PetCaretaker: Design to Detect Sleeping Behavior of Serious Mentally Limited People

Lang Bai
Eindhoven University of Technology
Eindhoven, Netherlands
lilbailang@gmail.com

ABSTRACT
There has few product or research on designing for mental limited people. The caretaking company have no enough caretakers to keep an eye on the mentally limited clients, who are suffering circadian rhythm disorder. We report on the design of a system to detect the sleeping behavior of mentally limited people in the day time. The system utilizes inexpensive 2D camera trackers and a pet-appearance to get accompany with clients. The system multi-stage wakes up the clients when a long-time sleeping is detected. Our study explores taking EAR, head orientation as inputs, SVM as an approach to managing complex situations in real life.

Author Keywords
Circadian cycle; eye aspect ratio; SVM; mentally limited.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

BACKGROUND
Siza is a company running for providing care of hundreds of disabled people in serval resident pots in the Netherlands. In the summer of 2015, Siza collected the 24-hour heart rate data of three group of clients. These clients are assigned in different by the level of cognitive disability. By analyzing the activities in day and night, the Serious Mentally Limited clients (from here called SML clients) and Moderate Serious Mentally Limited clients (from here called MSML clients) are with difficulty in understanding the routine activities. SML and MSML clients have no idea on what to do according to a regular timetable. It mainly shows from their above-normal heart rate at night and low heart rate at daytime. Therefore, the clients are considered to be circadian rhythm disorder. Since the clients are seriously mentally limited, even pressing a button will build up a barrier. It's difficult to use a system with complicated interaction to educate them to be aware of the appropriate routine. The primary step in this problem, it's to detect their sleeping behavior at daytime and awake them. Due to the limited number and full-loaded work of caretakers, it's impossible to provide one-on-one or one-on-two supervise. Thus, an autonomous device of detecting and waking would be in demand.

There are multiple methods in drowsiness detection, however, in our problem setting, there are constraints in choosing a method. Additional hardware should be with cost control, EEG based methods are eliminated on this rule. From the observation, clients are sitting in the wheelchairs all day. They are mostly quiet, with no movement in large amplitude. The quickest and easiest way to detect motion is to compute the difference between the initial frame and the subsequent frame. However, due to the complexity of the real-life situation, namely, some people might pass or stand behind the clients. To recognize the face, especially the eye, could boost the detection part in this scenario. According to Roenneberg[1], an alarm clock is hazardous to health. An alternative way is to set multiple wake-up stages, with gradually increasing intensity level.

RELATED WORK

1. Facial landmark
Facial landmark is a computer-based function for automatically dealing with detecting distinctive features in human being faces. According to the recent surveys, facial landmark detection has long been impeded by the problems of occlusion and pose variation [2]. What is noticeable is that most of these works assume a densely connected model which can support the need for matching algorithms. For example, based on a mixture of trees with a shared pool of parts, every facial landmark can be used global mixtures to capture topological changes due to viewpoint [3]. Another team presents an interactive model that allows for improving on the appearance fitting step by introducing a Viterbi optimization process which can operate along the facial contours [4]. Other notable recent works exploring how they use appropriate priors exploiting the structure of image data to help with efficient feature selection[5].

2. head pose estimation
Estimating the head pose of a person is a common human ability that presents a challenge for the computer system. With the recent increasing technologies, a few notable

works have shown the usefulness for solving this problem. Some use a unique cue [6], which use random regression forests for giving their capability to handle large training datasets. Another method can be combined with 2D image data, such as [7]. A regularized maximum likelihood deformable model fitting (DMF) algorithm is developed, with special emphasis on handling the noisy input depth data.

3. External influences

There are several elements that can influence the robustness of the algorithm. The light environment can be an assignable part of that. The light field is considered as the set of features on which to base recognition, analogously to how the pixel intensities are used in appearance-based face and object recognition [8]. Shelters that cover the face can also influence the detection a lot. People from certain religious groups and occupations occlude faces where only their particular regions are visible for biometric identification [9].

APPROACH DESIGN

We present an approach that can detect the sleeping and tenderly wake up the clients. It consists of eye recognition and multi-stage waking up. Our study draws upon research in activity room environments. It used a co-design approach involving people with mental limitation, their caregivers, and students from TU/e. We adopted this approach to get meaningful insights regarding the participants' and other stakeholders' sleeping behavior detection, their reaction to the alarm.

Observation and interviews setting up

We conducted semi-structured interviews and observations in two care centers in Arnhem. To get insights about the everyday lives of people with mental limitation, we directly observe the clients' behavior and also interview the caretakers.

There are plenty of mental limitation clients while we only focus on the SML and MSML clients because there are already on-going projects aiming at changing other clients' behavior due to their better ability to do basic interaction. While SML and MSML are the clients are in a more urgent need to have behavior changing. There are in total 12 clients being observed by us, 5 females and 7 males, aging from 19 to 62. we conducted interviews with 2 professional managers and 2 caretakers from care centers. All of them visited and participated in the group-based sessions or the training sessions at home on a regular basis.

After the observation and interviews, the system tailored to their behavior and environments was implemented. Another round of test had been conducted in the care center on three clients. We want to see the performance of the system and find out the complexity of the environments that we haven't concerned about. In the meanwhile, video data are recorded with permission for further development.

Data collection

In terms of data collection, we conducted semi-structured interviews with the sample of 12 SML and MSL clients and 4 of their caretakers. Interview questions concentrated on: their illness condition; daily routines; and the interaction between caretakers and clients; the difficulties of caretaker to conduct; the effect of previous research. Participants were encouraged to elaborate freely upon these topics. All interviews were conducted face-to-face by researchers. The interviews lasted 30 to 60 minutes and were transcribed. Researchers conducted a direct observation of participants' interactions and took field notes while observing.

To increase the usability of data, we took the two methods: (1) talking about the similar topic to all the caretakers, (2) note down the specific infrastructure only when we observe the real objects, (3) test our system with the accompanies of the manager of caretakers, communicating about the satisfactory and the aspects he thought might be important.

To examine the sleeping behavior, it's time-consuming to record the video of clients. Also considering privacy, we take our own video as inputs. We mimicked the sitting posture and blinking frequency of the clients to provide the parameters and train data for our system. The parameter is fined tuned after test on three clients.

Qualitative Data Analysis

There is two stage of data analysis. The first stage is to analyze the qualitative data from observation and interviews, the second stage is to analyze the data for system parameter design. In this section, we emphasize the qualitative data analysis since it provides the guideline of algorithm design. Quantitative data analysis will be mentioned in the system design section. In this case, thematic analysis is applied. After generating the initial codes, researchers searched for the themes and reviews them, the theme with refine specifics and clear definition is: user privacy; the cost of the system; mobility of the system; benefits and drawbacks of using the system; humanization of the system; learning curve of caretakers; workload of caretakers.

In the meanwhile, the biological clock of the SML and MSML clients are shown in Figure 1. The circadian rhythm disorder happened due to the sleeping behavior when they were in the activity room. To prevent the sleeping behavior in the daytime could lead to more sleep at night.

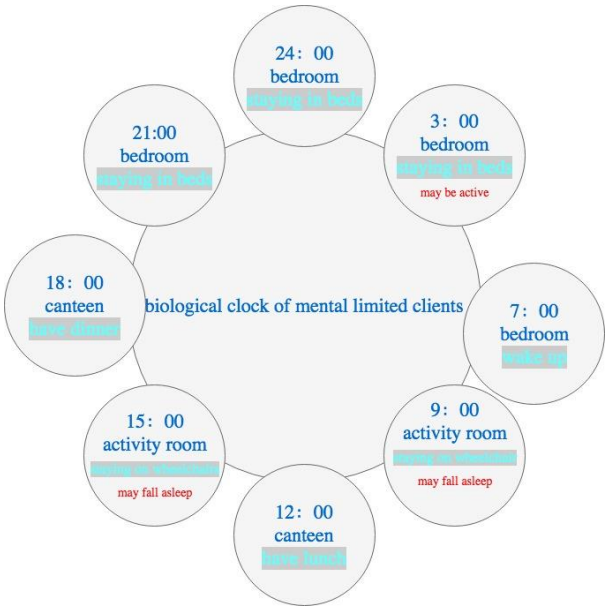


Figure 1. current biological clocks of SML and MSML clients

Detection algorithm

Based on the result of qualitative data analysis, we take the activity room as the environment of use case. Most of the time, clients are sitting on the wheelchair and listening to the soft music. With 12 clients in a room, the sound-based method is sensitive to the noise in the soundtrack, which is not appropriate to apply in this scenario. Another decision should be made on the algorithm chosen, in other words, how to detect the sleeping behavior from the camera facing to the clients? Concerned with the speed and computation cost, the initial choice was to compute the absolute difference between the previous frame and current frame. As caretakers narrated, these clients keep quiet and with only small movement, which brought two constraints when using the absolute difference of frames: (1) people walk behind the client from time to time, and the background may change because clients could be possibly moved position. The approach to predicting the background by computing the mean of previous frames had to be excluded due to the change of backgrounds. (2) body recognition can ignore the effect of background changing, however, clients' body movements are very small even when they're awake, instead of using body movement as the metrics for sleeping behavior, eye recognition is not only robust but also follow the intuition of human decision process.

Eye aspect ratio [10] is an efficient tool to determine whether the eye is closed or not. The eye aspect ratio involves a simple calculation based on the ratio distance between facial landmarks of eyes. In this method, each eye is represented by 6 points, starting at the left corner of the eye and going clockwise around the remainder of the eye region (seen in Figure 2.)

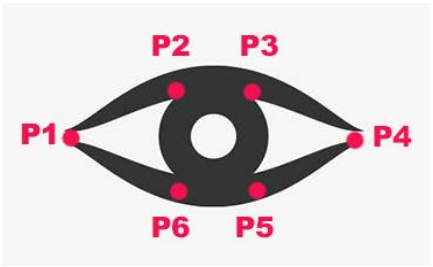


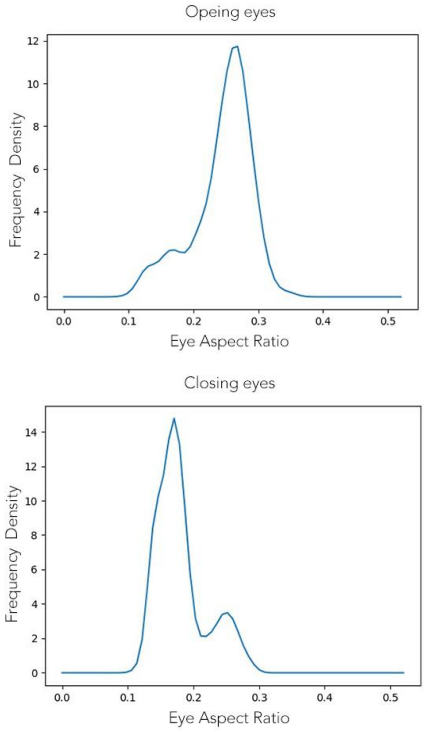
Figure 2. The 6 facial landmark points associated with the eye

The EAR can be regarded as the relation between the width and the height of these points. The equation is presented as follows:

$$\mathrm{EAR} = \frac{||\mathrm{P2}-\mathrm{P6}||+||\mathrm{P3}-\mathrm{P5}||}{2||\mathrm{P1}-\mathrm{P4}||}, \quad (1)$$

Where p1, …, p6 are 2D facial landmark locations. The numerator of this equation computes the distance between the vertical eye landmarks (height), and the denominator computes the distance between horizontal eye landmarks (width).

We use our own sleeping and awake video as inputs. Each video has around 10 seconds, 30fps, with a resolution 640 × 480 pixels. We have three types of behaviors to examine, closing eyes for the whole time, opening eyes for the whole time, blinking for the whole time. The landmark recognition is implemented through OpenCV and dlib library in Python. The EAR distribution histogram can be seen in Figure 3.



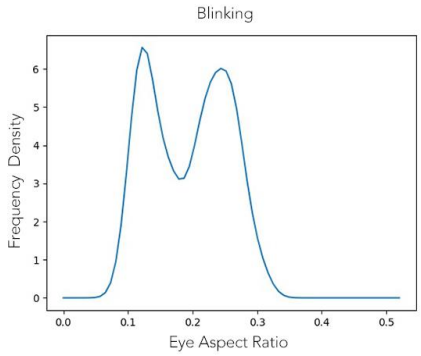


Figure 3: Examples of EAR distribution histogram of opening eyes, closing eyes, and blinking.

There is an interesting threshold value from the examples in Figure 3, which is supposed be between 0.18-0.22.

However, from the result of the first-round test, the drawback of thresholding has been exposed. In the real-life application, the clients are in a much more complicated situation than the training video, for example, with or without glasses, the light condition, head poses direction, etc. To build a more robust detector, the following approaches have been taken: (1) use video with different light conditions; (2) use video with and without glasses; (3) estimate 3D head pose direction. The examples of new training data are shown in Figure 4.

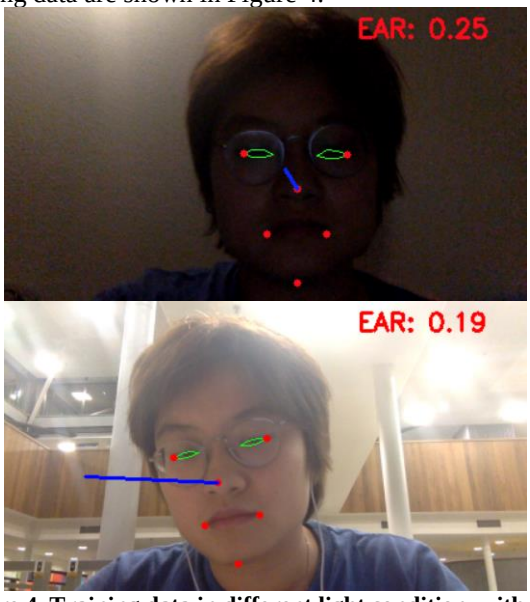


Figure 4. Training data in different light condition, with head pose estimation.

After the field testing, new dataset is annotated with EAR, roll angle, yaw angle, pitch angle, and class of sleeping or being awake, This dataset contains light setting in both bright and dark conditions, subjects with and without glasses, different head orientations in up, down, left, right direction, with maximum 45 degrees in each direction.

Head orientation estimation

The implementation of head orientation estimation follows the solution of the Perspective-n-Point problem in OpenCV. The aim of this problem is to find the pose of an object from a 2D image, given the camera intrinsic parameter (focal length, optical center), the 3D model of n points and the corresponding 2D features in the image.

For each PnP problem, there are three coordinate systems. World coordinates (U, V, W), camera coordinates (X, Y, Z), and 2D image coordinates (x, y). The 3D points in world coordinate can be transformed into 3D points in camera coordinates through rotation R (a 3x3 matrix) and translation t (a 3x1 vector). 3D points in camera coordinates can be projected on 2D image coordinates through the known camera intrinsic parameters.

The location (X, Y, Z) of the point P in the camera coordinate system can be yielded from the following equation:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = [\boldsymbol{R} \quad | \quad \boldsymbol{t}] \begin{bmatrix} U \\ V \\ W \\ 1 \end{bmatrix}, \tag{2}$$

Which can be also written as:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} r_{00} & r_{01} & r_{02} & t_x \\ r_{10} & r_{11} & r_{12} & t_y \\ r_{20} & r_{21} & r_{22} & t_z \end{bmatrix} \begin{bmatrix} U \\ V \\ W \\ 1 \end{bmatrix}, \tag{3}$$

For the point P in camera coordinates (X, Y, Z), the 2D image coordinates (x, y) can be conducted from the equation involving camera matrix (M), where $M = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$, where (f_x, f_y) is the focal lengths and (c_x, c_y) is the optical center. In most cases, the lens distortion is regarded not existed. And the equation is given by

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = s \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{4}$$

After combining the equation 3 and equation 4, the position in 2D image coordinates can be predicted from a given 3D point in the world (a facial landmark), if the right pose ($\boldsymbol{R}$ and $\boldsymbol{t}$) is given.

For the equations above, Direct Linear Transform [11] is used for getting the pose $\boldsymbol{R}$ and $\boldsymbol{t}$. However, the DLT algorithm does not minimize the reprojection error. Reprojection error represents the sum of squared distances between the projected 3D face points and 2D image points. Since the 2D points of face and body can be detected from dlib, an intuitive way to converge the error curve is to adjust the $\boldsymbol{R}$ and $\boldsymbol{t}$ to align the projected 3D points with detected 2D image points. Levenberg-Marquardt optimization is widely used to minimize reprojection error. In our implementation, CV_ITERATIVE in OpenCV's function solvePnP was chosen, which is based on Levenberg-Marquardt optimization.

Influence of head orientation, glasses, and light condition

The SVM classifier has four input: EAR, the angle of roll, the angle of yaw, the angle of pitch. Radial Basis Function (RBF) kernel SVM approach performs best among all the SVM methods, with an accuracy of 0.84 on the total of 7390 frames. After a deep dive into the detail of data, we found facial landmark detection algorithm we use works terribly when the head orientation is downward. Thus, we reach the accuracy of 0.94 on the remaining 5485 frames, after excluding the downward orientation data. The ROC curve of SVM with RBF kernel is shown in Figure 5. Since there is no strict rule of mounting a camera on the wheelchair, a relatively low position is recommended.

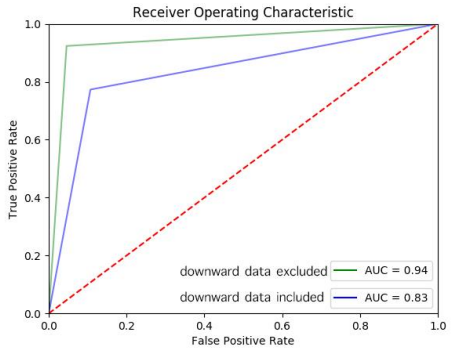


Figure 5. ROC curve of SVM with RBF kernel on including and excluding downward orientation data.

Influence of light condition is also examined in our study. Using between-group experiment design, we divide data into dim condition group and bright condition group. In the dim condition, the accuracy is 0.81 on 2959 frames, in the contrast, our approaches accuracy rises to 0.90 on data in bright condition (shown in Figure 6.). A good light condition dramatically increases the detection accuracy.

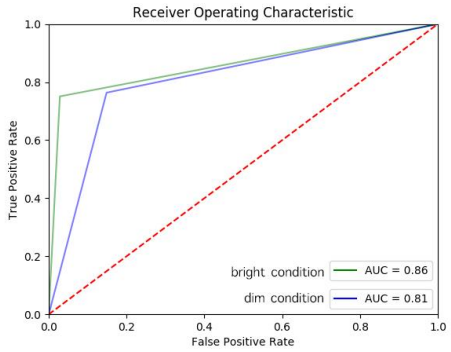


Figure 6. ROC curve of SVM with RBF kernel on different light conditions.

A pair of glass is widely considered to lead to the decrease in detection accuracy. However, our experiment on the data shows the opposite situation. When subjects are wearing glasses (3650 frames), the accuracy is 0.88, while it drops to 0.82 when subjects are not wearing glasses (3740 frames), as Figure 7 shows. However, a detailed inspection of the data provides us the insights that the exceptional conditions happen when subjects turn head downward while wearing glasses. After excluding the downward data, the accuracy on without-glasses group reaches the highest number, 0.96; while the with-glasses group also climbs to 0.94. In this case, the difference between glass conditions can be ignored.

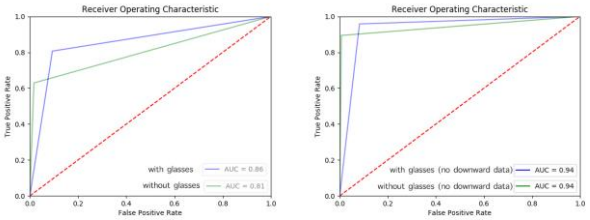


Figure 7. ROC curve of SVM with RBF kernel on different glasses conditions.

In summary, the crossing contrast experiments provide us the insights of the position of mounting the camera, a relatively low position is better than the high position. Besides, if under the cost control, low power and diffused light source would be recommended. Glasses are not a barrier to our system.

SYSTEM OVERVIEW

System feature

The section above talks about determining whether the eyes are closed or not. A sleeping behavior within a certain time is not simply only involves closing eyes, for example, people might close eyes for the first three-quarters of the time while blinking their eyes at the last quarter. To make our approach flexible, we set a window of 200 ms, which move at the step of 100 ms. Each window has a flag and highest EAR. For the EAR values in every window, we use the Gaussian Kernel Density Estimation to make histograms smooth. The average of the highest three EAR is regarded as the highest EAR of this window. Since each flag (predicted label) is assigned to one EAR value, the flag of the window is aligned with the flag of the highest three EAR value (the flag that appears at least two times). For every 10 seconds, a stop-alarm flag is on when the half of the last-2.5-second window flags are labeled as eyes-open. Otherwise, if half of the window flags are eyes-closed, the level-1 alarm is on, which is the sound of birds singing and colorful blinking lights. The volume of the alarm increases gradually if the client is still asleep. The alarm soundtrack turns into the sound of waves at level 2. When the volume increases to a certain value, the system turns off the alarm and texts to the caretaker. There is a limit on the volume since multiple clients sitting in one room, it's necessary to avoid bothering or scaring others.

System infrastructure

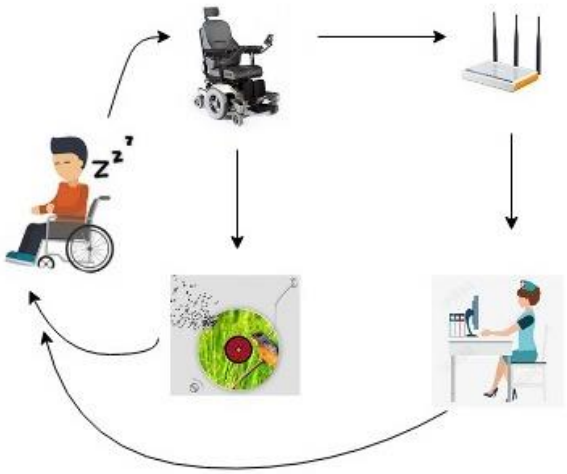


Figure 8. System structure

The system consists of detection, alarm, and communication components (see Figure 28. Camera, a diffused light source, and micro-speaker are mounted on the wheelchair. The system collects the EAR and 3D head orientation data, and store it on the cloud, no image is allowed to record. The caretakers can access the cloud to review the long-term status of clients, as well as get the notification from the system. The system is shaped like a cute pet, which imitates the pet on the wheelchair (shown in Figure 9).

Figure 9. System as a pet

CONCLUSION

In this paper, we have analyzed the approach of helping mentally limited clients build a better biologic clock. We found the trait of closing and opening eyes from the EAR values. After the field test, we found the drawback of thresholding method. Thus, the SVM approached is introduced on the data we collected from the different conditions of lighting, head orientation, and wearing glasses. We discussed the influence of different condition and yielded the position of mounting the camera, which should be relatively low. Besides, we also recommend a low power and diffused light source. Ideally, our detection approach can reach the accuracy at 0.96. However, our approach should be fine-tuned under more subjects, since different facial features would also affect the detection accuracy. Being informed that few products are designed for the serious mentally limited people, we hope to inspire work for this target group and beyond this domain.

REFERENCES

1. Roenneberg, T., Allebrandt, K. V., Merrow, M., & Vetter, C. 2012. Social jetlag and obesity. *Current Biology*, *22*, 10, 939-943.
2. Zhang, Zhanpeng, et al. "Facial landmark detection by deep multi-task learning." *European Conference on Computer Vision*. Springer, Cham, 2014.
3. Zhu, Xiangxin, and Deva Ramanan. "Face detection, pose estimation, and landmark localization in the wild." *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.
4. Le, Vuong, et al. "Interactive facial feature localization." *European Conference on Computer Vision*. Springer, Berlin, Heidelberg, 2012.
5. Kazemi, Vahid, and Sullivan Josephine. "One millisecond face alignment with an ensemble of regression trees." 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, United States, 23 June 2014 through 28 June 2014. IEEE Computer Society, 2014.
6. Fanelli, Gabriele, Juergen Gall, and Luc Van Gool. "Real time head pose estimation with random regression forests." *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011.
7. Cai, Qin, et al. "3d deformable face tracking with a commodity depth camera." *European conference on computer vision*. Springer, Berlin, Heidelberg, 2010.
8. Gross, Ralph, Iain Matthews, and Simon Baker. "Appearance-based face recognition and light-fields." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.4 (2004): 449-465.
9. Juefei-Xu, Felix, and Marios Savvides. "Subspace-based discrete transform encoded local binary patterns representations for robust periocular matching on NIST's face recognition grand challenge." *IEEE transactions on image processing* 23.8 (2014): 3490-3505.
10. Soukupová, T., & Cech, J. 2016. Real-time eye blink detection using facial landmarks. In *21st Computer Vision Winter Workshop*.
11. Sutherland, I. E. 1974. Three-dimensional data input by tablet. *Proceedings of the IEEE*, *62*, 4, 453-461.