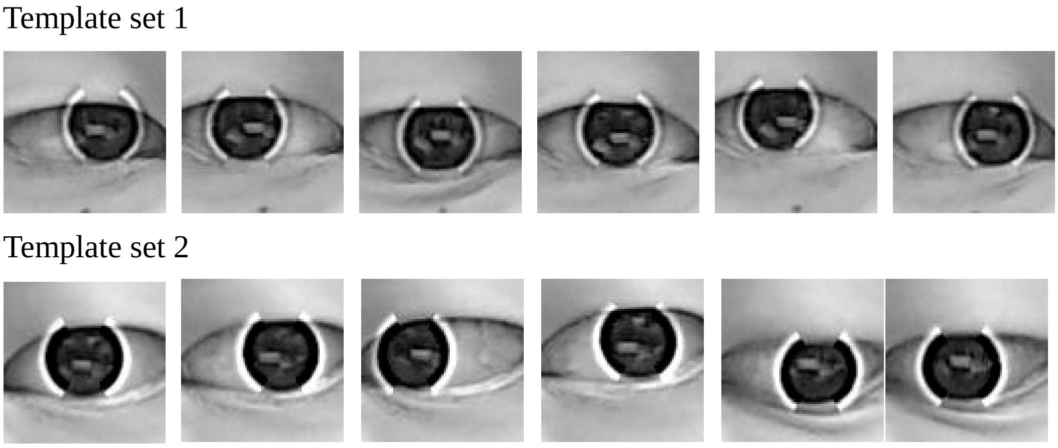
Template set 2

Ellipse fitting

Figure 12: Horizontal (red dot) and vertical (green dot) movements of the iris.

- High resolution image

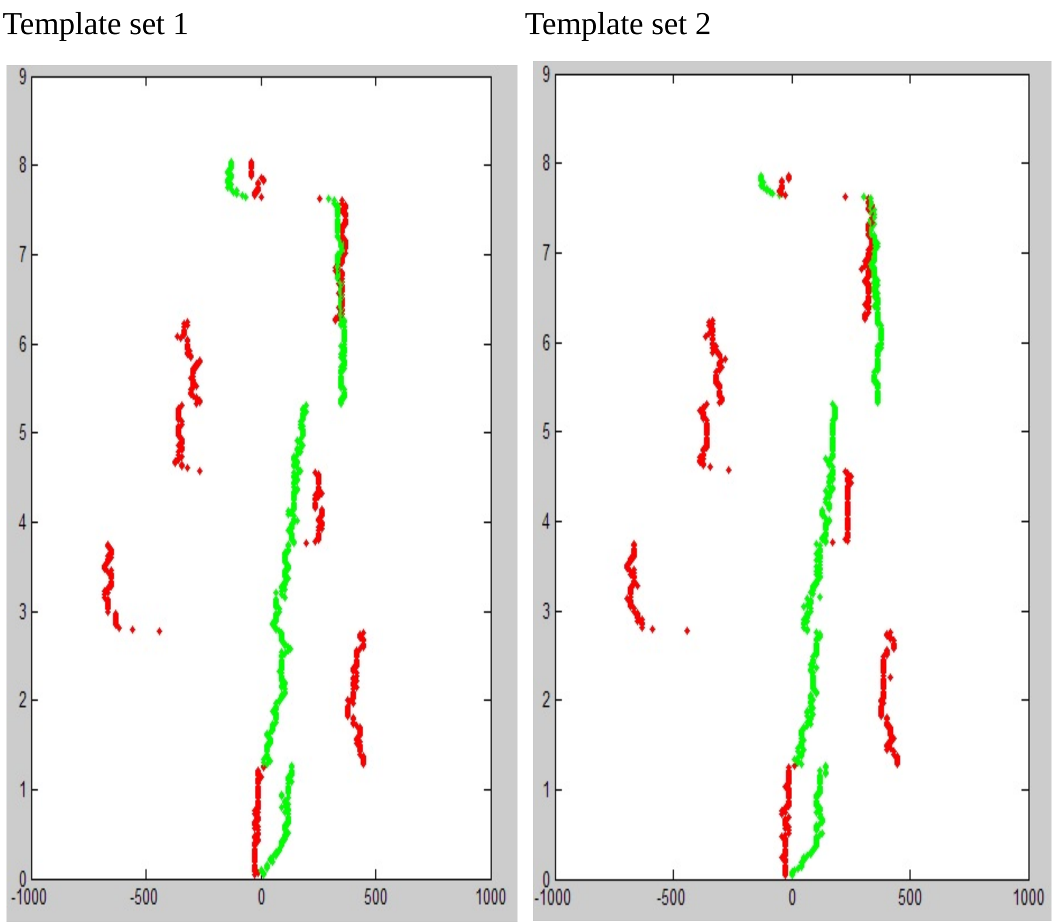
The high resolution images are captured by a much better webcam, which has brightness correction. So the overall result is much better than the lower resolution image case. Figure 13 shows some examples of overlaid images.


Template set 1

Template set 2

Figure 13: Templates overlaid on eye images. The top row shows the first template set. The bottom row shows the second template set.

In the first trial, the test subject fixed the gaze direction at one specific point at a time, so the horizontal and vertical angles should be constants during this period. The results are shown in Figure 14.



Template set 1

Template set 2

Figure 14: Horizontal angle (red dots). Vertical angle (green dots).

In Figure 14, each line segment represents one fixation, which means the gaze is fixed at a specific direction. As a result, most line segments are approximately straight lines. Ideally they should be vertically straight. However, without considering the drift error, those straight lines are tilted to the right, which means the drifting error direction of the tracking process is to the left.

The result of ellipse fitting method on the same dataset is shown in Figure 15. We can see that the lines are almost straight in the vertical direction compared to the template matching result, thanks to the drift cancelation mechanism.

Ellipse fitting result

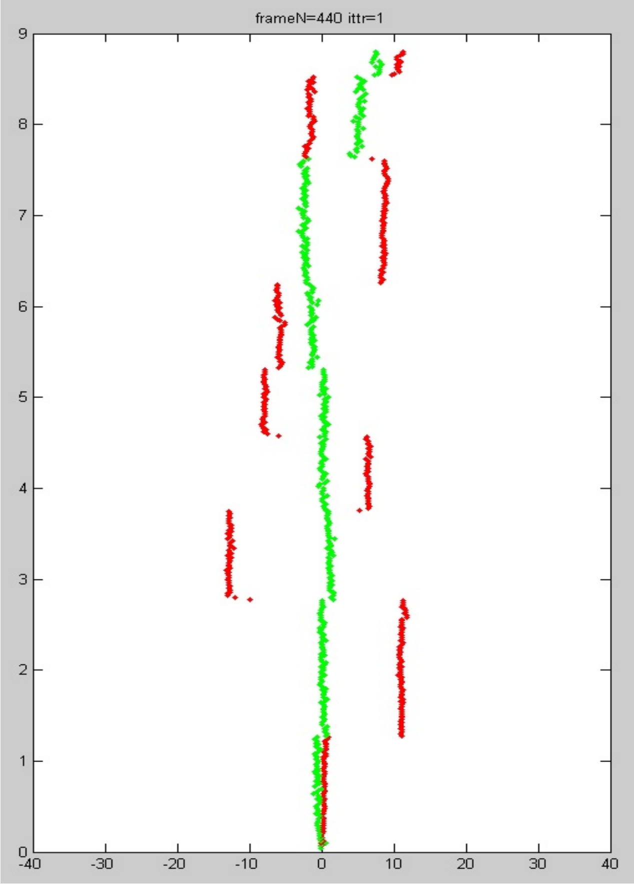


Figure 15: Compared to the result without drift cancelation, this time the drift effect has been greatly suppressed, evinced in the greater verticality of the lines. Note that this drift cancelation mechanism can only deal with drifting error, whereas head drifting caused by user movement cannot be canceled.

Chapter 6: Discussion

6.1 3D Avatar

In our system, an audio-visual emotive 3D avatar is proposed to deliver medical messages to patients to help them understand their medical results. The use of an avatar enhanced the portal messages by delivering additional emotive information through the avatar's face expression and prosody of speech. Along with the semantic information, the enhanced portal messages should be more easily understandable by the patients, especially those with low medical and numeric literacy.

To summarize, the avatar enhanced PHR system has the following features:

- Additional information can be delivered by the avatar portal, such as expression and prosody of speech, whereas traditional patient portals only have text or graph information. This extra information has great potential to help patients better understand the medical information.
- The system is highly flexible. Almost any frontal image of a person's face with neutral expression can be converted into a 3D avatar in only a second, which means we can customize the system for each patient and let them build their own avatar. This feature gives the user a great deal of variety as well as offering the patients more choices and making it easier for them to accept the interface.
- The construction of the avatar is fully automatic, which means the user only needs to give a 2D frontal image of a person. There is no need to specify where the head is or the location of some important feature points like eye center or nose tip. As a result, this approach is very convenient, especially for users without any knowledge of avatar construction.

Despite the many advantages discussed above, there are many aspects that need to be improved. Currently the avatar's facial expressions are rule based, which means the expressions and lip movements during speech are controlled by pre-defined templates. However, different people might have different facial expressions and lip movements so that the templates built for one person may not suit another. One solution is to develop a method to learn a set of templates for each new person's avatar by tracking the facial feature points when the person is talking. Then, construct a set of templates for this person. If successful, the resulting avatar should be more realistic.

Another possible direction for improvement is patient-avatar interaction. Currently, the avatar is only a narrator without any kind of interaction. However, patients might have questions when listening to their medical results. One approach to improve this system is to add in speech and emotion recognition which monitors the patient reaction, as Lisetti et al. did in [5]. For example, if the patient looks confused or asks a question, the system will pause and give more detailed background information or instructions. This will make the system much more user friendly and as a result, greatly increase the effectiveness of the PHR systems.

6.2 Gaze Estimation

We have proposed an appearance template-matching based eye gaze estimation method, which can be realized using a single RGB web camera. The experiment results show that this method is able to track the user's eye movement even for low-resolution images with poor lighting.

However, this method has a major drawback. The algorithm depends heavily on the alignment of each eye image. The current evaluation was done for a fixed head without rotation or translation, and the eye images are stabilized by tracking the nose. But in practice the user may want to move his head and the tracking can be inaccurate due to lighting change and facial landmark shape change, which will negatively impact the gaze estimation result.

This drawback can be solved if a consistent eyeball center can be estimated. One possible solution is to use the Active Appearance Model to get a consistent eyeball center when the head translates or rotates, as studied by Chen and Qiang [11]. Following their approach, a better estimation result should be expected.

For the ellipse fitting method, the most important factor affecting the accuracy is iris boundary extraction. It turns out that the ellipse fitting is much more accurate when the eye is looking at the center and less accurate when the eye is rotated away, which blurs the iris boundary. Possible ways to solve this problem include using a more robust ellipse fitting algorithm which deals with noise better, or using a better camera to get a sharper iris boundary.

The evaluations performed in this thesis are qualitative, showing that the algorithm is able to track the eye movement of the user. To truly evaluate the usefulness of this method, a quantitative test still needs to be done. Specifically, the user can look at several fixed points on the screen during the calibration phase and the resulting output from the proposed algorithm can be recorded. During the testing, the target gaze point can be estimated by comparing the current output to the recorded output from calibration and just picking the most likely point.

Chapter 7: Conclusion

Computer vision has been an active research area for many years. From the 1960s when the concept of computer vision was born until today, great advancement has been achieved. Recently, the advancement of powerful hardware systems has enabled the technologies in CV to be successfully applied to many practical uses. Despite the limitation of CV technologies, where in most cases the environment is required to meet certain requirements, CV has many advantages that human beings do not possess, such as high reliability, high consistency and high speed. Those advantages make CV an ideal tool for automation and labor intensive applications, which could free us from a large amount of tedious work.

In this thesis, two applications of computer vision are studied. The first application, which is to enhance PHRs with 3D avatars, demonstrates a promising approach to increase the usability of PHRs and promote self-care. With the help of a 3D avatar, the patients are expected to have a better understanding of their test results and the portal messages, by means of interpreting the emotive information delivered by the avatar. This improvement greatly enhances the remote communication between the patients and their physicians, which could reduce the frequency of patients returning to the hospital. As a result, the physicians are able to focus on more important and urgent cases without being overwhelmed by answering questions, which greatly increases the efficiency of the medical system.

The second application, eye gaze direction estimation, is a new template matching based method of gaze estimation with low image resolution and contrast. Although this method cannot achieve high accuracy, it greatly reduces cost by requiring only a single RGB camera. The proposed method is suitable for situations where only a rough gaze direction is sufficient to make decisions. For example, the experimenters only need to find which one of the nine pictures on the screen the tester is looking at, or maybe the experimenters only need to know whether the tester is looking at the center, left or right. Another possible application is driver fatigue detection, where only a rough gaze direction is sufficient to decide whether the driver is looking forward or looking down.

The work done in this thesis is only preliminary. There are many possible directions for further study. For the 3D avatar, the appearance can be further improved to increase authenticity. Or the avatar system can be made interactive, which means the avatar is able to interpret what the patient is asking and monitor the patient’s emotion and take appropriate actions. This improvement should be able to boost the flexibility and usability of the avatar-enhanced PHR systems. For the gaze estimation, the next step should be to carry out quantitative evaluation while enabling the system to take head translation and rotation into account, making the system more reliable and robust.

References

[1] P. C. Tang, J. S. Ash, D.W. Bates, J. M. Overhage, and D. Z. Sands, "Personal health records: Definitions, benefits, and strategies for overcoming barriers to adoption," *Journal of the American Medical Informatics Association*, vol. 13, no. 2, pp. 121–126, March/April 2006.

[2] P. Kuhn, “Patient portals,” *Health Management Technology*, vol. 29, pp. 43-44, Oct. 2008.

[3] Institute of Medicine (2004). Health Literacy: A Prescription to End Confusion. [Online]. Available: http://www.iom.edu/Reports/2004/Health-Literacy-A-Prescription-to-End-Confusion.aspx

[4] H. Mark, J. Leff, and G. Williams, "Avatar therapy: an audio-visual dialogue system for treating auditory hallucinations," *Interspeech*, 2013.

[5] C. Lisetti, U. Yasavur, C. de Leon, R. Amini, U. Visser, N. Rishe, "Building an on-demand avatar-based health intervention for behavior change," *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference*, pp. 175-180, 2012.

[6] T. W. Bickmore, L. M. Pfeifer, and B. W. Jack, "Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1265-1274, 2009.

[7] D. Li, D. Winfield, and D. J. Parkhurst, "Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, p. 79, June 2005.

[8] J. Wang, E. Sung, and R. Venkateswarlu, "Eye gaze estimation from a single image of one eye," *Computer Vision*, vol. 1, pp. 136-143, 2003.

[9] S. Cadavid, M. H. Mahoor, D. S. Messinger, and J. F. Cohn, "Automated classification of gaze direction using spectral regression and support vector

machine," *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1-6, 2009.

[10] H. Tang, Y. Hu, Y. Fu, M. Hasegawa-Johnson, and T. S. Huang, "Real-time conversion from a single 2D face image to a 3D text-driven emotive audio-visual avatar," *IEEE International Conference on Multimedia and Expo,* pp. 1205-1208, June/April 2008.

[11] J. Chen and Q. Ji, "3D gaze estimation with a single camera without IR illumination," *19th International Conference on Pattern Recognition*, pp. 1-4, Dec. 2008.