Response to Reviewer 1

We thank the reviewer for his time and comments. Below, we provide our answers to his comments and to the unclear points that were raised, and what we have done to clarify the paper and take comments into account. Note that in our answer, all references (Equations, Figures) refer to the new version unless stated otherwise.

Comment: *This paper addresses the problem of 3D gaze estimation. In particular, it solves the problem in a person-independent and head pose-invariant manner.*

The key of the proposed method is the use of the depth data. The depth information is used to obtain a 3D face model and then synthesize the frontal appearance. Therefore, the rest appearance-based gaze estimation can be done as if the head pose is stable. In addition, the paper achieves person-independent estimation by using training images from multiple subjects with an eye alignment method.

Response: Thank you for your careful review. Note that to address the particular points you have raised for discussion, in the new version of the paper, we have conducted new experiments and added several Sections (Section 6.1 -Implementation details and speed-,Section 6.2 -Head pose experiments- and Section 7.1 -Tracking results; Section 9 -Discussion and future work) that we hope should answer most of them. In addition, in the following link you can also find video results of the gaze coding problem:

`https://www.youtube.com/playlist?list=PLQeF4owrybh2s5bwHW8tpxLNPC1FjANu1`

Please notice the video/playlist settings are not public. Thus, we kindly ask you not to redistribute to protect the privacy of the participants[1].

In the following we will now detail our specific answer.

Comment: *1) It is unclear about the RGB-D sensor's details. How about its accuracy and performance? How will it affect the final accuracy? This should be examined in experimental part.*

Response: The sensor we used is a Microsoft Kinect, 1st generation, for XBox 360. This device is well known and available in the consumer market. Its specifications in terms of accuracy and performance are given by Microsoft. However, as reported in the paper describing the details of the EYEDIAP database [Funes Mora et al (2014)] we further used the calibration toolbox which implements this method:

Daniel Herrera C., Juho Kannala, and Janne Heikkila. *Joint depth and color camera calibration with distortion correction*. PAMI 2012. ,

and is provided by the same authors. In their experiments, they reported the standard deviation for the depth error is around 1.5mm at a distance of 0.9 meters (conditions for the screen target in our data) and of around 3mm for the floating target conditions, at a distance of $\approx$ 1.2m.
To account for your comment, we have updated Section 6.3 of the paper to comment on these accuracy:

"… *For our main task, we used the publicly available7 EYEDIAP database described in Funes Mora et al (2014) which provides data recorded using a Microsoft RGBD Kinect (1st generation, for XBOX 360)*

[1] Note that within youtube the viewers are anonymous (to the exception of the country access statistics, if one would check, which we will not)

that was further calibrated with the method of Herrera C. et al, (2012). …"

and later, discussing the floating target condition:

"… *As people were seated at a distance of* $\approx 1.2m$ *from the sensor, the typical depth error at this distance is around 3mm (according to Herrera C. et al, (2012)), the typical eye image size is* $\approx 13 \times 9$ *pixels, and the gaze space is as large as* $\pm 45^\circ \times \pm 40^\circ$. *The head pose variations follow a similarly large range. …*"

and similar clarification were provided for the continuous screen (CS) condition.

We agree with the reviewer that experimental conditions can greatly affect the reported errors. In the EYEDIAP database, we have tried to achieve this by comparing data at two different distance conditions (as it can clearly affect both the depth measurements for the tracking and rectification, as well as the eye image resolution), and for different tasks, but it does not allow for a proper evaluation of the link between the device noise and the algorithm performance, as you suggest. To evaluate the influence of a given sensor on the given algorithm would imply the collection of a database only for this task, which we feel is beyond the scope of the current paper. Furthermore, to our knowledge there is no database available which would allow to compare gaze estimation across RGB-D devices, as the EYEDIAP database is the only publicly dataset on gaze estimation using RGB-D data. Nevertheless, as we rely on the well established and well disseminated Kinect sensor (we have similar results with the Primesense sensor), we believe that the conclusions drawn in the paper are representative of what can be achieved.

In the paper, our main objective has been to document, as detailed as possible, the sensing conditions, i.e. device type, resolution, sensing distance, etc, as shown above. Furthermore we have made an important effort to make the EYEDIAP database publicly available, such that other researchers can have a close look at the quality of the data.

In addition, we have added new results on head pose tracking for the BIWI database, and the 3D Head Pose ICT database, which can also show that (at least on the head pose tracking side, which is one of the main part that depends on the depth data), we obtain similar results on other datasets.

Comment: *2) In sec 3.2, it is not clear whether the 3DMM is original. If not, please cite the related papers.*

Response: Thank you for pointing this out and we apologize for missing this important detail. The 3DMM we use is the *Basel Face Model*:

P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. *A 3D Face Model for Pose and Illumination Invariant Face Recognition.* In Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), workshop for Security, Safety and Monitoring in Smart Environments, Genova, Italy, 2009.

which we have been using and constantly citing in our previous work. We were surprised to find we forgot to put the citation in this paper. We have corrected this mistake in this revision. Within the new "Implementation Details" section, you can find the following:

"… *The 3DMM we used is the Basel Face Model (BFM) (Paysan et al, 2009). This model contains*

53490 vertices and has 200 deformation modes. The BFM was learned from high resolution 3D scans of 200 individuals (100 male/100 female) thus it span a large variety of face shapes with neutral expression. ..."

Comment: *3) This is my major concern: please provide details on the cost to learn the face mesh for each user. In particular, whether it is necessary for every user/every time? What is the time cost? Again, how will it affect the alignment/gaze accuracy?*

Response: The model has to be fitted only once as it is not expected for the subject's face shape to change from session to session. Indeed, the EYEDIAP database includes sessions that were recorded 9 months apart for 3 people. For these 3 persons, the face model built 9 months prior to the main database recordings was the one used to perform tracking and eye alignment on the newer recordings. So, in practice, we do not expect the re-fitting of the model at every session to improve (or degrade) the alignment or gaze accuracy. Refitting may be required only after months/years due to important weight variations or aging. Notice as well that the same fitting results can be used across different sensors (in our lab demonstration, we used the same model for Kinect and Primesense sensors), provided they are calibrated.

To address your concern, we have clarified the above points in the new Section 6.1 (Implementation details and speed), which addresses as well the time cost issue. Sample paragraphs from this Section include:

"... *Given a few annotated frames with landmarks (typically 1 to 5 frames), the fitting algorithm takes from 5 to 20 seconds to optimize. Note that since people face shape is not expected to change much, this step is only performed once per subject, which means that the fitted model can be reused across sessions. ...*"

In terms of the time cost, further optimizations could be done to our implementation. We used 41585 vertices, which could be heavily reduced and still maintain the fitting quality. Nevertheless, we think 5 to 20 seconds for optimization per user is very much acceptable for many applications. Further information is provided on the gaze tracking part:

"...*Overall, the gaze tracking takes around 100ms per frame. Note however that this is a research implementation, where the most time-consuming elements are the data pre-processing (3D mesh creation) and the head pose tracking. The head pose tracking is CPU based and alone takes from 20 to 100ms (depending on the amount of head pose changes during consecutive frames). A careful GPU-based implementation could greatly increase the speed. ...*"

Comment: *4) Another major concern: to do eye image alignment, several annotated samples are needed for each user. It is therefore not really calibration-free or person-invariant.*

Response: We understand the reviewer concern but do believe that we do propose person-invariant and calibration-free gaze models as we try to clarify below.

First, by person invariance we mean gaze model person invariance, i.e. where building a gaze model denotes the problem of *mapping the eye image appearance to gaze parameters when the test subject' specific eye appearance relation to gaze is unknown to the system.* Such regression model is thus trained

from a database containing as many people as possible, with appearance variations due to factors like ethnicity, specific eye geometry, illumination, etc. From this point of view, our proposed method is indeed person invariant.
While this might not have been clear in the paper, we want to emphasize that the proposed alignment step can be used in *two* different but related contexts:

- *Person invariant gaze model training*: we argue it is better to align the eye images across different subjects in the *training phase* such that the training data is consistent prior to the training of a model. The machine learning regression method can then learn to address further variations in geometry, illumination, eyelids shape, etc. In the literature, this alignment has often been done by bringing the eye corners to a canonical location, In the paper, our experiments show that eye corners based alignment even in the optimal case of manual annotations does indeed improve the results, but not as much as our alignment approach that is directly driven by gaze (and eye appearence) information.

- *Test subject adaptation*: in this case we assume there is a person invariant eye appearance (image) to gaze parameters regression model trained from a database of people. Note that this person-invariant model can directly be exploited for a test subject to predict her gaze direction without requiring any annotation. This is demonstrated in the gaze coding experiments, where good results are already obtained without applying a test user eye alignment. However what we have further proposed (*Alignment Procedure*, Section 5.2.2.) is that learning the person-specific alignment parameters of the test subject from a small set of gaze annotated samples (1 to 5) can make the test data more coherent to the learned person-invariant gaze estimation model. We argue this is a possibility which depends on the application, and we would like to raise the reviewer's attention on the gaze coding problem from natural interactions which we have addressed in this paper. This application clearly depicts a scenario in which it is not possible to train a gaze model for a given subject, but where we can obtain, from additional information, a few gaze annotated samples. This extra step improves the gaze coding accuracy.

To help better clarify the two above points, we have rephrased the last paragraphs of Section 5.2.2 in the paper as follows:

"... *We call the method described by Eq. 9 Synchronized Delaunay Implicit Parametric Alignment (SDIPA). In the paper, we have exploited it to address two related tasks.*

1. Person-invariant gaze model training. In this task, the goal is to align a gaze annotated training set comprising different subjects prior to learning the gaze regression models. We expect that exploiting aligned data will result in more accurate models. This is achieved by optimizing Eq. 9.

2. Eye image alignment for a test subject. The eye gaze model learned using the above alignment method (task 1) is person-invariant, and can readily be applied to any new test subject. However, in some situations (see for instance the gaze coding experiments in Section 8), there is the possibility to gather for a test user a few samples with gaze information (e.g. a person looking at known location like another person, or simply, looking at the camera) that can be further exploited to improve the result. In this case, the same method can be used to find the eye alignment of this user with respect to the already aligned training set using these gaze annotated samples. This is simply done by adapting Eq. 9 and conducting the optimization only w.r.t. a single subject (e.g. the parameters θ_j of subject j considered as our test subject) while the other $\{\theta_i\}_{i \neq j}$ remain fixed. This $2nd$ case can be seen as an adaptation step that is *highly valuable in an HRI and HHI scenario, where even if conducting a proper gaze model training session is not possible, it might still be feasible to detect in a supervised or unsupervised manner instants at which the subject is fixating a given (known) target and use the few collected samples to find the test subject's alignment. ...*"

Comment: *5) The warp function for alignment is described by θ. It is unclear what is the transformation type. What is the detail of θ? In L38/L50, the aligned eye images should be the same; can the alignment/transformation with θ guarantee this? How to handle the difference in color, shape and iris size in different people?*

Response: The warp function type can actually be of many types and include translation, scale, rotation, or even affine parameters. This is why we had purposely left it generic from the beginning of the paper, and had specified at the end of the *Alignment modeling* section in Section 5.2.2 that we were using a translation. We have kept this information there and further repeated it in the new "Implementation Details" section 6.1:

"... *The warping function f we used in this paper is a translation ($f(\mathbf{u};\theta) := \mathbf{u} + \theta | \theta \in \mathbb{R}^2$)) which we found sufficient to (implicitly) align the eyeball position across subjects after rectification. ...*"

We agree with the reviewer that there is no guarantee that the aligned image of a particular subject can perfectly match the aligned image of another subject, even for the same gaze direction, due to variations in color, shape, etc. Still, we expect our similarity assumption to hold *on average*. That is, we are not expecting Eq. 9 in the paper to reach 0, but by relying on a large pool of people and gaze directions in the cost function, we expect the optimal solution leading to the minimum of the objective function to properly align the eye images so that the main eye structures (eyelid and iris) are aligned. The many experiments conducted in the paper relying on this alignment assumption empirically demonstrate that this is a valid assumption.

To account for this comment, we have modified the paper when describing the alignment problem formulation, which reads now:

"...*To compute the parameters $\Theta := \{\theta_i\}_{i=1}^{M}$, we make the assumption that* when two subjects gaze in the same direction, *their aligned image (particularly the iris region) should match and their intensity difference should be minimal. Note that while this might not necessary hold for all gaze directions and pairs of people, we expect this assumption to be valid on average, i.e when considering a large number of people and gaze values to constrain the parameter estimation. ...*"

Comment: *6) There are previous methods based on RGB-D sensors. Comparisons and discussions should be added.*

Response: Do you refer to RGB-D methods for head pose tracking? or for gaze tracking?

It is true there are diverse methods in the literature for head pose estimation from RGB-D data, and even though this is less relevant according to the paper contributions which are centered on gaze estimation, we have made experiments comparing our head pose tracking method with two other methods using two publicly available benchmarks. The results are shown in Section 7.1. These experiments were also interesting to validate the need for a personalized face model.

W.r.t. gaze estimation, we thank the reviewer for pointing out to this fact. The only ones we have found are indeed very recent RGB-D methods:

Ref1: LI Jianfeng, LI Shigang. *Eye-Model-Based Gaze Estimation by RGB-D Camera*. Computer Vision and Pattern Recognition Workshops 2014

Ref2: X. Xiong, Q. Cai, Z. Liu, and Z. Zhang. *Eye Gaze Tracking Using an RGBD Camera: A Comparison with a RGB Solution*. PETMEI 2014

These approaches, even though very interesting, rely on geometric approaches of the problem and on the extraction of local features. Ref1 uses the approach by Timm and Barth (VISAPP 2011) to find the iris center through a gradient direction voting scheme, whereas Ref2 uses a variation of the Starburst algorithm (Li et al, CVPRW 2005) to fit an ellipse, and are thus suffering from the same limitations that we have described in our related work about geometric based methods. The usage of depth data in these methods is mainly for the head pose tracking: Ref1 uses the Microsoft Kinect SDK's tracker whereas Ref2 transforms 2D landmarks detection into 3D points based on depth measurements and compares it to a person specific 3D landmarks position model (to infer the head pose transform).

Notice in both cases, a Kinect of first generation was used, but it had to be configured to deliver an RGB resolution of 1280×960 (at the cost of reduced framerate) to achieve the features extraction. The users also had to sit closer to the sensor. In our experiments we used 640×480 and, motivated by possible applications, the users were at large distances from the sensor. Therefore, the difference in data quality is very large. In addition, both papers report the user was free to move the head however, according to the experimental settings described (e.g. Ref2 uses a screen as target), we can assume the head pose variations were actually minimal. This is radically different to the conditions we address, where in the data collection the participants were asked to deliverately do very large head pose variations, again to be representative of HHI, HRI like scenarios.

Furthermore, even when using this higher data quality conditions, the authors of Ref1 report "*But we also notice that our method could be more accurate if the pupil center can be detected better. Our future work would consist in finding a more stable and accurate algorithm on detecting pupil center.*". This indeed confirms the limitations of these method, as we have discussed in the related works section.

Appearance based methods, which is a main focus of this paper were indeed proposed as an alternative methodology to address the scenario in which using model-based, or geometric based, approaches have difficulties, as in the conditions we address.

In summary, these methods would have much difficulties to work under our low-resolution images, or in the floating target situation which has much greater variation in terms of head poses. As the data used by these author for experiments is not publicly available, a direct comparison with their approach is not possible. In addition, implementing their method is beyond the scope of this paper; however, as our data and experimental protocol is public, we hope that comparison will be possible in the future.

Nevertheless, it is indeed important to cite/discuss these works, as we have now included the following parragraph within the related works section:

"*. . . Yamazoe et al (2008) proposed a similar strategy with a reduced calibration session. Ellipse fitting was also used, but obtained from a prior segmentation of the eye region based on thresholding.*

Recent methods apply a similar strategy to RGB-D data. Jianfeng and Shigang (2014) infer gaze *based on iris center localization and the Microsoft Kinect SDKs head pose tracker. The eyeball center is refined from a calibration session, whereas the rest of eyeball parameters are fixed. Xiong et al (2014) used the same sensor, but relied on ellipse fitting and facial landmarks for 3D head pose tracking and to build a person specific facial landmarks position model. Their calibration method infers additional eyeball parameters. However, in both cases, the Kinect was configured for the highest RGB resolution of* 1280×960*, needed to allow local features tracking, the range of gaze directions was small and head pose variations were minimal.*

Nevertheless, an important limitation of the previous methods is the need to detect local features, which require high resolution and high contrast images. ..."

Comment: *7) In 'contributions', the authors state that 'almost no previous works address the problem of gaze estimation within a 3D space'. I am not sure whether the authors mention this as a merit. In fact, the difference between 3D direction and 2D position is only a transformation.*

Response: The reviewer is again right, in theory from the 2D to the 3D case there is only a transformation. However, this transformation requires to know additional information such as the 3D head pose, eyeball position, camera position w.r.t. world coordinate system, and screen pose (if relevant), etc. While it is true that geometric methods do allow by nature to do such 3D reasoning (but they have their own limitations in terms of sensing and feature extraction), within the literature of appearance based methods most authors avoid such 3D manipulation and instead estimate the mapping directly from the eye appearance to the point of regard whichcin most situations correspond to 2D screen coordinates: this way, they circumventing any geometric analysis, but has the disadvantage of poor generalization and of limiting the application to such creen gazing scenarios

To our knowledge, only the methods of [Lu et al BMVC 2011] and [Sugano et al CVPR 2014] have developed a 3D methodology. These methods nevertheless require training data from a diversity of head poses and furthermore, the addressed range of head poses and gaze directions is limited to a much smaller range than in our case. Our approach, which develops further an earlier publication of ours [Funes and Odobez at CVPRW 2012], stands out because it is able to address a much larger range of head pose and gaze directions combined in the 3D space. Furthermore, it allows for training from a single unrestricted head pose and generalizes well to the continuous range of head poses. This is validated in our experiments.

Note that thanks to this 3D property, this makes our approach attractive for other types of applications besides screen gazing in HCI, like Human-human interaction analysis (as our gaze coding experiment shows) or HRI. We therefore believe this methodology does stand out and has merit as *gaze estimation in the 3D space*.

To address your point and summarize the above comments, we have updated our contributions in the related work section as follows:

"... *As can be seen, few methods have addressed simultaneously head pose and user invariance. Moreover, to our knowledge, almost no previous works relying on the appearance framework have addressed the problem of gaze estimation within a 3D space, which allows the natural application of gaze techniques to more diverse HHI or HRI scenarios besides the traditional screen gazing case. ...*"

Comment: *8) To examine the person-invariant, leave-one-out experiments are used. This means to use all other subjects' training data for the test subject. I want to see the results by using training samples* *from no more than half of the subjects.*

Response: Unless we missed one important point, we were not sure about the motivation of this request.

Indeed, a leave-one-out evaluation is one of the most well established methodologies used in machine learning to perform evaluation in this type of settings (as far as the data used for training does not incorporate information from the test data, e.g. in particular here, the same user). Furthermore, for these type of methodologies, it is expected to use as much training data as possible in order for a model to generalize well to unseen data, and see how well the proposed system can perform in a realistic scenario. This is particularly true in our case where we only have data from 16 people (15 for training in the leave-one-out experiment) to perform training. In addition, conducting such experiments would be time consuming [2].

In summary, we believe that making such a proper evaluation, even if possible, would mainly show as in many other problems that using less training data would make the accuracy of a person invariant model to degrade. While we understand that seeing how much it degrades could be an interesting question (and in this case, we would need considering other amount of training data, e.g. 25, 50, 75% of the), we have left out such a proper evaluation for future work for the sake of time.

To account for this question, we have added in the new discussion section (Section 9) of the paper the following comment:

"*. . . In this direction, it could also be relevant to evaluate whether there is an impact of the pose-rectified eye image standardized size on the performance error, taking into account the distance at which the system is expected to operate; or similarly, evaluate the impact of the amount of training data, eg by using less than 15 persons, or by collecting more data to see at which level the method saturates. Such studies could be facilitated and compared to our work thanks to the use of our using publicly available database. . . .*"

Comment: *As mentioned in 3, 4 and 5, I have doubt whether it is proper to consider this method as person-invariant. Since this is the major contribution of this paper, please clearly describe the training cost for each subject. Other comments are also about the novelty of the paper, please address them.*

Response: We thank again the Reviewer for his/her reviews and comments. We hope we have properly stated why our approach is indeed person invariant and in particular further clarified the two different usage of the proposed alignment step.

In the training phase (when building person invariant models) the approach is indeed costly, but this

[2] To make experiment using, lets say half of the subjects, in a correct manner would imply not to be selective on which 7 of the other 15 subjects to use for evaluation of a given subject, but to use all possible combinations and then compute statistics on them. This would imply to evaluate, per subject, all possible half subset of their corresponding training subjects and for all algorithms we are considering in this paper, that is:

$$experiments = 16 \times 4 \times \binom{15}{7} = 411840, \tag{1}$$

which, assuming 1 minute of execution per experiment (which optimistically far from reality), the evaluation would take 2855 days of computation. This is a prohibitively large amounts of experiments. We could resort to sampling, but this would nevertheless be quite high. In addition, to study the the performance evolution in functions of the amount of training data, it would also be important to consider other numbers of training data, e.g.from the 5 to 15).

is not a problem as the main goal in practice is to reduce the cost at test time and more importantly, to address the situation when training from the test subject is not possible or minimal.

At test time the main element which currently require a user's input is the 3DMM fitting but, as we mentioned when addressing this point before, this is quickly done, it has to be done only once per subject and it can even be done by a third person by looking at a recording. Please consider the example application on gaze coding to put this into context. Furthermore, note that such fitting could be made fully automatic, as we now discuss in the new Section 9 ("Discussion") of the paper:

"... *The use of a personalized (3DMM fitted) face model was shown beneficial for accurate head pose tracking and consistent eye image cropping. The required manual landmarks annotation during training is in practice a simple procedure which has to be done only once per subject. Nevertheless, this step could leverage state of the art landmarks detection algorithms to make it fully automatic, such as Dantone et al (2012), Cao et al (2013) or, as used in Sec.7.4, Kazemi and Sullivan (2014). Although such detectors may also introduce noise, keeping the more consistent landmarks results could allow the fitting algorithm to be obtained automatically and online.*

..."

The alignment at test time is optional, which means that, when it is not used, it would not require any user input, and as we show in the gaze coding experiments, the results are already very satisfying. If used, then only a few samples for the test subject (3-5 in our experiments, both at EYEDIAP or the gaze coding experiments) are sufficient to improve the gaze accuracy. This has been validated by the many experiments we presented. Note that this is radically different from retraining an eye appearance to gaze regression model, which is what is done in the "person specific" case. As we included within the Discussion section, different applications could benefit from the proposed method and of this alignment step.

"... *Finally, we want to emphasize that our proposed approach could be exploited in diverse manners for many applications. It could be used with no cooperation from the user whatsoever, meaning the overall system and person invariant gaze models are used as is, for a new test subject. Alternatively, a minimal cooperation protocol could be defined to obtain the needed alignment data, either explicitly, e.g. requesting the participant to fixate at the camera for a few seconds (Oertel et al, 2014), or implicitly through an agent (e.g. a robot) persuading the subject to do such actions either by a direct request or by leveraging on gaze priors on non verbal human behaviors in a dialogue situation. In another direction, a third person could annotate higher level gaze semantics (people at gazing at known targets) as was shown in the previous Section.* ..."

Response to Reviewer 2

We thank the reviewer for his time and positive appreciation of the paper. Below, we provide our answers to your comments and to the unclear points that were raised, and what we have done to clarify the paper and take comments into account. Note that in our answer, all references (Equations, Figures) refer to the new version unless stated otherwise.

Comment: *This paper describes an approach to gaze estimation from 3D sensors. Using sensors like Kinect, the authors start with an offline step of fitting a 3D Morphable Model to the input point clouds. The 3DMM is a mean shape and a set of deformation vectors M, and each face would then be represented by a set of linear coefficients α. As far as I can tell the paper never mentions the origin of the 3DMM or whether it is possible to learn it from data. The fitting approach is an iteratively re-weighted least squares approach where the correspondences are estimated using ICP given the best fit model, and then the model parameters are estimated using LS given the correspondences. It was not made explicit how the model parameters were initialized or how λ_n (the stiffness parameter) is chosen in each iteration. Following the offline setup where a person-specific head model is generated, the system first estimates the head pose for the current frame, then the face texture is rectified to a front-facing head pose. The eye region is then extracted and aligned to have the eye center in the same location for all the users. The gaze vector is finally estimated using a supervised learning approach, and the authors present multiple options including kNN, linear regression, as well as using a multi-level HOG descriptor + SVR. Estimating the gaze direction as the head pose is also used as a baseline. The training involved presumably using a subject-specific training set, which is limiting, so the authors explore a person-invariant approach where they can use data from all or a subset of the other subjects to train the model. The paper also explores different approaches to eye alignment. The experiments evaluate many of the combination of the algorithms on the EYEDIAP dataset, showing how specific choices improve the gaze estimation error. Additional experiments are performed on the SONVB dataset to find gaze events in job interviews.*

Overall the presentation of the paper is clear, and the sections are well-ordered. The technical details are well-presented and sound, but some of the implementation details are missing. Additionally some of the design choices were never evaluated empirically. I believe the following points need to addressed:

Response: Thank you for your comments and careful review. Note that to address the particular points you have raised for discussion, in the new version of the paper, we have conducted new experiments and added several Sections (Section 6.1 -Implementation details and speed-, Section 6.2 -Head pose experiments- and Section 7.1 -Tracking results; Section 9 -Discussion and future work) that we hope should answer most of them. In addition, in the following link you can also find video results of the gaze coding problem:

https://www.youtube.com/playlist?list=PLQeF4owrybh2s5bwHW8tpxLNPC1FjANu1

Please notice the video/playlist settings are not public. Thus, we kindly ask you not to redistribute to protect the privacy of the participants[3].

In the following we will now detail our specific answer.

Comment: *1. There was no empirical evaluation of why fitting a person-invariant template (or the mean*

[3] Note that within youtube the viewers are anonymous (to the exception of the country access statistics, if one would check, which we will not)

shape μ) would not be sufficient for gaze estimation.

Response: This is a very good and interesting suggestion. In the new version of the paper, we have now added experiments which can bring clarity to this point in Section 7.1. We evaluated the head pose tracking accuracy in two different and publicly available benchmarks: the BIWI Kinect head pose dataset and the ICT 3D Head Pose dataset. The main motivation is that the face model has a direct impact on the accuracy of the head pose tracking, and we know accurate head pose tracking is an important requirement for appearance based gaze tracking

For both benchmarks, using a person specific face model (from the 3DMM fitting) leads to higher accuracy in comparison to using the mean face shape. This already is a strong point in favor of using a person specific face model. However a less obvious problem, which we depict in Fig. 7 is that, a mean shape-based tracking can lead to inconsistent eye cropping. This is due to the mismatch in shape between the model and the data causing convergence to local minima which are semantically inconsistent. This possibility is reduced when using a person specific face model, as it is shown in Fig. 7.

In the paper, the discussion on this point can be found in Section 7.1:

"*. . . Table 1 and 2 also compare the results using either the 3DMM-based personalized template or only the mean face shape. Even though using the mean shape leads to good results, using a personalized template do lead to more accurate head pose estimates.*

Note that accurate pose estimation is important for our method, as pose estimation impacts gaze estimations in two ways. First, as a direct input to the estimation of the line of sight in the 3D space (see step h) in Fig. 3). In this case, an error made in the estimation almost immediately translates into an error in gaze estimation. Secondly, in the extraction of the cropped rectified eye images, which needs to be consistent over frames (which means having the image projection of the eyeball center always at the same position) for the same person, since a displacement error will translate into gaze estimation errors[4].

To qualitatively illustrate the impact of this second point on gaze tracking, we present in Fig. 7 for a representative sequence the eye cropping resulting from using the fitted model or the mean face shape. As can be seen, since the mean shape do not fit well the given subject, the pose tracking results oscillate even for similar head poses, generating an inconsistent frame by frame cropping of the eye image. Notice in contrast the more stable results obtained when using a personalized face model.

Thus, overall, the better tracking results validate the use of a personalized template over the simple use of the mean face shape. . . ."

Comment: *2. The value of using the landmark term, which requires additional labeling for each frame.*

Response: The benefit of the landmarks term is for semantic consistency, as we have now explained in the revised version of the Manuscript (Section 3.2):

"*. . . The term E_l is similar to the E_d cost, but applies to a set of landmarks points (which form a subset L of the 3DMM vertices) whose position $\mathbf{l}_i$ is assumed to be annotated in the data. This term foster a semantic fitting of the 3DMM (eye corners, eyebrows, mouth corners, etc.) which, due to depth noise in the data, could be otherwise poorly localized. . . .*"

In the new Implementation details section (section 6.1) within the experiments section of the revised

[4] And the global alignment strategy only correct a systematic displacement bias error for a person, not per-frame errors.

Manuscript, we also report how the γ value (landmarks weight) is set:

"… *The γ parameter was set as $0.5\frac{N_v}{Card(L)}$, that is such that the landmarks term has 0.5 the cost of the data term, taking into account the number of data points-landmarks ratio.* …"

Notice as well that in this same new Implementation details section, we emphasize the fact that this has little practical cost per subject:

"… *Given a few annotated frames with landmarks (typically 1 to 5 frames), the fitting algorithm takes from 5 to 20 seconds to optimize. Note that since people face shape is not expected to change much, this step is only performed once per subject, which means that the fitted model can be reused across sessions.* …"

In future work, these landmarks could probably be obtained automatically with state-of-the-art landmark detection methods. This possibility has been included within the new Discussion section (section 9) as follows:

"… *The use of a personalized (3DMM fitted) face model was shown beneficial for accurate head pose tracking and consistent eye image cropping. The required manual landmarks annotation during training is in practice a simple procedure which has to be done only once per subject. Nevertheless, this step could leverage state of the art landmarks detection algorithms to make it fully automatic, such as Dantone et al (2012), Cao et al (2013) and Kazemi and Sullivan (2014). Although such detectors may also introduce noise, keeping the more consistent landmarks results could allow the fitting algorithm to be obtained automatically and online.* …"

Finally, as explained in the reply to your point 4, the landmarks are also valuable for initializing the optimization procedure.

Comment: *3. How the 3DMM was created, or if it is possible to learn it from data.*

Response: We very much apologize for missing this important point. The 3DMM we used is not original work of ours, it is the Basel Face Model:

P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. *A 3D Face Model for Pose and Illumination Invariant Face Recognition.* In Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments, Genova, Italy, 2009. IEEE

The missing citation is a mistake, as we thought it was already in the paper. The Basel Face Model is indeed learned from data as it is now discussed in the implementation details section:

"… *The 3DMM we used is the Basel Face Model (BFM) (Paysan et al, 2009). This model contains 53490 vertices and has 200 deformation modes. The BFM was learned from high resolution 3D scans of 200 individuals (100 male/100 female) thus it span a large variety of face shapes (neutral expression).* …"

This model is available on request for academic purposes.

Comment: *4. How the model parameters $\mathbf{X}_0$ were initialized and how the λ_n were chosen.*

Response: In Section 3.2 of the revised Manuscript, it is now explained how these values are set:

"*... The initialization ($\mathbf{X}^0 := \{\alpha^0, \mathbf{R}_1^0, \mathbf{t}_1^0, \ldots, \mathbf{R}_J^0, \mathbf{t}_J^0\}$) is given by the mean face shape ($\alpha^0 = \mathbf{0}$) and its -per image sample j- rigid transformation $(\mathbf{R}_1^0, \mathbf{t}_1^j)$ parameters minimizing the landmarks term E_l^j alone assuming $\alpha = \alpha^0 = \mathbf{0}$, i.e. the rigid transform that best fit the mean face shape to the annotated landmarks. ...*"

With respect to how the λ_n were chosen, this is now explained within the implementation details section 7.1:

"*... The λ_0 value was set empirically, such that its initial value is high enough to keep the α parameters close to $\mathbf{0}$ ($\lambda_0 = 0.1$ in our implementation) then $\lambda_n = 0.5\lambda_{n-1}$ within the iterative process. ...*"

Comment: *This is a well-written paper with generally sufficient experiments, but I believe the authors should fill in the missing technical and implementation details, which would also make the work easier to reproduce. I recommend a minor revision.*

Typos:
alignement -> alignment (all over)
trainign -> training (line 15, sec 5.1)
Overal -> Overall (line 24, page 14)
Set-up -> Setup (line 52, sec 8)

Response: Thank you again for your review and comments. We hope we have well addressed your concerns and we have taken care of the typos you pointed out.

Response to Reviewer 3

We thank the reviewer for his time and detailed read of the paper. Below, we provide our answers to his comments and to the unclear points that were raised, and what we have done to clarify the paper and take comments into account. Note that in our answer, all references (Equations, Figures) refer to the new version except stated otherwise.

Comment: *The paper presents a system for eye gaze estimation using an RGBD sensor, like the (first generation) Kinect. The approach is to build a person-specific model of the user by fitting a 3DMM to some training frames (using standard method and with the need of manual intervention), then use such person-specific mesh to track the rigid head motion of the user via standard ICP. Given the personalized template the rigid tracking, the authors can extract pose-normalized images of the eyes and employ standard regression techniques to infer the gaze from the eye appearances.*

The main contribution over previous work of the same authors is an additional alignment step, where errors in the face model building or in the rigid tracking are alleviated by learning a function which aligns the pose-normalized eye images to the ones used for training.

Experiments are carried out on the EYEDIAP database, which was collected by the same authors and made publicly available. Several methods are compared in several different settings. The results mainly confirm the thesis that trying to normalize for head pose changes increases the accuracy of the gaze estimation.

Even though the problem at hand is important and the work done is of good quality, I find this to be a system paper, with limited novelty, so I consider it to be borderline in terms of importance for a journal such IJCV.

Response: Thank you for your very careful review. We very much value the work and time you have devoted which was evident from the many insightful items you have brought to discussion.

Please note that we believe our paper has sufficient novelty and contributions, and very much deserves publication in IJCV. As you have mentioned, the problem at hand is very important. To come up with a solution is of great benefit for areas such as HRI, HHI and HCI, which have a significant need for algorithms which can sense human activity. Gaze is one of the most important human communication mechanism and most commercial systems are not suitable for the conditions required in such scenarios, such as poorer data sensing quality and minimal user cooperation.

In terms of the technical aspect, it is true that part of the methodology is a direct extension of previous and original work of ours, nevertheless there are many questions which were left unaddressed in this previous work, which needed to be explored more carefully and validated on a larger dataset.

For example, to what extend does the eye rectification method gives head pose invariance (quantitatively) or which strategy is more adequate to generate the said rectification? Another example is the regression algorithm and features used, for both person specific and person invariant models.

Please note that in the literature it has not been a practice to compare methods on the same data (perhaps until very recently, with the work of Schneider et al (2014)) and certainly not under the very challenging conditions which we address in this paper, motivated by the possible applications. We believe that this is in itself is a contribution of interest to the community.

We have also raised the awareness on the eye alignment problems across subjects, which has been rather neglected in the literature. To address this problem, we have proposed a novel alignment algorithm which brings the data from a set of subjects to an implicit cross-subject aligned frame. This algorithm builds and extend computer vision concepts through a sound modelling strategy.

Finally, we have depicted a clear example of how the proposed method can be employed in a human-human interaction scenario. This is indeed of high interest for researchers from diverse fields, as these ideas can be directly applied to other scenarios, for example, when interacting with a robot.

In the following link you can find video result examples of such a coding to allow you better judge the quality of the results:

`https://www.youtube.com/playlist?list=PLQeF4owrybh2s5bwHW8tpxLNPC1FjANu1`

Please notice the video/playlist settings are not public. Thus, we kindly ask you not to redistribute to protect the privacy of the subjects[5].

We can somewhat understand your concern of why to call it a "system" paper, nevertheless we strongly argue the contributions are more than sufficient to be published in a journal such as IJCV, and of particular relevance for the special issue on Human Activity Understanding from 2D and 3D data.

Before addressing the specific points you have brought to discussion, we would like to summarize some of the most important changes:

- we have conducted head pose tracking experiments on two benchmark dataset, to provide an idea of the accuracy of this step and motivate the use of a person specific 3DMM. These experiments are presented in Section 6.2, with results in Section 7.1.
- we have conducted experiments using the state of the art landmarks detection algorithm of Kazemi and Sullivan (CVPR 2014), which, as you suggested, is capable of handling lower resolution conditions (and it's implementation is available through Dlib). This to further validate and contrast the alignment strategy we have proposed.
- we have added a section on Implementation details and speed (section 6.1);
- we have added a Discussion section (Section 9) presenting limitations of the work as well as pointing out to extensions and ways to improve the approach.
- at several places we have updated the text to clarify some of the contributions or some of the questions by the reviewer.

Comment: *The use of the English language is mostly OK, but there are parts which are not crystal clear, several grammar mistakes, and some inconsistencies. For example, verbs often lack the ending 's' when conjugated in the third person singular, dashes are sometimes used to combine words in a strange, and often inconsistent way (e.g., the text includes both "head pose" and "head-pose", or "person-invariant" and "person invariant"). The paper needs a thorough review in this regard.*

Response: Thank you for your observation. We have revised the manuscript and we hope it is now improved with respect to the previous version.

[5] Note that within youtube the viewers are anonymous (to the exception of the country access statistics, if one would check, which we will not)

Comment: *- Page 2, lines 42-44, first column: the authors write "Depth measurements provide information about the shape of a 3D scene...". I disagree: shape information is also present in the RGB data.*

Response: We agree with the reviewer that there is indeed shape information in the RGB data. Please be assured that our intention was not to state otherwise, and that the sentence was not exclusive. In the case of designing an algorithm which uses shape information explicitly (as we do here), state-of-the-art algorithm on inferring shape from monocular RGB data can not be considered as a mature and reliable technology yet. Depth-cameras on the other hand can provide this information more reliably.

We have thus changed the manuscript (Introduction) as follows:

"*... or facial expressions recognition (Weise et al, 2011). Through depth (D) maps, these sensors provide explicit and reliable measurements of the scene's shape, as opposed to the implicit shape information embedded within the RGB data. Notice it is still difficult and costly to infer shape information from the visual domain (RGB) alone (Barron and Malik, 2013).*

Therefore, depth sensing enables the use of shape information in further processing stages. In particular, depth data has been shown to be valuable for accurate head pose estimation (Fanelli et al, 2011; Weise et al, 2011), a necessary step prior to determining the gaze direction. On the other hand, gaze itself requires ..."

Comment: *The commercial software faceshift Studio is an obvious competitor to the proposed method: their system, like the one proposed, uses consumer depth cameras (like the Kinect), trains person-specific 3D models fully automatically, and, among other things, outputs the user's eye gaze. I realize that it's not easy to make a quantitative comparison of the two methods (faceshift Studio only processes the live stream coming from the camera and does not allow loading pre-recorded data like those acquired by the authors), but I would at least expect the software to be mentioned in the related work, or a qualitative comparison shown.*

Response: Yes, the faceshift Software is a very good solution readily available in the market. Unfortunately, it is indeed very difficult for us to make a comparison, be it quantitative or qualitative. First, faceshift does not allow to input data for evaluation. More importantly, as faceshift's main goal is facial expression tracking and animation, we did not find information (on the scientific publications by the faceshift founders) about the principle used for eye/gaze tracking. We can only guess the following from the manual:

- within the training stage, the user must rotate the head such that the system is able to build a person specific face model from a sequence of RGB-D frames. This is a fully automatic and alternative process to that of using a 3DMM, but requires user collaboration.

- As it is required for the user to fixate at the sensor (established within the faceshift protocol), it seems the system undergoes a training specific to the eyeball. Nevertheless, it encourages, a further manual refinement of the position and size of the eyeball (see `http://doc.faceshift.com/studio/training/index.html`). Given that this information is manipulated explicitly, we are almost certain a geometric based method is being used, rather than an appearance based method which is the focus of our paper.

Therefore, we included faceshift within the relevant discussion within the related works as follows:

"*... or even complex shapes incorporating the eyelids (Yuille et al, 1992) or the full eye region (Moriyama* *and Cohn, 2004) could be used. As an example, the commercial system faceshift* {**footnote:** `www.faceshift.com`}*, intended for facial motion capture from consumer RGB-D sensors, makes use of this principle for eye tracking. . . .*"

We hope you understand this is the best we can do given what is in the public domain.

Comment: *- A paper which I think could find its way in the related work is Egger et a., Pose Normalization for Eye Gaze Estimation, 2014, where the authors also use a 3DMM to normalize the eye region of the face before gaze estimation, though only using 2D images.*

Response: Indeed, this is a relevant reference. We have added this reference and modified the Related Works section as follows:

"*. . . In another direction Funes Mora and Odobez (2012) proposed to use depth data to rectify the eye appearance into a canonical head viewpoint, an approach which was later used by Egger et al (2014) relying on a 3D face model fitted to the 2D image, rather than depth data. . . .*"

Comment: *- The work of Valenti and Gevers, 2012, is listed among the ones which fir an ellipse to the iris, but that is not correct.*

Response: It is true that the isophotes curvature is mostly intended to locate the iris center in Valenti's work, rather than doing an explicit fitting of the iris. Thank you for your observation, we have rephrased this part in the Related Works as follows:

"*. . . Under natural light conditions, many proposals also leverage local eye features to build geometric models of the eyes. Features such as the iris center (Valenti and Gevers, 2012), an ellipse fitted to the pupil/iris (Li et al, 2005), or even complex shapes incorporating the eyelids (Yuille et al, 1992) or the full eye region (Moriyama and Cohn, 2004) could be used. As an example, the . . .*"

Comment: *- Given today's state-of-the-art methods for 3D face tracking (e.g., faceshift), I would expect a system like the one presented here to run in real time. Instead, I am surprised that here's no mention about the speed of the implementation. Only the ALR alignment method is presented as prohibitively computationally expensive.*

Response: The reason we were not discussing it is because indeed, a system like faceshift has proven these type of methodologies can be very fast if well implemented. The ALR case was exceptionally slow (for a large training set), reason why we mentioned it in the previous version of the Manuscript. However, you are definitely right and this is valuable information which should be presented. We therefore have added the following in Section 6.1, "Implementation details and speed":

". . . **Speed.** *Overall, the gaze tracking takes around 100ms per frame. Note however that this is a research implementation, where the most time-consuming elements are the data pre-processing (3D mesh creation) and the head pose tracking. The head pose tracking is CPU based and alone takes from 20 to 100ms (depending on the amount of head pose changes during consecutive frames). A careful GPU-based implementation could greatly increase the speed.*

The OpenGL based rectification takes 15ms. The gaze regression computation time depends on the used algorithm. For the particular case of using 1200 *training samples (e.g. for an experiment from Sec. 7.2) the kNN's method takes 25ms per eye, the H-SVR method takes 15ms per eye, whereas the R-SVR method takes 11ms per eye. ALR's speed is heavily dependent on the training set' size, as discussed in Sec. 7.2. ...*"

In the same section, you can also find information about the computation time for the offline steps.

Comment: *- The head pose tracking is really not the main focus of the paper, but, because the same method is used to annotate the database as well as to track the faces in order to remove the head pose, I would suggest that the authors test their 3DMM + ICP method on other publicly available databases of head poses recorded with an RGBD camera, such as the one of Fanelli et al., DAGM'11, (Biwi Kinect Head Pose Database) and of Baltruaitis et al., CVPR'12, (ICT 3D HeadPose Database).*

Response: We understand your concern. In this revision we conducted experiments on both datasets. The results can be found in Sec. 7.1 of the revised Manuscript.

As it can be seen from these experiments, the head pose tracker obtains high accuracy for the ICT 3D Head Pose Database (annotated using a flock of birds sensor) and it also has comparable results to the annotations of the BIWI head pose database ($1.61°$ of difference for 20 out of 24 sessions). Notice that the annotations for the BIWI dataset were obtained from the faceshift software, which means that we are obtaining similar accuracy.

Comment: *- I find that the words on page 6, lines 23-24, second column, are incorrect: The natural human variations in eye localization should be taken care by the 3DMM fitting; I believe that the need for introducing the proposed, person-specific, alignment transform is only due to the inaccuracy of the 3DMM fitting in the training step, or of the rigid tracking. An interesting experiment would be to artificially and separately control the errors in the 3DMM modeling (i.e., deforming the personalized template so that the eyes are translated away from the right location) and in the rigid 3D tracking and see how the proposed eye alignment methods would cope with such errors.*

Response: We agree with you that the 3DMM fitting is indeed trying to solve for the eye localization. This is fostered by the landmarks terms from *manual* annotations at training and implicitly by the 3DMM being fit to the depth data (if it indeed encode these natural variations, which in principle is the case here). The alignment is implicitly exploiting this information as it is used for initialization, when solving Eq. 9.

Nevertheless, there are two important reasons why an additional alignment is needed: i) the depth data noise level does not allow for the 3DMM to use actual eyeball information; we thus agree with your statement: "...due to the inaccuracy of the 3DMM fitting in the training step...". This is also mentioned in the paper and, more importantly, ii) we argue the eye corners, even if perfectly localized, are not sufficient to determine the eyeball center (the eyeball center, although correlated, is not the mean position between the eye corners). It was indeed one of our motivations in this work to find parameters which attempt at matching gaze-aligned images across people, rather than to go through the usage of eye corners. Note that this last point is validated through the experiments where our alignment performs better than when using eye corners obtained either through automatic methods (as we have done in this revision) and even when using *manual* eye corners annotations to obtain said alignment.
In the paper, we have rephrased the paragraph to account for your comment.

"... *In principle, due to the 3DMM fitting, this window should capture the same part of the eye for*

different users, if head pose tracking errors are not considered. However, due to the uncertainty affecting the accuracy of the 3DMM fitting, or the natural human variations in the eyeball localization, which are not perfectly correlated to the position of facial features (e.g. eye corners), this may not be the case, as illustrated in Fig. 12. ..."
The experiment you propose is indeed very interesting. However, it is not possible to control the resulting 3DMM fitting without negatively influencing the head pose tracking, making it difficult to draw conclusions based on this strategy. Notice accurate head pose tracking is a requirement for gaze estimation. To corroborate this point, please see Section 7.1, where we evaluate the accuracy of the head pose tracker while contrasting to the results obtained when using the mean face shape (instead of the more precise 3DMM fitting). Nevertheless, there is a strong link associating the obtained alignment to the 3DMM semantics, as we now describe in the manuscript (Sec. 3.4):

"... *Notice that coordinate transformations in the* $\mathbf{I}^R$ *image domain can be directly reinterpreted within the* $\mathbf{HCS}$ *domain. Therefore, the alignment transform could be seen as a transformation of the 3DMM fitted model itself, as a refinement step. Notice, if* θ *defines a translation, such refined face model would generate the same eye image to* $\mathbf{I}^C$*, in particular for the* DDM *case. ...*"
Meaning the alignment is actually correcting 3DMM fitting semantic errors. Again, from the experiments based on automatic and manual landmarks detection, it has been demonstrated our alignment strategy goes even one step further than that, as it provides more accurate results on gaze estimation than landmarks.

Comment: *- The footnote of page 7 says that the cropped eye images are 55x35 pixels in size. But when the EYEDIAP database is presented, the distance of the subjects from the camera makes the eye regions vary between only 13x9 and 19x14 pixels. Why choosing such a larger size for the normalized eye images? Did the authors up-sample the images to get the desired 55x35 images? I wonder what differences in accuracy and speed would be achieved by keeping the images smaller.*

Response: Indeed, this might be larger than needed for the EYEDIAP database. However, this was merely a conservative decision on our side, with the purpose of not losing information when rectifying the image. As this image is obtained for a canonical head viewpoint it has the property that the size in pixels of the eyeball is constant (for the same subject) no matter the original image resolution or subject distance. Therefore, 55×35 was simply a "good" size which would fit diverse conditions such as faces closer to the sensor, higher resolution sensors (eg Kinect 2), or diverse head poses without losing information, with the side consequence of being an up-sampled version of the original image (for most of the cases).

To clarify this point, we have modified the footnote as:

"... **footnote**{*Note that in all methods* $\mathbf{I}^C$ *is a gray-scale image of size* 55×35. *This is a conservative choice, since in our experiments eye image sizes almost neither go beyond* 20×15. *It should however not be harmful in principle.*} ..."

As described within the Implementation details section, the gaze estimation step is not the most costly element in terms of speed (for most conditions), thus we do not think that reducing the size this would have an important impact. In addition, the ALR and H-SVR features have a fixed size which are independent of the image size. Although this is an interesting point, conducting this experiment would be time consuming, and might not bring much in terms of the contributions. As this is an interesting point for future work, we therefore included it in the Discussions section, where you can find the following:

"*… In this direction, it could also be relevant to evaluate whether there is an impact of the pose-rectified eye image size on the performance error, taking into account the distance at which the system is expected to operate; or similarly, evaluate the impact on accuracy of the amount of training data, eg by using less than 15 persons, or by collecting more data to see at which level the method saturates. Such studies could be facilitated and compared to our work thanks to the use of our publicly available database. …*"

Comment: *- The authors convert the cropped eye images from RGB to greyscale straight away. I wonder whether there is some valuable information in the color channels which is being discarded?*

Response: We agree that there might be valuable information in the color channels. However, we have been consistent with the literature, as there are no previous works exploiting color information (to our knowledge), except from a recent work of ours with takes a completely different direction (Funes Mora and Odobez, CVPR 2014). Furthermore, it is not necessarily trivial to exploit this information (e.g. which color representation, which channel, possible saturation, fusion, etc...). Adding such feature would mean performing again the experiments for all conditions, which is still time consuming given all the cross-validation. Even though this could indeed be interesting, we believe it it not so much in line with the paper contributions, and have left this as future work.

In Section 9 (Discussion and future work) or the manuscript, the following was added:

"*… An exhaustive comparison in terms of features and regression algorithms has not been conducted in this paper, as our purpose was but to validate our contributions using the best and representative features and algorithms found in the literature, as motivated in Section 4. This leaves room for future studies evaluating whether in our framework and scenarios, other types of features such as local binary patterns, the possible exploitation of color information, or combination of features (as done by Schneider et al, 2014 for instance), could improve the results. …*"

It is important to mention that one of our objectives by releasing the EYEDIAP database is to allow direct comparisons in the future, either by us or by other researchers which would like to explore these possibilities.

Comment: *- In section 4.2, regarding ALR, the authors write that the optimization in eq 7 is "conducted for seven predefined values of epsilon". It's not really clear from the text: do the authors mean seven times *per test sample*? This does not seem to be the case for Lu et al 2011a, where a constant value is found using a leave-one-out experiment on the training data.*

Response: Quoting (Lu et al 2011): "*the optimal value of* ϵ *should be determined when the minimized* $||\mathbf{w}||_1$ *just converges to 1*", this being done per test sample. Their decision of an ϵ value obtained from leave-one-out cross validation was merely due to computational reasons. We indeed tried the approach proposed by Lu et al, but we found the solution of Eq. 7 to be the empty set $\emptyset$ in many cases when evaluated in our data/experiments (using CVXOPT for solving Eq. 7). Note that ϵ is a *strict bound* on a reconstruction error. So, in the presence of large variations due to, for example, head motions, illumination changes, or when the training data does not even correspond to the same subject, the ϵ found through cross validation might be too restrictive. Note that our data drastically differs from the conditions evaluated in Lu et al 2011, where the data is of higher resolution, from the same subject, using a chin rest, no illumination changes and carefully aligned images. Note as well that increasing ϵ to avoid the empty set situations led to accuracy degradation in pilot experiments.

We therefore decided to stick to the original claim by Lu et al and find (through grid search) the ϵ for which the optimum $||\mathbf{w}||_1$ is as close to 1 as possible, at the expense of larger computational time.

We have modified the Manuscript to make this clearer:

"... *In the above formulation, the parameter ϵ plays an important role. Lu et al, 2011a recommended to obtain it from cross validation on the training set but, in our much noisier data, which drastically differ from the well controlled conditions used by Lu et al, 2011a, the resulting value usually happened to be too restrictive at test time. We therefore resorted to the original proposition by the same authors, where the optimal value of ϵ should be determined when the minimized $\|\mathbf{w}\|_1$ is equal to 1. In practice, we evaluated this using seven predefined values of ϵ, at the cost of longer computation time.* ..."

Comment: *- In section 4.3, mHoG Features, the authors cite Schneider et al 2014 as supporting the use of mHoG. However, the paper of Schneider reports slightly higher results when mHoG features are used together with LBP. It would be good if the authors also tried LBP as a feature.*

Response: Referring to Table I from Schneider et al 2014, this difference is negligible. The reported error for mHoG is of 3.55° whereas mHoG+LBP presents 3.53° error, that is 0.02° difference. The proposed dimension reductionality method by Schneider has a negative impact for both features when used for an SVR method (see Table II); we thus didn't try this and, moreover, to be consistent with Martinez et al, 2012 and Noris et al, 2010 we kept it this way. Again, please note that in this paper we are not aiming to come up with the best feature/machine learning algorithm combination. We are not claiming that as a contribution, but we are using a set of representative algorithms to validate the actual contributions.

Comment: *- In sec 5.1, a solution is proposed to alleviate the person-generic problem, where a "small" number of frames (how many?) from the test subject are used at test time to select a subset of the subjects in the training database which most closely resemble the person being tracked. For these frames, the whole database needs to be used, so I wonder: how slow is this initialization process? Also, I assume that this "trick" is always used in the ALR experiments, but there's no mention about it after section 5.1.*

Response: This is done only for "person invariance" experiments. For the rest of the experiments (person specific), as we mentioned in Section 7.2 (footnote), we simply fixed the ALR training set to 150 samples.

This model selection strategy is indeed terribly slow and costly, taking up to 10 minutes per test sample when using around 1200 training samples. This is the reason why we proposed this strategy, such that it can be done for a very small set of the test samples (1 out of 50 samples in our experiments) and then execute ALR with the selected subjects on the rest of the data at faster speed. Nevertheless, clearly this is too limiting for real-time applications.

To clarify this point, we modified the Manuscript in Section 7.4 (before commenting about the alignment methods) as follows:

"... *Note that in particular, ALR performs poorly, and the mandatory selection process described in Sec. 5.1, which was here obtained from 1 out of 50 samples, was prohibitively slow. For these reasons, we did not evaluate the alignment techniques with ALR.* ..."

Comment: *- In Section 5.2.1, the authors quickly discard using automatic methods for localizing the eye corners and thus align the eye image, claiming that such methods would not work on their low resolution* *images. I find that there have been quite some steps forward in the localization of facial features (such as eye corners) in recent years (e.g., Cao et al, Face alignment by explicit shape regression, CVPR'12, Dantone et al., Real-time Facial Feature Detection using Conditional Regression Forests, CVPR'12, Kazemi and Sullivan, One Millisecond Face Alignment with an Ensemble of Regression Trees, CVPR'14, etc...), so I wouldn't be so categorical in rejecting an automatic feature-based approach.*

Response: Actually, we were not discarding these methods. On the contrary, the experiments on alignment parameters estimation based on *manual* eye corner annotations are intended to be representative of such methods and other variants, as these experiments would represent the best case scenario.

Nevertheless, this is indeed a very interesting point which we took even further as it would help us to address a request from the editor. Due to its good performance, handling of low resolution data and accessibility through the dlib library (`http://dlib.net/`), we now conducted experiments using the method of Kazemi and Sullivan (CVPR 2014) as an alternative strategy to infer the person specific alignment. To clarify this point and the motivation behind each strategy, we made the following modification in the manuscript (results section):

"… *Notice we evaluate four types of alignment:* "FL" *correspond to an alignment based on an automatic facial landmarks detection algorithm,* "EC" *correspond to an alignment based on manually annotated eye corners……*
… *The first tested method is* FL. *In this case we applied the facial landmarks detection method of Kazemi and Sullivan (2014) on the pose-rectified facial images, as shown in Fig. 12; although in practice this method showed good stability on this type of images, we obtained the eye corners position for over 100 frames and computed their average to account for minor variations. Considering future improvements on automatic landmarks localization algorithms, we also evaluated the* EC *case, which means that 10 to 15 eye image samples were annotated* manually *with the eye corners* {**footnote:***Doing such annotation was not so easy in practice. Given our image resolution, determining visually the location of an eye corner is difficult, thus the need for multiple annotations.*}. *In both cases this was used to register the eye images in a canonical view from the average eye corners position. …*"

Indeed, the automatic landmarks detection provided lower accuracy than the manual eye corner annotations, as expected, and now mentioned in the manuscript:

"… *Overall, the* FL *strategy brings minor improvements to the* FT *scenario; although it has a similar behavior in the* CS *case, it actually degrades the accuracy for the H-SVR method. The gain is nevertheless larger for the* EC *strategy, with a gain of* 1° *in* FT, *but surprisingly almost no gain in* CS, *except for the kNN method. …*"

Moreover, in the revised version of the manuscript, we now discuss how these methods could be valuable for the proposed methodology. This was included within the *discussion and future works* section:

"… *The use of a personalized (3DMM fitted) face model was shown beneficial for accurate head pose tracking and consistent eye image cropping. The required manual landmarks annotation at training is in practice a simple procedure which has to be done only once per subject. Nevertheless, this step could leverage state of the art landmarks detection algorithms to make it fully automatic, such as Dantone et al (2012), Cao et al (2013) or, as used in Sec. 7.4, Kazemi and Sullivan (2014). Although such detectors may also introduce noise, keeping the more consistent landmarks results could allow the fitting to be obtained automatically and online.*

Furthermore, during head pose tracking, the same facial landmarks estimates could further constrain the ICP cost function to improve the tracking accuracy, esp. in near frontal poses. Fusion algorithms *and experiments would then be needed to evaluate whether the landmark extraction is robust and precise enough and can lead to the reduction of the gaze tracking errors. …*"

Comment: *- I am a bit concerned about the number of images discarded from the database: given the average clip length of 2.75 minutes and assuming that the Kinect acquired 30fps data, the average clip should contain almost 5000 frames. When the authors indicate as 2400 the average number of frames per session retained after rejecting wrongly annotated frames, that would mean that more than 50 of the recorded frames were discarded. After reading subsection 6.2, I wonder how many of the excluded frames were actually discarded because the automatic ground truth annotation failed, and how reliable in general such annotation procedure was. In page 6 an error of 5mm is mentioned, but it's not clear how that number was computed and also whether it relates to the head pose tracking or to the target tracking. Lastly, samples were also eliminated when "the eye was not fully visible"; what do the authors mean? That when even only one eye was covered, the whole frame was rejected? In general I think the database would benefit from a more thorough analysis of the quality of the annotation process.*

Response: We understand your concern, and the text was missing some information regarding this aspect. Please note that we do not say 2400 is the number of frames obtained after removing "wrongly" annotated frames. This number also takes into account the amount of frames for which the ground truth *can not* be determined, either due to some immediate sensing issue (e.g. the ball target is not in the field of view), or due to some conservative measures to keep very high the certainty that people look at the target (eg in case of gaze shifts).

The detailed reasons are now explained in Section 6.2 (also the selection of valid frames is indeed done separately for each eye) where we have now:

"… *For each session a file provides the frames that can be considered as valid. In brief, this was obtained by automatically, semi-automatically, or manually, excluding frames to account for the following: i) it is important to note that* not all *frames in a session are annotated with* $\mathbf{v}^{\mathrm{gt}}$ *as the EYEDIAP dataset consists of non stop video recordings. There were moments in the* FT *case where the GT could not be determined, like when the ball's position could not be retrieved as it was either out of the sensor's field of view, or so close that the sensor does not output depth data (needed to determined its 3D position). ii) for the* CS *case, each time* $\mathbf{p}_{PoR}$ randomly *starts a new trajectory (i.e. a new dot is displayed on the screen), a set of frames were systematically removed from the annotation to allow sufficient time for the gaze shift and ensure that the participant is fixating at the target. iii) self-occlusion situation. In sessions with head pose variations, frames where the eye was not fully visible and occluded by the nose (as estimated from the head pose) were removed. iv) extreme gazes. Frames were removed in situations with head pose and* $\mathbf{p}_{PoR}$ *measurements, but with a* $\mathbf{v}^{\mathrm{gt}}$ *almost impossible anatomically (yaw beyond 45 degrees), making it unlikely that the person was actually gazing at the target. v) finally, manuel inspection was conducted to eliminate frames with blinks and obvious distractions (the participant is not fixating at the visual target). Note that the criteria iii and iv were applied to each eye separately, meaning that in a given frame one eye annotation can be considered as valid while the other is not.*

As a result of these validity checks, the average number of valid frames per session is around 2400. …"

When EYEDIAP was released we decided to keep it in the video format as temporal information can be very interesting to provide to other researchers, even when ground truth is not available.

The automatic elements used for annotation (such as the ball tracking) are very reliable. As explained in the EYEDIAP documentation, its position was determined by color detection followed by ICP fitting. There are no false positives, except for a few frames in one of the 94 sessions, which were not detected at the time of release. By visual comparison of the 3D rendering of the ball's 3D model (used for ICP) on the data we found it to be consistently a very good fit. The 5mm estimate is a qualitative but confident estimation (it is not possible to obtain quantitatively)

To account for this, we have clarified the origin of our estimation:

"... *As an indication, the location errors were estimated to be around* $5mm$ *on average*[6] *leading to an estimated accuracy of the gaze direction of around* 0.25^o*.* ..."

Note that, when describing the two situations in Section 6.3, we have also provided the depth uncertainties (between 1.5mm and 3mm for the *CS* and *FT* conditions) as reported by the authors of the calibration software we are using.

Comment: *- The way missing values were handled in the DDM method is to simply ignore the pixels or replace them with the average of the rest of the cropped eye region. I wonder if a better solution might be to replace the missing depth values with some approximation (of the depth) based on the neighboring regions: by doing so, the actual RGB data could still be used for the areas where depth sensing was not possible (often the case around the eyes).*

Response: We agree with you, and included this in the new Discussion and future work section:

"... *The eye image pose rectification plays a role as well in our approach, esp. when eye goes towards more profile views. The* TDM *template method would profit from a 3DMM model with a tighter fit in the eye region. Local non-rigid registration methods, or unsupervised frame matching and averaging could be used there. As the depth noise level makes this challenging, RGB information could be exploited as well. Also, depth information could help handling self occlusion by the noose. The* DDM *depth driven method could alternatively make use of depth filling methods and depth smoothing, to maximize the region with texture information in the pose rectified image, and to reduce artifacts (see Fig. 6). Note here that the* TDM *approach implicitly has this function.* ..."

Comment: *- Page 11, first column, at the bottom: the authors write about an additional selection of the test data, done by removing the frames falling outside the convex hull of the training annotations, the motivation is not clearly explained though. A mismatch is mentioned for some subjects in the CS setting, but such mismatch is not well explained nor could I find the percentage of test frames which were excluded in the CS case (this number - 5 - is only mentioned for the FT sessions, where there had been no mismatch).*

Response: This is a decision we have made to report errors that are meaning full in terms of algorithm performance. Note that this happens mainly in the *CS* setting with a single user, where we train from gaze data recorded with a static pose (not necessarily frontal), and test on data with a moving head. So, for instance, for the yaw, in the training set we may observe gaze directions (in the head coordinate system) only in the range (-5, 15), while in the test data, due to the large head poses taken by users, we typically have yaws ranging in (-30, 30). Then, making predictions on test samples belonging to the ranges (-30;-5)

[6] This is an educated estimation. Location errors for the ball target or the screen dot center is considered as 0, but we needed to add the depth uncertainties or calibration errors. For the eyeball center, we evaluated the error by comparing in a few frames the manual annotation of the eyeball center with the projection of $\mathbf{o}^{\text{WCS}}$.

or (15,30) would be difficult, and errors of these samples would overwhelm the reported errors without allowing us to really understand what is going on, and draw conclusions.

This is why we mention in the footnote that "In a given application, training data would need to be collected appropriately".

In order to further clarifiy this point, we have rephrases the paragraph as follows:

"... *During evaluation, valid frames were further filtered to exclude the test samples in which the gaze ground truth was not within the convex hull of the training data (in terms of gaze angles defined w.r.t. the* HCS*). This was motivated by the fact that in the head pose invariance tests with the screen (*CS *case), there was a mismatch between the gaze training and test data for some participants. Since it is known a priori that the regression methods in appearance based apporoaches can not handle well extrapolation to unseen data*[7]*, consider samples out of the convex-hull of the training data would introduce much (random) 'noise' in the evaluation, deflecting the reported errors to convey the actual performance achieved by the different algorithms. In the (*CS*) case, as the screen is a small object (within the larger 3D space), the training samples collected using a static head pose only cover a small region of the gaze space, whereas sessions with head pose variations induced a larger coverage as the screen region would move within this space following the head movements, causing the data mismatch.*

For head pose invariance experiments in the CS *case, this could discard up to 40% of the test samples. Nevertheless, the remaining samples are still diverse in terms of head pose and gaze direction combinations. Note that (i) in the other experiments (*FT *target, person invariance), excluded frames represented less than 5% of the test frames and (ii), in all cases, as the training and test samples are the same across different gaze regressions methods, results between methods are directly comparable.* ..."

Comment: *- I would personally separate the results for the FT and CS conditions, to produce tables which are smaller and easier to understand.*

Response: Thank you for the suggestion. It is indeed a more clear way to report the experiments, we have done the modifications leading to Tables 3 and 4.

Comment: *- I appreciate that the authors included the HP experiment, i.e., taking the head direction to be the gaze, but I believe it's enough to show its errors only once in the tables. In plots like the ones of Fig 8 and 10, reporting the HP error simply makes it very hard to see the differences among the other, more interesting methods.*

Response: We have removed the HP error from Fig. 9. We kept it for Fig. 10 as we believe it is still interesting for the reader. Nevertheless, we enhanced these plots by adjusting the vertical axis (from 40° max to 30°), allowing a better view on how does the other methods compare. Please have a look at the revised Manuscript.

Comment: *- It should be highlighted in Fig 8 that the simple kNN achieves lower errors than the other methods for large gaze angles.*

Response: That is indeed an interesting observation. The Manuscript has been updated as:

[7] In a given application, training data would need to be collected appropriately.

"… *The plots show that errors are well distributed over the large range of gaze values. Interestingly, we can note that kNN has a flatter error distribution w.r.t. head pose. In particular, it has the lowest errors at large angles, followed by R-SVR.* …"

Comment: *- Mean errors are always presented alone, but I would want to see also a measure of the variance of the errors, especially because the differences between the various methods are often very small.*

Response: The tables have been updated and the standard deviations provided.

Comment: *- Fig. 9 shows the error as a function of the number of trainings samples, but I could not find how many samples were used to compute all other errors in the paper. Also, because ALR is so expensive when the number of training data increases, a comparison of the different algorithms in terms of processing speed is needed.*

Response: The amount of samples can be inferred from the protocol used for each experiment. We have made the following modifications to make this information explicit now:

For the static pose and person specific conditions:

"… *In this section we compare the regression methods assuming the model is trained and tested for the same person and under minimal head pose variations. There are 19 sessions for the* FT *target, 14 sessions for the* CS *target. In each session, the algorithm was trained using the first -temporal- half, while the evaluation was done in the second half. This means that on average, around 1200 samples are used for training*[8] *and around 1200 are used for testing.* …"

For the head pose invariance experiments:

"… *Note that for each of the 19 (*FT*) or 14 sessions (*CS*) used in Section 7.2, there is an equivalent recording session (same person and visual target) involving head pose variations rather than a static pose. Therefore, for each of the person, we used as training set the session involving a static head pose and as evaluation set the equivalent session with head pose variations, each of them comprising 2400 valid samples on average.* …"

For the static head pose, person invariance experiments:

"… *To evaluate the person invariance case, we conducted a leave-one-person-out cross-validation on the sessions involving minimal head pose variations (*SP *data). This means that in each of the N experiments (where is N is the number of session for the* FT *or* CS *case), there are around* $2400 \times (N-1)$ *samples available for training*[9] *and around* 2400 *for testing.* …"

For pose variations and person invariance:

"… *The size of the training and test sets are the same than for the* SP-PI *case.* …"

[8] Note that for ALR, the number of training samples was limited to 150. Otherwise the test time is prohibitively large (it is 3,5secs per sample when using 150 training samples).

[9] For the SVR methods we limited the training set to 1200 samples as using the full set was prohibitively slow.

With respect to the processing speed of the other methods, we have now included this information within the new Implementation Details section. Where you can find the following:

"... *The gaze regression computation time depends on the used algorithm. For the particular case of using* 1200 *training samples (e.g. for an experiment from Sec. 7.2) the kNN's method takes 25ms per eye, the H-SVR method 15ms, whereas the R-SVR method takes 11ms per eye. ALR's speed is heavily dependent on the training set size, as discussed in Sec.7.2.* ..."

Comment: *- In the plots of Fig. 10, I am surprised by the distribution of jaw angles, which reach 50 degrees on the right side, but only -20 on the left. I would expect some explanation about this imbalance in the section describing data acquisition.*

Response: This figure is indeed asymmetric and intended to be like this. As we mentioned in the paper, at an angle of around $-10°$, $-20°$ the nose starts to occlude the eye, making it difficult (or even impossible at larger angles) to infer gaze. For positive yaw angles, this does not occur as the nose "goes away" from the eye w.r.t. the sensor, and remains visible up to large yaw angles. This is why in the caption of this figure we mention that this is the distribution for the right eye only, as the left one would have an antisymmetric behavior. To avoid confusions we have modified the Manuscript as follows:

"... ...*This is confirmed by comparing the error distributions according to the head pose, shown in Fig. 10: ALR errors are higher in the* DDM *case than in the* TDM *for a head yaw angle near to -10 or -20°. Notice at head yaw angles further that -20°corresponding to the right eye getting more and more occluded by the nose, whereas at positive and larger angles (up to* $\approx 50°$*) the right eye remains visible.* ..."

Comment: *- Looking at Fig. 6, there are some small artifacts also for two of the images produced by the TDM rectification. Why is that?*

Response: The *TDM* strategy captures the texture by projecting every mesh point into the RGB image, and thus obtaining its 2D coordinates. At certain head poses, self occlusion can occur, for example one side of the nose occluding the other side. A computationally efficient way to detect this (not 100% accurate) that we experimented on was to monitor when the surface normal vector (per vertex) goes away from the camera. When this occurred we treated the points as occluded and thus become equivalent to the missing data in the *DDM* approach. The Basel Face Model (which we used in the paper) is however detailed enough to cause a few points around the eyelids or eye corners to match this criteria, even when close to frontal head poses. This is what produced this minor "artifact" in Fig.6. In practice this effect is very negligible and we prefer not to mention it in the paper.

Comment: *- Page 14, lines 17-18, first column: the table number should be mentioned, and Table 1 only has 13 columns, not 14. Here the authors should also explain the EC, A, and A5 acronyms, whose meaning is only mentioned in the caption of Table 1 before this point.*

Response: Thank you for the suggestions, we have modified the Manuscript as follows:

"... *The results for* FT *and* CS *are reported under the* SP-PI *columns from Table 3 and 4 respectively,* *and differ on which alignment strategy was used, if any. The obtained results can be compared to the person-specific case on the same data (*SP-PS*)*[10].

*Notice we evaluate four types of alignment: "*FL*" correspond to an alignment based on an automatic facial landmarks detection algorithm, "*EC*" correspond to an alignment based on manually annotated eye corners, "*A*" is our proposed synchronized delaunay implicit parametric alignment whereas "*A5*" is the same approach but using only 5 samples for the alignment of the* test *subject.* NA *correspond to no alignment. . . .*"

Later, we further explain the eye corners based alignments as follows:

"*. . . The first tested method is* FL. *In this case we applied the facial landmarks detection method of Kazemi and Sullivan (2014) on the pose-rectified facial images, as shown in Fig. 12; although in practice this method showed good stability on this type of images, we obtained the eye corners position for over 100 frames and computed their average to account for minor variations, Considering future improvements on automatic landmarks localization algorithms, we also evaluated the* EC *case, which means that 10 to 15 eye image samples were annotated* manually *with the eye corners*[11]*. In both cases this was used to register the eye images in a canonical view from the average eye corners position. . . .*"

Comment: *- In Table 1, a difference between A and A5 should be an increased processing speed, so a comparison in such terms is really needed to justify the proposal of the A5 method.*

Response: The motivation to compare *A* and *A5* is not to reduce processing speed since once the eye alignment translation is estimated, both methods have the same processing speed at test time (the same image warping step is applied).

Indeed the motivation is to evaluate whether using few annotated samples in the alignment step can still lead to performance improvement, as compared to not having any alignment for a test subject (both the EYEDIAP and gaze coding experiments show it is the case). Or more precisely concerning your comment, comparing *A* and *A5* evaluates how much we loose when using only 5 samples for alignment. This is important in practice, as the need for less annotated samples sets lower requirements in a given application like the gaze coding in natural dyadic interactions we present.

To further clarify this we modified the Manuscript as follows:

"*. . . "*A5*" is the same approach but using only 5 samples for the alignment of the* test *subject. . . .*

Notice the A *vs.* A5 *comparison is motivated by possible applications where we want to reduce the load for annotation. In this context "*A*" can be interpreted as the best case scenario whereas "*A5*" is representative of a conservative scenario where only a few gaze annotated samples can be obtained for the test subject. Please see the gaze coding experiments in Section 8 for an example. . . .*"

Comment: *Smaller comments/errors:*

Eq 4., the target point u_i is never explained.

Page 9, line 29, second column: why use the word "simplex", when triangle is appropriate (and easier to understand) for the 2D case at hand?

[10] With the slight difference that the evaluation is conducted on all samples of the test subject's session, instead of only the second half in the person-specific case.

[11] Doing such annotation was not so easy in practice. Given our image resolution, determining visually the location of an eye corner is difficult, thus the need for multiple annotations.

*Page 8, line 55, I would not include the case when there is *little* training data available as part of a "person-invariant problem".*

Fig 14 should be better explained in the text.

Section 7.2, gaze accuracy, table 2 is mentioned when table 1 is actually meant.

Response: Thank you for the observations, we have taken care of each one of these comments in the revised Manuscript.