March 16th, 2015

To: Guest Editor: IJCV special issue on Human Activity Understanding from 2D and 3D cameras

Obj: #VISI-D-14-00453 submission - *Gaze Estimation in the 3D Space Using RGB-D sensors. Towards Head-Pose And User Invariance.*

Dear Guest Editor,

This letter is in response to the review of our submitted manuscript referenced above on gaze estimation using RGB-D sensors. We would first like to thank the reviewers and guest editor for their time and valuable comments. We have taken into careful consideration each one of these comments, and have prepared a detailed response in a separate document adjoint to this letter. We have made this answer as self-contained as possible to facilitate the review process. Furthermore, addressing these comments led to many improvements of the manuscript. Before summarizing the main changes in the paper, we would like to recall the main characteristics and contributions of the paper:

- we propose a methodology for 3D gaze estimation under the challenging conditions of non collaborative users, directly addressing the problems of head pose and user(eye) appearance variations. We emphasize the 3D capabilities of the approach as the addressed range of head pose variations and gaze directions are larger than any previous works on gaze estimation.

- we propose to exploit RGB-D data to track the head pose and to rectify the eye appearance into a canonical head pose viewpoint, bringing head pose invariance to the appearance based gaze estimation paradigm, even for important head pose variations.

- We address the problem of user appearance variations through the training of user invariant models for appearance based gaze estimation. In this context we propose a novel eye alignment method which implicitly address the problem of inter-person eyeball center misalignment, circumventing feature detection.

- The overall method including our head pose and user invariant strategies was validated through extensive experiments. In particular, our proposed alignment consistently outperformed strategies based on eye corners, whether extracted *manually or automatically.*

- Finally, we demonstrate the usefullness and the accuracy of our approach on one representative application, namely automatic gaze coding in natural interactions. This application is of *high relevance*, as it helps to motivate, further validate and to demonstrate our methodology's applicability to challenging scenarios.

Major modifications: We now would like to summarize the main changes done to the manuscript:

- *Head pose tracking experiments*: We evaluated the head pose tracker on two publicly available benchmarks: the BIWI Kinect head pose dataset and the ICT 3D Head pose dataset. This had two purposes: i) to validate the high accuracy of the head pose tracker, in comparison to representative RGB-D methods, as suggested by Rev3 and; ii) to quantitatively and qualitatively confirm the need of the 3DMM offline fitting, by also evaluating the usage of a mean face model. The experiments confirm the 3DMM's need for accurate head pose tracking and time-consistent eye cropping. This address at once concerns by all reviewers.

- *Automatic landmarks detection-based experiments*: We have performed additional alignment experiments in which a state-of-the-art landmarks detection algorithm (Kazemi and Sullivan, CVPR2014) is used to detect eye corners. Results show that such an alignement performs much worse than our alignment approach. Note that the experiments we had already conducted using *manually* annotated landmarks subsume such eye corner alignment strategies.

- *Implementation details section*: all reviewers enquired details of our system, such as the 3DMM we used, the hyperparameters values, the computational cost of different parts of the methodology and overall speed. We therefore included an "Implementation details" section where all these elements are described in detail to the reader.

- *Discussion and future work section*: among the many insighful reviewers comments, there are some which are not possible to address directly due to either time constraints, their relevance with respect to the contributions or because the data at hand is not adequate. Nevertheless, as a way to address these points and present the limitations of our work, which we considered of interest to the reader, we added a "Discussions and future work" section.

Minor modifications: We have addressed all reviewer's comments, most with direct modifications in the paper. We would like to highlight a few relevant points:

- *Person-invariance* (cf Rev1): We have clarified the person invariance aspect of our method within the eye image to gaze parameters mapping function. Therefore, we better described the contexts in which the alignment is used, which does not contradict the invariance claim.

- *EYEDIAP*: We clarified to Rev3 and the reader that the automatic annotation process is of high quality. Many frames are ignored because the visual target (ground truth) can not be determined like when the ball target leaves the sensor field of view. EYEDIAP consists of *non-stop* recordings, where such events are common.

- *RGB-D based methods* (cf Rev1): We cited and discussed the only and very recent RGB-D based gaze estimation methods. These rely on local features tracking from higher resolution data and were tested on less challenging conditions than in our paper.

We hope that these new experiments, clarifications, and paper modifications will satisfy the reviewers as well as address your own comments.

We thank you again for your time and consideration of our manuscript.

Sincerely,

Kenneth Alberto Funes Mora and Jean-Marc Odobez