Computer Vision for StudyBuddy: Face Detection and Eye Tracking

Maria Irene Gonzalez Martinez
Department of Electrical and Electronic Engineering at Imperial College London
Email: mig14@ic.ac.uk
Supervisor: Prof. Yiannis Demiris

***Abstract*— Procrastination can be a serious issue while studying. With an overly extended duration, it can lead to increased stress levels, a higher propensity to illness and lower academic performance [1]. As a result, we propose StudyBuddy; an interactive mobile healthcare system which can monitor study sessions via the means of a smartphone's built-in frontal camera. This is to be used to generate a real-time attentional forecast. We seek for our system to be accessible, user-friendly and non-invasive. Therefore, we focus on the estimation of visual focus of attention (VFOA), which we can estimate via computer vision techniques. We expect our attentional forecast to be useful to deliver personalized recommendations and feedback to the user. These intend to actively counter procrastination.**

I. INTRODUCTION

StudyBuddy seeks to battle student procrastination by monitoring its users' behavior while studying, and to use this to deliver personalized recommendations before, during and after a studying session. Additionally, users will be able to obtain feedback from past sessions and track their progress from historical statistics. This system is to be delivered in the form of a user-friendly and intuitive Android application.

StudyBuddy calls for the design of a component which extracts a robust, real-time, attentional estimate. This is essential to understand users' behavior while studying, and hence for its recommendation component to be able to generate individual guidance for each user. In this report, the preliminary design for such a component will be discussed.

The remainder of this document is structured as follows: section II discusses the requirements that the computer vision component fulfills within StudyBuddy. Next, section III outlines the background of our work, including state-of-the-art techniques for the determination of VFOA. The preliminary face detection and eye tracking system is described together with a simple attentional forecast algorithm in section IV. These are evaluated in section V. Finally, section VI discusses the next steps to be followed in terms of the computer vision component towards the final implementation of StudyBuddy.

II. REQUIREMENTS

As explained in section I, the computer vision component lies at the heart of the StudyBuddy system. It is essential for our system to have a means of interpreting users' behavior from an input video stream. This is precisely what the computer vision component will achieve. Our intention is for our product to be easily accessible to users, and for any attentional monitoring to be non-invasive. Therefore, we focus on the estimation of visual focus of attention (VFOA) from an input a video sequence, recorded with a smartphone's built-in frontal camera. The requirements outlined are put together in the following problem statement:

"The computer vision component intends to deliver an accurate, real-time attentional forecast. This is to be exclusively based on a video sequence captured by a smartphone's frontal camera, and results must be delivered to subsequent components within StudyBuddy."

Estimating attention exclusively from an input video sequence is achievable [2], as long as suitable (and accurate) computer vision methods are employed. The challenge related to this component resides fundamentally in two aspects: the input restriction (exclusively a video sequence captured by a smartphone's built-in frontal camera) and the need to implement computer vision techniques to interpret the input data. Note that the latter is essential for extracting the necessary information required for a robust attention estimate.

Literature shows that face detection, eye gaze and center tracking, and head pose estimation can be used for this purpose [2], [3]. Therefore, we aim at extracting these parameters accurately, and to use them as inputs to our attentional forecast algorithm.

III. BACKGROUND

We examine voluntary focus of attention. By definition, this originates from a human consideration or thought [4], and is scientifically classified as *'top-down'* attention.

VFOA estimation is currently a strong research topic in the field of computer vision. This is motivated by the scope of its applications which include video surveillance, navigational aids and replication of human intelligence [3].

Several methods have been designed to determine VFOA from video sequences. These involve extracting high-level information from recorded human eye regions, such as eye gaze, movement and center location [3], [5]. Other methods aim at forecasting attention from head pose, which has been proven to be strongly correlated to gaze and head orientation [2], [6], [7], [8]. Hybrid methods combine these different pieces of data for an increased robustness [2], [9] .

The component this report describes, combines face detection with eye-center tracking in order to conclude an attentional forecast. Both face detection and eye tracking

have been thoroughly investigated in the field of computer vision.

Remarkable methods for computerized face detection include Viola and Jone's machine-learning approach for object detection using a multi-layer cascade of simple Harr-like features [10]. This inspired local binary pattern (LBP) descriptors described in [11], and Dalal and Triggs' histograms of oriented gradients descriptors (used for training a linear SVM) [12]. In turn, these have inspired the deformable parts model [13], which accounts for interclass variation by representing objects as a collection of parts (arranged in a deformable configuration), and Dollar's combination of aggregate channel features and fast feature pyramids to achieve accurate yet rapid face detections [14].

Regarding eye tracking, Haar feature-based cascade classification has (again) been proven useful [15]. Additionally, this can be used to determine a user's gaze point on a screen, similarly to the method described in [16]. The latter discusses a gaze point evaluation algorithm consisting of an estimation of iris center followed by an optimal matching of the iris region. In [17], the authors describe a method which initially uses [10] to locate a face. Then, eye regions are preliminary estimated from standard human face proportions. Template selection, to account for different face shapes followed by template matching, is ultimately used to perform eye tracking.

IV. SYSTEM DESIGN

The first proof-of-concept for the computer vision component is inspired by the work discussed in section III, and consists of three subcomponents: face detection, eye tracking and attentional forecasting. The particularity of our work consists in the integration of these in an Android device.

OpenCV's Android SDK is used for the initial system design aiming at an early proof-of-concept which is essential given the time frame of the StudyBuddy project. This acts as a catalyst for camera interfacing and the integration of standard computer vision techniques.

The face detection and eye tracking blocks are inspired on the code available at [18]. Below, each of the subcomponents will be discussed in detail.

A. Face Detection

Due to the time constraint associated to the StudyBuddy project, no new method for computerized face detection will be designed, but the methods outlined section III will be used instead.

The preliminary face detector uses an implementation of [10]. The reason behind this choice is a combination of suitability (in terms of the performance of the method), and availability of resources. OpenCV's Android library contains functions which can make use of pre-trained Haar-feature based cascade classifiers available at [19] to perform real-time human face detection. Both training and implementation of this method are out of the scope of this project, again, as a result of the time constraint.

This submodule delivers as output the location of the user's face (given his or her presence) relative to the whole captured input frame.

B. Eye Tracking

Similarly to the previous submodule, eye tracking has been implemented via a series of functions available from OpenCV's Android library.

The methodology behind the selected approach follows from [17]: first, the user's face is located. Next, the system takes five input video frames to select an optimal eye template for the user. Eye center locations are then determined by the means of template matching. OpenCV provides several methods for template matching. Qualitatively, careful observation of the performance of each of these on real-time input video sequences does not allow for conclusions to be drawn regarding the choice of ideal matching method. Due to the unavailability of a precisely labeled database, at this stage, the performance of each of these cannot be quantified. Thus, following from [17], the normalized sum of square deviations algorithm is used for matching. This design choicer aims at optimizing tracking in terms of performance and computational complexity.

This submodule delivers as output the location of the user's (whole) eye regions as well as that of their pupils, relative to the captured input frame. This allows for the determination of the direction the user is looking towards relative to the location of his or her face.

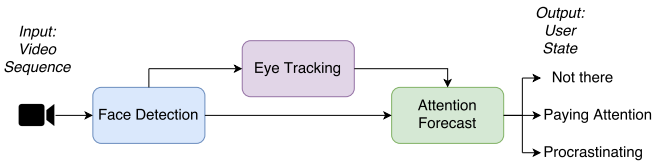

Fig. 1: Preliminary Computer Vision System Block Diagram

C. Attention Forecasting

As shown on figure 1, the outputs of both submodules A and B feed into submodule C. The purpose of the latter is to deliver an attention estimate. For the preliminary design, this is based on the following assumptions:

- Attention is defined as visual focus of attention.
- The user movement is constrained; (s)he is sitting down during the study session.
- The recording smartphone is placed at a fixed position where the user is perceivable by the frontal camera.
- The user's face can be seen at all times; (s)he does not cover his or her face.
- The user's eyes are coordinated in their movement.
- In front of the user there is nothing but study material.
- When the user is looking down, (s)he is studying.
- When the user looks away from the studying material, (s)he is procrastinating.
- When the user stops being perceivable for the camera, (s)he is procrastinating.
- If more than one user is detected, the attention estimate is undefined.
- The user is not wearing glasses while studying.
- The illumination of the scenery is optimal: the user can be seen clearly, there is no excess brightness.

This submodule, illustrated in figure 2 (enclosed by the dotted line), operates in the following manner: two variables are computed initially; the number of detected faces, and the boolean *'looking down'*, which is set to *true* when the user looks down at his or her studying material. The latter is computed by (empirically) defining the relative positions of human eye regions and pupil locations which correspond to a user looking downwards. Combinations of these two variables, as shown in figure 2, determine the output. This is always set as one out of three possible states: *not there*, *paying attention* or *not paying attention*.

In order not to have undefined outputs for any input combination, a second person on the recorded scene is assumed to be a distraction. Therefore, as soon as the system detects more than one face; the user is assumed to be procrastinating.

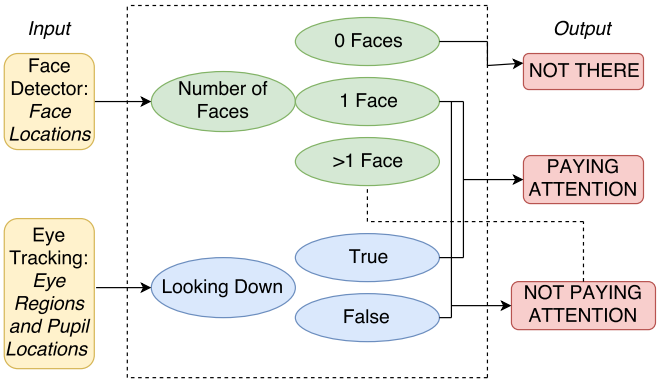

Fig. 2: Block Diagram of Attentional Forecast Module - we assume a second person in the scene to be a distraction

V. SYSTEM EVALUATION

A. Face Detection

This subsection explains the reasoning for the choice of Haar features as image descriptors for the face detection submodule.

Pre-trained Haar and LBP classifiers are both available from OpenCV's library. For the purpose of the StudyBuddy project, the face detector is required to run on an Android device, and operate in real-time while minimizing false-positive detections. The computer vision component is expected to ultimately deliver an output (user state) at an approximate rate of 2fps. Therefore, real-time becomes a (relatively) soft constraint. Haar and LBP cascades are compared experimentally using OpenCV to process still (face) images captured by a smartphone's built-in frontal camera. Three key observations are extracted from this experiment:

1) Haar descriptors require a longer processing time (per image) on average. However, the empirical upper-bound for processing time for these descriptors does not exceed 0.3 seconds.
2) LBP descriptors yield a larger proportion of false positive detections.
3) Both pre-trained classifiers only detect frontal faces (this implies room for improvement prior to the delivery of the final design, which should be capable of detecting tilted faces and profiles).

These conclusions drive the choice of Haar descriptors for the preliminary face detection submodule.

B. Eye Tracking

Due to the unavailability of an accurately labeled testing dataset for eye tracking, the performance of this submodule is not quantified in great precision. Instead, its behavior is evaluated qualitatively.

Given the user is facing front, four directions are defined: up, down, left and right. This submodule is tested on 10 human subjects, by asking each individual to direct their eyes towards each of the defined directions. Results are gathered in the form of a confusion matrix, as shown in figure 3.

The results obtained proof accurate eye tracking possible

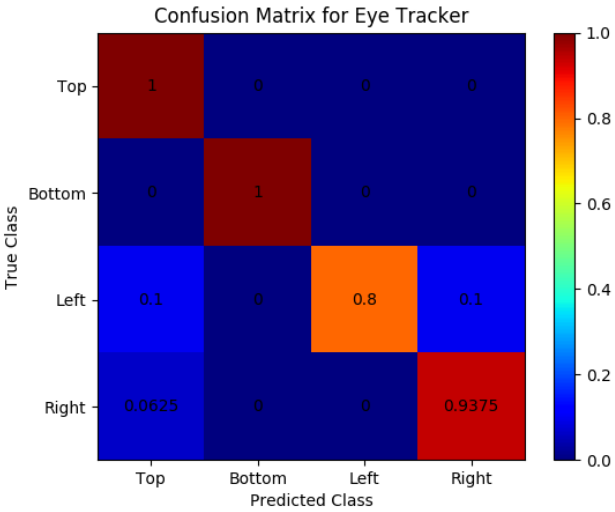

Fig. 3: Confusion Matrix for Eye Tracking Submodule

in our context and hence imply potential to deliver a robust attentional forecast.

C. Attention Forecasting

This submodule is tested in a similar manner to B; 10 human subjects are asked to replicate their behavior while studying from paper material. Each individual is asked to (at least):

- Replicate being focused
- Replicate being distracted (looking away)
- Leave the scene

The results obtained from this experiment are shown in figure 4. The following conclusions can be drawn:

1) It is feasible to deliver an estimate for VFOA.
2) The preliminary system can accurately determine whether the user is present (in scene) or not.
3) Given the user is in scene, the preliminary attentional forecast requires refining (special case must be drawn to the case when the user is paying attention).

VI. FUTURE WORK

Prior to the final delivery of StudyBuddy, the work of the computer vision team will be merged into a single component. In comparison with the preliminary system

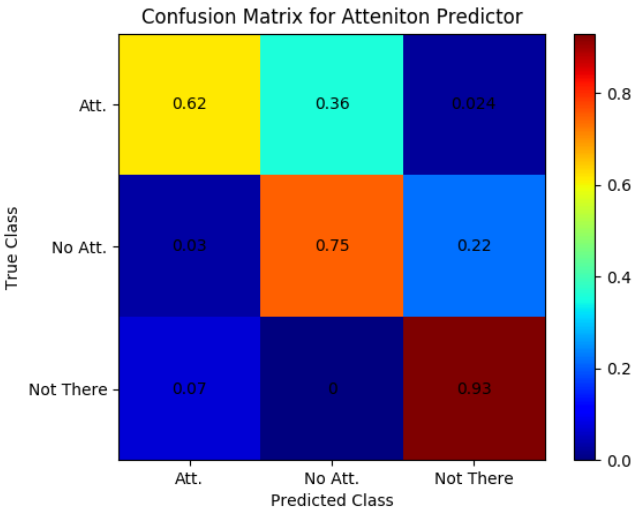

Fig. 4: Confusion Matrix for Eye Tracking Submodule

introduced in this report, the face detector submodule is expected to improve with the replacement of OpenCV's Haar feature-based cascade with an SSD neural network with MobileNet implemented in Tensorflow.

We recall from section III that head pose estimation has been proven useful in the estimation of VFOA. We therefore intend to strengthen our attention forecast by incorporating this, increasing the variety of data received by the attention forecast module. These changes will allow for a more robust attentional forecast.

Finally, we will introduce the concept of '*time outs*' (pre-determined time periods where the user is given to think, hence (s)he can look away from his or her study materials without it being interpreted as not paying attention). Figure 5 shows the block system diagram for the final computer vision component design.

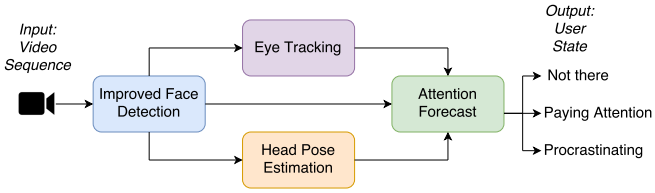

Fig. 5: Final Computer Vision System Block Diagram

VII. CONCLUSION

This report has introduced the preliminary design for the computer vision component of StudyBuddy, together with the requirement it intends to fulfill and the background of our work. Attentional forecasting exclusively from an input video sequence has been proven feasible and the steps to be taken towards the completion of the StudyBuddy project have been defined.

REFERENCES

[1] D. M. Tice and R. F. Baumeister, "Longitudinal study of procrastination, performance, stress, and health: The costs and benefits of dawdling," *Psychological Science*, vol. 8, no. 6, pp. 454–458, 1997. [Online]. Available: https://doi.org/10.1111/j.1467-9280.1997.tb00460.x

[2] S. O. Ba, H. Hung, and J. M. Odobez, "Visual activity context for focus of attention estimation in dynamic meetings," in *2009 IEEE International Conference on Multimedia and Expo*, June 2009, pp. 1424–1427.

[3] H. Cai, B. Liu, J. Zhang, S. Chen, and H. Liu, "Visual focus of attention estimation using eye center localization," *IEEE Systems Journal*, vol. 11, no. 3, pp. 1320–1325, Sept 2017.

[4] 2018. [Online]. Available: https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/what-is-visual-attention/

[5] K. Smith, S. O. Ba, J. M. Odobez, and D. Gatica-Perez, "Tracking the visual focus of attention for a varying number of wandering people," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1212–1229, July 2008.

[6] M. Danninger, R. Vertegaal, D. P. Siewiorek, and A. Mamuji, "Using social geometry to manage interruptions and co-worker attention in office environments," in *Graphics Interface*, 2005.

[7] M. Katzenmaier, R. Stiefelhagen, and T. Schultz, "Identifying the addressee in human-human-robot interactions based on head pose and speech," in *Proceedings of the 6th International Conference on Multimodal Interfaces*, ser. ICMI '04. New York, NY, USA: ACM, 2004, pp. 144–151. [Online]. Available: http://doi.acm.org/10.1145/1027933.1027959

[8] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel, "From gaze to focus of attention," in *Visual Information and Information Systems*, D. P. Huijsmans and A. W. M. Smeulders, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 765–772.

[9] B. Massé, S. Ba, and R. Horaud, "Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 14 p., Dec. 2017. [Online]. Available: https://hal.inria.fr/hal-01511414

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–511–I–518 vol.1.

[11] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, Dec 2006.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, June 2005, pp. 886–893 vol. 1.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, Sept 2010.

[14] P. Dollr, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, Aug 2014.

[15] Y. Li, X. Xu, N. Mu, and L. Chen, "Eye-gaze tracking system by haar cascade classifier," in *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, June 2016, pp. 564–567.

[16] W. C. Kao, C. Y. Lin, C. C. Hsu, C. Y. Lee, B. Y. Ke, and T. Y. Su, "Optimal iris region matching and gaze point calibration for real-time eye tracking systems," in *2016 IEEE International Conference on Consumer Electronics (ICCE)*, Jan 2016, pp. 443–444.

[17] Z. Zhao, S. Fu, and Y. Wang, "Eye tracking based on the template matching and the pyramidal lucas-kanade algorithm," in *2012 International Conference on Computer Science and Service System*, Aug 2012, pp. 2277–2280.

[18] "Simple sample, for eye tracking with OpenCV," 2018. [Online]. Available: https://github.com/hosek/eyeTrackSample

[19] "Open Source Computer Vision Library," 2018. [Online]. Available: https://github.com/opencv/opencv/tree/master/