Łódź, 10 March 2026

Dr hab. inż. Jacek Kucharski, prof. nadzw.
Lodz University of Technology
Institute of Applied Computer Science

REVIEW
of the doctoral dissertation of Mohd Faizan ANSARI, MSc, entitled:
„Using a Camera to Determine Human Gaze Point"

The review has been prepared based on the resolution of the Council of the Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, represented by the dean Prof. Dariusz Kania, PhD, DSc.

The dissertation has been prepared under the supervision of Prof. Paweł Kasprowski, PhD, DSc. This monography counts 128 pages organized in 7 chapters, bibliography and lists of figures and tables. Bibliography consists of 183 entries, among which two are coauthored by the Mr M.F. Ansari.

1. Topic, scope and objectives of the thesis

The presented dissertation concerns the development of gaze identification methods using low-cost built-in cameras and represents a significant engineering effort to democratize eye-tracking technology. The central premise of the work is to enable robust gaze estimation using standard, unmodified webcams found in consumer laptops and desktops, thereby removing the reliance on expensive, specialized hardware or head-mounted displays. This objective is well-aligned with the broader trends in Human-Computer Interaction (HCI) and Computer Vision observed throughout the last decade, where the drive toward ubiquitous, non-intrusive computing has spurred a transition from laboratory-grade equipment to edge-device implementation.

Defining in Section 1.1. the goal of the thesis the author states that the main objective of his work is „*to investigate whether a person owning a standard webcam, typically on a laptop or desktop, can perform gaze estimation on their machine without the need for any additional software or hardware*." Due to the low-quality images usually obtained from this kind of cameras (especially where the eye region might occupy only a small fraction of the frame) along with variable and uncontrolled lighting conditions a special data processing methodology is required. The author directed his efforts towards Deep Learning (DL) techniques, specifically to Convolutional Neural Networks (CNNs). In my opinion, the objectives outlined in the dissertation are ambitious and undoubtedly constitute a challenge commensurate with the scope of a doctoral dissertation, and the solution obtained represents an original achievement of the doctoral candidate.

According to the above defined goal the three research hypothesis have been formulated:

Hypothesis 1 (H1): *It is feasible to develop a CNN-based model that classifies a person's gaze into screen regions using low-resolution, low-quality eye and face images captured by a standard webcam.*

Hypothesis 2 (H2): *A model trained on low-quality eye and face images from a standard webcam, tailored to a specific individual, can achieve gaze estimation accuracy comparable to that of state-of-the-art eye trackers.*

Hypothesis 3 (H3):*Gaze estimation using low quality images collected from webcams that utilize models pre-trained with data of multiple users and fine-tuned for a specific person using transfer learning requires less data, and converge faster during fine-tuning compared to models trained from scratch for the specific person.*

These hypothesis indicate that the research has been structured progressively. Hypothesis 1 explores the fundamental feasibility of estimating gaze direction from noisy data by simplifying the problem into a classification task (region prediction). Hypothesis 2 advances to high-precision regression, postulating that limiting the domain to a specific individual (person-specific modelling) can compensate for sensor limitations, achieving accuracy comparable to commercial trackers. Hypothesis 3 addresses the scalability of person-specific models, proposing Transfer Learning as a mechanism to achieve high performance with minimal data collection.

I positively assess the completeness of the dissertation in which the candidate presented in particular:

- theoretical background concerning the most important areas covered by the research i.e. description of the human eye in terms of understanding the vision and where the person is looking, presentation of selected problems of artificial neural networks from their basis to CNN and transfer learning;
- comprehensive analysis of state-of-the-art methods for gaze estimation, including different approaches to this task, like: model based, feature based, and appearance based gaze estimation, supplemented by the overview of possible usage of deep learning and comparison of different gaze datasets;
- proposal of original structures of CNN networks as well as data acquisition and preprocessing procedures;
- experimental verification of the proposed solutions and the analysis of the obtained results.

In my opinion, the dissertation constitutes a comprehensive study with a clearly defined objective and its consistent implementation.

2. General evaluation of the dissertation

The main part of the thesis is preceded by a concise but comprehensive introductory chapter (Ch. 2) where the candidate presents some background information related to the subject of his research. First, the anatomy of the eye as a human organ is introduced along with some specific notations and an explanation of the mechanism of seeing as well as eye movements is offered. What follows is the second part of the introduction which is concerned with an

overview of artificial neural networks (ANN) and includes some basic notions and concepts of ANN, their structures and formal mathematical description alongside more advanced solutions and techniques such as convolutional neural networks (CNN) and transfer learning. Generally, this theoretical material is correctly presented, well balanced and provides necessary mathematical insight into the methods applied in the research. It also proves a good level of competence of the candidate in the areas covered by the thesis. The introductory part is followed by an insightful overview of the state-of-the-art methods for gaze estimation (Chapter 3). The candidate has presented an exhaustive review of the research based on numerous publications dated mostly in the last decade which proves that the undertaken task is reasonably up-to-date. He has also included a historical account of the research dating back to the beginning of the 20th century when the first trials of gaze estimation were documented. The importance of the paradigm shift around 2015 towards deep learning methods has also been pointed out and solutions based on CNN and transformer based methods have been discussed. Additionally, a description of some of the datasets, which have been created over the last decade thanks to intensive work on gaze estimation, have been presented and some characteristic features as well as the collection methodology details have been given. This chapter constitutes a good justification for Mr Ansari's research and focuses on several unresolved problems in the area of gaze estimation which have become a goal of the work presented in the dissertation.

The following three chapters (Chapter 4, 5 and 6) refer directly to the hypothesis formulated at the beginning of the dissertation (H1, H2 and H3, respectively) and constitute the main and most important part of the thesis. The structure of all these chapters is similar as they are based on papers already published (ch. 4 and 5) or prepared for publication (ch. 6). This naturally supports the solutions and results presented (at least for ch. 4 and 5), but as the content of these chapters is taken almost fully from the cited papers some repetitions are noticeable which makes the entire dissertation slightly redundant. For example, Sections 4.2.1 and 4.2.2, which are devoted to data collection and preprocessing, contain very similar methodology descriptions as shown in 5.2.1 and 5.2.2 as well as in 6.2.1 and 6.2.2. The same concerns apply to introductory sections in these chapters (namely sec. 4.1, 5.1, 6.1), where similar remarks on the importance of gaze estimation alongside typical problems with data collection and processing are repeated several times.

The core assumption in Chapter 4 is that while exact pixel-level regression might be too ambitious for a noisy webcam signal, a coarse-grained classification into discrete regions should be achievable. Two datasets were created collecting images form a single person and from four subjects, respectively. In all cases the test subjects were asked to distribute their attention equally to 20 discrete points on the screen. The acquisition procedure was organized in such a way that a variety of expositions were taken into account, which makes the datasets suitable for proper training of neural networks. The machine learning task was defined as a multi-class classification problem to predict which of the 20 zones the user is fixating on in a given input image. Three architectures of CNN have been proposed for processing a one eye image, face image and both eye images as input of the network. Although the author claims on page 52 that "*the proposed architectures were chosen from a set of several possibilities that were initially studied*", the final choice is arbitrary and not discussed in the dissertation and therefore it should be regarded as a weakness of the proposed solution. The results presented in section 4.3.2 demonstrate the possibility for achieving reliable eye-tracking accuracy in diverse environments (in most cases the accuracy exceeded 80% with the highest value of over 88%), which provides support for hypothesis H1. However, table 4.3 reveals that the sharpening of the images improves the total accuracy significantly, which confirms

that this type of preprocessing can compensate, to some extent, the "low quality" of webcam images and it aids the CNN. It is not also clear what the detailed data are presented for in tables 4.3 – 4.5, namely accuracy values during training, as there are no comments to that in the text.

The main focus of Chapter 5 is regression, and it covers the training of models on the data from one user, which corresponds directly to hypothesis H2. Two CNN different architectures were examined with one image or two images as input. As in the previous case the structures were chosen arbitrarily but in this case the candidate analysed sets of values of some parameters, e.g. learning rate, dropout and kernel size. The results presented in section 5.2.5 and discussed in section 5.2.6, with best-case accuracy in the range of 1.98 to 1.14 degrees, prove undoubtedly high performance of the proposed models and support hypothesis H2. However, this was achieved by training the network on images of a single user which can be a strong reason of network overfitting to the subject, which significantly reduces the generality of the proposed solution. Moreover, collecting ~12,000 labelled samples is a massive burden for a user and it thus makes the solution hardly useful in practice. Some of the conclusions drawn in this chapter seem to be quite obvious and do not necessarily give valuable insight. For example on page 65 the author claims that "*when the model was transferred to a machine equipped with a graphic processing unit (GPU) for training, there was a notable decrease in training time*"; (p. 66) "*Based on this experiment, it becomes clear that the learning rate plays a crucial role in determining the performance of the network.* " and (p. 73) "*This comparison demonstrates that training the model with a single user can result in higher accuracy compared to models trained on data from multiple individuals*". This part of the dissertation exhibits also some editorial weaknesses which make the reading difficult. The text directly below Table 5.6 is exactly repeated below Table 5.8 as well as parts of the text repeated above and below Table 5.7. Additionally, Table 5.8 has not been discussed and referred to in the text.

A key element from the perspective of the research objective is Chapter 6 of the dissertation, devoted to the use of transfer learning for gaze estimation purposes. The candidate applies a well-known strategy to reuse knowledge learned from large-scale datasets and fine tune the model for a specific task rather than training it from scratch. This chapter investigates the performance of gaze estimation under limited data scenarios. For these purposes the data from 19 participants were collected with substantial diversity in terms of ethnicity, gathered across multiple sessions and under varying lighting conditions. The participants were asked to click on a set of 54 points distributed across an entire screen. Around 90.000 images gathered in such a procedure constitute a reliable and well prepared material for further steps of developing the gaze estimation method. Moreover, this dataset is a tangible contribution of the thesis to the research on computer vision and image analysis. The transfer learning procedure was organized in such a way that the model was pre-trained using data from all but one participants and after that the model was fine-tuned specifically for the excluded participant. Both performance of the models and their convergence were analysed for different dataset sizes (form 500 to 100 images, and in extreme cases down to 10 images). Results for models with and without pre-trained weights prove main assumptions of hypothesis H3, exhibiting in most cases higher performance and better convergence of pre-trained models. The most important finding of this part of the research is the indication of capability of transfer learning to obtain a reasonable accuracy of gaze estimation having limited data available. Even though the candidate states on page 97 that "*there is room for further refinement to achieve comparable accuracy of high-end system*" it improves the applicability of the proposed solution in real systems. This approach seems to be especially

important for edge computing where small models fine-tuned on the user's device is needed. Chapter 6 is also not free from unnecessary obvious statements. For example, on page 87 the author includes the information that "*The relationship between inches and centimetres is 1 inch=2.54 cm*" and on page 91 states that "*These results lead us to the conclusion that the results can be further improved by increasing the number of training images*".

Concluding the dissertation in Chapter 7 Mr M.F. Ansari summarizes in a proper way its contribution to the area of gaze estimation, pointing out that he developed a CNN-based models for discrete classification of a person's gaze, evaluated the influence of personalization of the solution as well as studied gaze estimation solutions which use the transfer learning. All these contributions refer directly to the goal of the research and the formulated hypothesis.

3. Remarks and questions

The reviewed dissertation is not devoid of passages that raise certain doubts or give rise to specific questions.

1. The outputs of CNNs applied to gaze estimation, particularly in the regression cases, contain the coordinates of the person's gaze at the screen in 2D space (namely, x and y coordinates, as shown e.g. in Figure 5.3, 5.4 and 6.4). However, the Mean Absolute Error (MAE) loss function utilized for CNN training defined by equations 5.1 and 6.1 includes only the y coordinate. What is the reason of omitting the x coordinate in these formulae? What is the difference between $\overline{y}$ and $\hat{y}$ notions in equ. 5.1 and 6.1, respectively?
2. What was the influence of inaccuracy of determination of the distance between the screen and the viewer's eyes (*Dist*) while recalculating the errors from pixels (centimetres) to degrees by equations 5.3 and 6.2? It was assumed to be 30-50 cm, but was not measured?
3. Could the candidate comment on the value of the comparison of the developed gaze estimation method to other state-of-the-art methods presented in Table 5.9, taking into account significantly different characteristics of the datasets used in analysed cases, particularly as the dataset used in Chapter 5 was collected from only one person?
4. Do the candidate see the possibility of using different pre-processing methods instead of Viola-Jones classifier and Haar cascade for objects' detection? How beneficial could it be for the results of gaze estimation in the proposed solution?
5. When discussing partial results of the transfer learning method the candidate states in section 6.3.2 on page 87 that "*For right eye, transfer learning, the MAE started at 164.51 for 500 images, increased slightly to 197.47 with 100 images, as shown in Table 6.3. The steady increase shows the model's capability to maintain stable performance, despite a decrease in dataset size*". It is not clear how this statement should be understood.

4. Detailed remarks and editorial evaluation

The reviewed dissertation by Mr M.F. Ansari, MSc, presents the most important results of the conducted research in a fairly clear manner. However, the candidate did not avoid a number

of stylistic shortcomings in the language as well as editorial errors, some of them are listed below:

1. p. 17 there is "*each of its each serving*", but should be *each serving*;
2. p. 18 there is a reference to equ. 2.4.1 but should be 2.12;
3. p. 19-20 there is "*have been become*" but should be *have been* or *have become*;
4. p. 64 and p. 65 there are references to Table 4.2 but should be Table 5.1;
5. p. 86 there is "*which we can be observed*" but should be *which can be observed.*

5. Final conclusion

The reviewed doctoral dissertation by Mr M.F. Ansari, MSc, constitutes a complete solution of scientific problem and, in my opinion, contains the author's original contribution within the discipline of information and telecommunication technology. The doctoral candidate has demonstrated a good level of knowledge in the areas covered by the dissertation, namely machine learning and image processing. The critical remarks formulated in this review do not change my overall positive opinion of the dissertation.

In the light of the Act Law of higher education and science of July 20, 2018 I hereby state that the doctoral dissertation of Mr M.F. Ansari, MSc, has satisfied the requirements, therefore, I recommend to submit the dissertation for public discussion and defence.