Highly Accurate Gaze Estimation Using a Consumer RGB-D Sensor

Reza Shoja Ghiass[†] and **Ognjen Arandjelović**[‡]
[†] Université Laval, Québec (QC) G1V 0A6, Canada
[‡] University of St Andrews, St Andrews KY16 9SX, United Kingdom

Abstract

Determining the direction in which a person is looking is an important problem in a wide range of HCI applications. In this paper we describe a highly accurate algorithm that performs gaze estimation using an affordable and widely available device such as Kinect. The method we propose starts by performing accurate head pose estimation achieved by fitting a person specific morphable model of the face to depth data. The ordinarily competing requirements of high accuracy and high speed are met concurrently by formulating the fitting objective function as a combination of terms which excel either in accurate or fast fitting, and then by adaptively adjusting their relative contributions throughout fitting. Following pose estimation, pose normalization is done by re-rendering the fitted model as a frontal face. Finally gaze estimates are obtained through regression from the appearance of the eyes in synthetic, normalized images. Using EYEDIAP, the standard public dataset for the evaluation of gaze estimation algorithms from RGB-D data, we demonstrate that our method greatly outperforms the state of the art.

1 Introduction

The need to know the direction of gaze of a person is a challenge encountered in many human centred computer applications. It is of pervasive interest in marketing [Horsley *et al.*, 2014], in human-computer interaction [Yuan *et al.*, 2011], gaming [Corcoran *et al.*, 2012], psychological research [Ba and Odobez, 2009], face recognition [Arandjelović, 2012; Ghiass *et al.*, 2013; Arandjelović *et al.*, 2010], and many others. Therefore it is unsurprising that the problem of inferring gaze is a popular and well established research topic in computer vision which continues to challenge the state of the art [Hansen and Ji, 2010].

Most of the published methods on gaze estimation precede the emergence of cheap and readily available depth sensors such as those addressed in the present paper. Therefore these, which we shall for the sake of brevity refer to as 'conventional' approaches, rely purely on visual (in general colour or more often simply pixel intensity) information. Amongst these conventional approaches two broad classes of approaches can be recognized: (i) model based, and (ii) learning based. The former group of methods uses an explicit 3D model of the eye to estimate gaze direction. Almost invariably methods of this group require calibration which is a significant practical limitation, in that it is cumbersome and tedious to the user. Recent and notable methods of this group include [Yamazoe *et al.*, 2008; Yang *et al.*, 2012; Taba, 2012; Sigut and Sidha, 2011; Hung and Yin, 2010; Model and Eizenman, 2010; Nagamatsu *et al.*, 2010]. For further detail the reader is directed to a recent comprehensive survey [Hansen and Ji, 2010].

The second major group of conventional approaches adopts a more explicit, learning based model. Generally speaking algorithms of this type attempt to learn the mapping from the space of eye appearance images to the space of screen gaze points or gaze directions. Similar in their general approach, methods of this type exhibit differences in terms of eye appearance representation and the specific statistical models employed to learn the aforementioned mapping. Notable methods include [Tan *et al.*, 2002; Sheela and Vijaya, 2011; Orozco *et al.*, 2009; Lu *et al.*, 2011a; Sugano *et al.*, 2012; Coutinho and Morimoto, 2012].

Notwithstanding this continued major research effort, practical gaze estimation remains a significant research challenge. In particular, the competing requirements of usability, accuracy, and robustness, amongst others, have proven difficult to achieve. Recent advances in the availability and affordability of sensors of alternative modalities in the consumer market offer if not a possible solution, then certainly a major source of potential improvement in the aforementioned aspects of gaze estimation systems. In particular in the present work we are interested in inferring gaze direction using a combination of conventional RGB image data and low quality, noisy depth data provided by devices such as Microsoft Kinect. This problem has so far received little attention, save for the work by Funes Mora and Odobez [Funes Mora and Odobez, 2012] which is the current state of the art.

2 Proposed method

On the coarsest level the method we introduce in this paper comprises two stages. In the first stage the sensed depth data is used to reconstruct a person specific 3D model of the user's face, and then to create a synthetic frontal image by simulat-

ing a rigid 3D transformation of the face and its re-rendering. This normalizing step is used to constrain the subsequent learning stage which estimates the user's point of gaze by regression. The two stages of the algorithm are explained in detail next.

2.1 Pose estimation

Unlike much of the previous work we perform pose estimation explicitly, that is, the three pose parameters (Euler angles) are explicit parameters of the underlying model rather than inferred through regression. This removes the need for elaborate training which necessitates extensive and laboriously labelled data. We use the so-called 3D morphable model [Sun and Yin, 2008] which we fit directly to the sensed depth data as in [Ghiass *et al.*, 2015], that is, without the use of RGB appearance (unlike e.g. [Blanz and Vetter, 1999]). The fitting is formulated as an optimization task solved iteratively using a type of iterative closest point algorithm [Wollner and Arandjelović, 2011; Bouaziz *et al.*, 2013].

Basic fitting framework

In this first stage of the proposed method our goal is to fit a 3D morphable model [Romdhani *et al.*, 2002; Romdhani and Vetter, 2003] to the sensed ("target") point cloud $Y = \{y_1, \ldots, y_m\}$. The 3D morphable model captures shape variation through a linear combination of the principal shapes each of which is a dense triangulated mesh of vertices which correspond to identical anatomical and semantic loci across faces. A shape vector $\mathbf{s}$ which contains stacked 3D coordinates of model vertices can be written as:

$$\mathbf{s}(\theta) = \mu_s + \mathbf{S}\theta \tag{1}$$

where μ_s is the mean face shape, $\mathbf{S}$ a column matrix of shape basis vectors, and θ the set of parameters of the model i.e. the coefficients associated with the shape basis vectors [Blanz and Vetter, 1999].

Herein the task of fitting is posed as a registration problem whereby the aim is to register the "source" point cloud $X = \{x_1, \ldots, x_n\}$, generated by sampling from the 3D morphable model synthesised surface, with the target by minimizing an error function which comprises a weighted summation of three terms:

$$E_{\text{fit}} = E_{\text{match}} + \omega_1 \, E_{\text{rigid}} + \omega_2 \, E_{\text{model}}. \tag{2}$$

The first term, E_{match}, quantifies the proximity of source and target point clouds. The second term, E_{rigid}, seeks to impose rigidity of registration by penalizing point correspondences between the two clouds which correspond to non-rigid deformations. Lastly E_{model} penalizes unlikely intrinsic model parameters. The parameters ω_1 and ω_2 determine the relative contributions of the three error terms.

To increase its robustness to incorrect or misleading point cloud correspondences, our method uses a robust metric to quantify the goodness of alignment of two point clouds. We shall explain this shortly. For clarity we start with a description of the process using the simpler Euclidean distance based metric which captures the spirit of the process. In this case, the three terms in (2) can be written as follows:

$$E_{match} = E_{\text{match-fast}} + E_{\text{match-accurate}} \tag{3}$$

$$= \sum_{i=1}^{n} (\mathbf{n}_i^T (\mathbf{z}_i - C_Y(\mathbf{z}_i)))^2 + \tag{4}$$

$$\sum_{i=1}^{n} (\mathbf{z}_i - C_Y(\mathbf{z}_i))^2 \tag{5}$$

$$E_{rigid} = \sum_{i=1}^{n} \|\mathbf{z}_i - (\mathbf{R}\mathbf{x}_i + \mathbf{t})\|_2{}^2 \tag{6}$$

$$E_{model} = \sum_{i=1}^{n} \|\mathbf{z}_i - (\mathbf{P}_i\mathbf{d} + \mathbf{m}_i)\|_2{}^2. \tag{7}$$

Here $Z = \{z_1, \ldots, z_n\}$ is a deformed point cloud X (hence for all indexes i, the point z_i corresponds to x_i) which is being aligned with Y, $\mathbf{R}$ and $\mathbf{t}$ respectively the rotation matrix and the translation vector which describe the rigid transformation of the source point cloud, $\mathbf{P}_i$ and $\mathbf{m}_i$ the parts of respectively the matrix of morphable model principal components of shape and the mean shape, $C_Y(\mathbf{z}_i)$ the point in the target point cloud closest to $\mathbf{z}_i$, and $\mathbf{n}_i$ the surface normal at $C(\mathbf{z}_i)$. In words, the term E_{match} comprises two contributions: $E_{\text{match-fast}}$ and $E_{\text{match-accurate}}$. The first of these can be seen to accumulate point-to-plane errors between the point cloud Z and the surface described by Y. For reasons of efficiency this has in the past been used as a linearized version of the point-to-point error of $E_{\text{match-accurate}}$. However, we found that the inclusion of both $E_{\text{match-fast}}$ and $E_{\text{match-accurate}}$ provided the best trade-off between the two. Continuing with the term in (6), E_{rigid} penalizes large non-rigid deformations between X and Z i.e. differences between different x_i and z_i which cannot be explained by simple global rotation and translation. Lastly E_{model} can be understood as implementing the distance-from-feature-space metric [Arandjelović and Cipolla, 2006; Arandjelović, 2014; Wang *et al.*, 2012] where the feature space is spanned by the morphable model principal components of shape; the further the shape described by Z is from the best reconstruction by the morphable model, the greater the corresponding penalty is.

The optimization problem described by (2) is solved iteratively. In particular the first step is to linearize the updates to the rotation matrix using first order Taylor expansion – since the updates are by their very nature assumed small (this is particularly true in our algorithm given that the localization of facial features described in the previous sections allows us to initialize the model well) all cosine terms are approximated by 1 and all sine terms by the corresponding angles. This results in the rotation update matrix $\widetilde{\mathbf{R}}$ of the form:

$$\widetilde{\mathbf{R}} = \begin{bmatrix} 1 & a & b \\ -a & 1 & c \\ -b & -c & 1 \end{bmatrix}. \tag{8}$$

Thus the iterative process can be described by the following

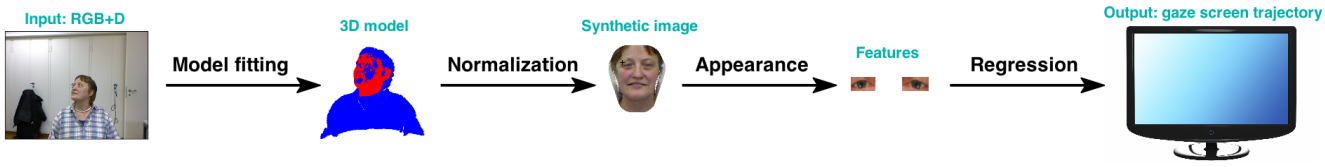

Figure 1: A schematic summary of the main steps of the proposed algorithm.

equation:

$$\arg\min_{Z^{k+1},\mathbf{d},\widetilde{R},\widetilde{t}} \sum_{i=1}^{n} \Big\{ (\mathbf{n}_i^T(\mathbf{z}_i^{k+1} - C_Y(\mathbf{z}_i)^k))^2 + (\mathbf{z}_i^{k+1} - C_Y(\mathbf{z}_i^k))^2 + \omega_1 \left\| \mathbf{z}_i^{k+1} - (\widetilde{\mathbf{R}}(\mathbf{R}\mathbf{x}_i + \mathbf{t}) + \widetilde{\mathbf{t}}) \right\|_2^2 + \omega_2 \left\| \mathbf{z}_i^{k+1} - (\mathbf{P}_i\mathbf{d} + \mathbf{m}_i) \right\|_2^2 \Big\} \quad (9)$$

where k is the iteration number which modifies each of the iteratively updated variables to denote their values in the corresponding iteration (such as Z^{k+1} for example), and $\widetilde{\mathbf{R}}$ and $\widetilde{\mathbf{t}}$ are respectively the update to the rotation matrix and the translatory adjustment between the source and the target. The iteration is initialized with $Z^0 = X$ and, as we explained before, $\mathbf{R}$ and $\mathbf{t}$ computed from the 3D loci of the detected facial features. Notice that in the computation of the closest points in Y to each $\mathbf{z}_i$, for tractability reasons it is the previous set of estimates Z^k that is being used rather than Z^{k+1} i.e. $C_Y(\mathbf{z}_i^k)$ instead of $C_Y(\mathbf{z}_i^{k+1})$.

Increasing fitting robustness

When applied on real-world data, the model fitting error function (2) formulated using the Euclidean distance metrics (4)–(7) is readily found to exhibit difficulties posed by noise and incorrect matches between two point clouds. As already noted, the sensed depth data is highly noisy so this is a major practical challenge. On the other hand the latter problem of incorrect matches may occur when there are missing point cloud data such as when a part of the sensed surface is occluded. In this case in (4) a model point (recall: obtained by sampling from the surface generated by the 3D morphable model) may be matched with a point which corresponds to an entirely different (and hence incorrect) part of the face surface. The Euclidean metric allows such misleading matches to contribute greatly to the overall error function thereby misleading the iterative process. This has negative consequences both to the accuracy of the fit as well as the efficiency of the fitting process i.e. its speed of convergence.

To dampen the effect of noisy and incorrect matches we employ a robust metric ψ to modulate the contribution of each term in the summations in (4) and (5):

$$E'_{match} = E'_{\text{match-fast}} + E'_{\text{match-accurate}} \quad (10)$$

$$= \sum_{i=1}^{n} \psi(|\mathbf{n}_i^T(\mathbf{z}_i - C_Y(\mathbf{z}_i))|)\ (\mathbf{n}_i^T(\mathbf{z}_i - C_Y(\mathbf{z}_i)))^2 + \quad (11)$$

$$\sum_{i=1}^{n} \psi(\|\mathbf{z}_i - C_Y(\mathbf{z}_i)\|_2)\ (\mathbf{z}_i - C_Y(\mathbf{z}_i))^2. \quad (12)$$

For ψ we use Tukey's biweight function [Huber and Ronchetti, 2009]:

$$\psi(d) = \begin{cases} 1 - (d/d_t)^2 & d \le d_t \\ 0 & d > d_t \end{cases}, \quad (13)$$

where d is an input distance argument such as $|\mathbf{n}_i^T(\mathbf{z}_i - C_Y(\mathbf{z}_i))|$ in (11) or $\|\mathbf{z}_i - C_Y(\mathbf{z}_i)\|_2$ in (12), and d_t the threshold which governs the breadth of the function's influence. We used $d_t = 0.01$ which corresponds to the physical distance of 0.01 m. The iterative process summarized by the expression in (9) remains unchanged so we do not repeat the equation which is understood to include the described weighting terms.

Adaptive descent

The three terms in (2) differ substantially in terms of their ability to compensate for fitting errors of different magnitudes. While the point cloud matching term (3) and the rigidity-constraining term (6) can effect fitting parameter changes which cross large spatial distances, the term (7) which corresponds to the goodness of fit of the 3D morphable model is far more spatially constrained. This is a consequence of relatively small variability of faces and in particular their shape [Craw *et al.*, 1999; Gross *et al.*, 2000]. Hence even the faces of different individuals are sufficiently similar to be registered well using a rigid transformation only.

The aforementioned observations lend a useful insight. Firstly, from the point of computational efficiency, if all of the terms in (2) are included from the very start of the fitting process, resources are unnecessarily wasted on the estimation and updating of the 3D morphable model parameters – if the rigid registration parameters are too far from their optimum values, the tuning of the model which describes intricate inter-personal differences is not done with sufficiently good data. Secondly, the misguided adjustments of the 3D morphable model parameters which occur in the early stages of the fitting process can accumulate and make the final fitting stages (when only fine refinements of the rigid registration may be needed) excessively slow and possibly produce a result of lower accuracy than one which would be achieved if no prior adjustments had been made.

Guided by the analysis above we implement an error term re-weighting scheme which at the same time achieves fast and robust convergence, and accurate fitting. In particular we start the fitting process with the model error term (7) entirely suppressed by setting ω_2 to 0 – we shall denote this initial weight as $\omega_2^{(0)} = 0$. When the error function (2) reaches a local minimum we declare that the first stage of the process is complete and that the generic shape model is approximately registered with the target. At this point the value of ω_1 is gradually reduced with a concurrent increase of ω_2 from its initial value of $\omega_2^{(0)} = 0$. The changes to ω_1 and ω_2 stop when ω_1 reaches a preset minimum value ω_1^{∞} and when ω_2 reaches its preset maximum ω_2^{∞}. The fitting process itself continues until convergence.

2.2 Gaze from synthetically generated images

After the fitting of a morphable 3D shape model to the sensed depth data, both the inherent (i.e. person specific) and relative (i.e. pose specific) geometric configurations are known and accessible explicitly. Moreover the complementary RGB information can be readily used to associate a texture map with the inferred face shape. We use these observations to render a synthetic frontal face image by normalizing the head pose i.e. by simulating a rigid transformation of the head which places it in front of and facing the camera. This is the second step in the pipeline shown in Figure 1.

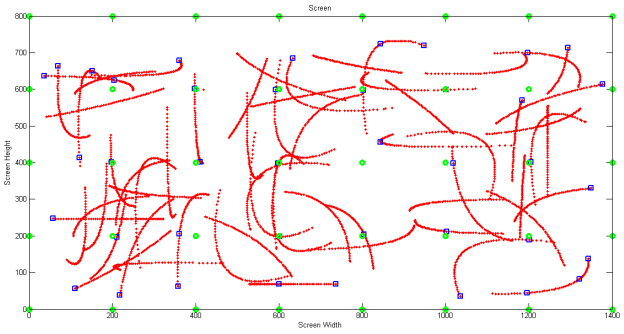

Figure 2: A sparse set of training points of gaze and the corresponding eye appearance images are chosen from the original gazing tracks (red) by sampling the screen space uniformly (green points), and using only those points (blue) from the original gaze tracks closest (in the Euclidean distance sense) to the sparse samples.

Hereafter the steps performed by our algorithm follow under the umbrella of appearance based estimation methods discussed in Section 1. However at this stage it is important to highlight a few important distinctions. Firstly having obtained a highly accurate 3D model of the face our algorithm does not need to inter the locations of the eyes. Rather these are known *a priori* seeing that the 3D morphable model comprises mixing shapes which are semantically mutually co-registered and annotated. This allows us to extract eye appearance images with high accuracy and effectively perfect reliability.

The second important effect of the preceding pose normalization step is that the learning space is greatly reduced. In particular, while the existing 'conventional' appearance based gaze estimation algorithms have to learn the mapping from eye appearance to gaze direction space over a broad range of different head poses and relative pupil loci, having normalized pose using photorealistic 3D rendering our learning is constrained to learning pupil movement only. This makes the learning process both inherently easier and in terms of practical demands reduces the amount of data required to perform the aforesaid learning.

2.3 Feature extraction

Thus this second stage of our algorithm begins by extracting greyscale eye appearance patches. As mentioned earlier this is readily achieved because the locations of the eyes are explicitly given by our 3D morphable model. To reduce the dimensionality of the representation we downsample the patches to the uniform scale of 3×5 pixels thus obtaining 15D feature vectors. Previous work suggests that this scale is sufficient for accurate gaze estimation; our results presented in the next section corroborate this finding.

2.4 Learning the appearance to gaze mapping

In principle any of a number of regression approaches can be applied at this stage. For the sake of comparison we chose to adopt two well known, widely used, and well understood methods of different complexities. These are (i) simple k-nearest neighbour (kNN) regression, and (ii) adaptive linear regression (ALR) [Lu *et al.*, 2011b]. These are summarized briefly next.

k-nearest neighbour regression As other k-nearest neighbour based algorithms [Khan and Ahmadb, 2004; Arandjelović, 2013], kNN regression is a non-parametric technique. Given an input independent variable x (in our case this is an eye appearance image), the corresponding dependent variable value y (in our case this is gaze direction) is predicted by finding the k-nearest neighbours $x_{i(1)}, \ldots, x_{i(k)}$ to x in the training set, and then by computing a weighted summation of the dependent variable values $y_{i(1)}, \ldots, y_{i(k)}$ associated with them:

$$y = \sum_{j=1}^{k} w_j y_{i(j)}, \tag{14}$$

where the values of the weight w_j is inversely proportional to the distance between x and $x_{i(j)}$:

$$w_j = \|x - x_{i(j)}\|_2^{-1}. \tag{15}$$

As per (15) we used the Euclidean distance though any of a number of alternatives, such as the Minkowski distance, could be employed just as readily.

Adaptive linear regression (ALR) Linear regression relates an input independent variable x with the corresponding output dependent variable y through a linear transformation:

$$y = Ax. \tag{16}$$

Adaptive linear regression draws from this idea and the observation that if the number of training samples is greater than the dimensionality of the independent variable, a more input specific mapping can be found by exploiting the structure of the input space. Specifically in the present case images of eyes can be considered to lie approximately on what somewhat loosely may be described as an eye manifold. Much like the better known face manifold [Lee *et al.*, 2003; Lui and Beveridge, 2008; Wang *et al.*, 2012], the eye manifold is approximately smooth and highly non-linear. Therefore, rather than learning the global projection matrix A in (16), adaptive linear regression adaptively learns this mapping for the specific sample of interest i.e. for the specific region of the independent variable space. Relevant training samples from which learning is performed are chosen on the basis of the criterion which attempts to maximize the linear representability of the input sample. Full detail on this technique can be found in [Lu *et al.*, 2011b]. Following this work we used a sparse training set, as illustrated in Figure 2.

3 Evaluation

For the evaluation of the proposed method and its comparison with the state of the art we adopted the well known EYEDIAP database [Funes Mora and Odobez, 2012] [1]. It is a freely and publicly available standard benchmark for the evaluation of algorithms for gaze estimation from RGB-D data. A detailed description of the database, and the protocol used for its acquisition and ground truth labelling can be found in the original paper [Funes Mora and Odobez, 2012]; for completeness herein we summarize the key aspects of the database of relevance to the present work.

EYEDIAP contains RGB-D sequences acquired using Microsoft Kinect at VGA resolution of 640×480 pixels and at 30 fps. The total number of individuals in the database is 16, of which 12 are male and 4 female. Each user participated in multiple acquisition sessions of 2 to 3 minutes, resulting in the total number of 94 sequences (total duration of over 4 hours) across the database. Of particular interest to the problem addressed in the present work is that the control over head motion which was imposed during data acquisition. Specifically, in some sessions the users were asked to track a screen target while keeping the head still, while in others natural head movement was not constrained. Because the location of the screen target was controlled by the experimenters and its tracking was not challenging (its motion was not excessively fast or erratic) the ground truth is considered *ipso facto* known.

3.1 Results and discussion

We start our analysis of the experimental results by comparing the performance of the proposed method with that from [Funes Mora and Odobez, 2012] on the less challenging subset of videos in the EYEDIAP database, in which as we noted previously the users kept their head stationary and effected gaze changes by means of eye movement only. A summary of the key findings can be found in Tables 1 and 2. Considering that some of the key strengths of the proposed method are effected by the highly accurate 3D face model fitting, we found it rather surprising that even though in this simpler challenge the required pose normalization was small in extent, our method already exhibited significantly superior performance than the current state of the art [Funes Mora and Odobez, 2012]. When simple kNN regression was used the average reduction in the error of the gaze direction estimate was approximately 22% (being 7.7° in comparison with 9.9°, see Table 1); even better performance (average error of 7.2°, see Table 2) and even greater reduction (nearly 30%) in the estimate error was attained with the application of the more complex adaptive linear regression.

Following the highly promising results obtained already on the simpler task of gaze estimation from video sequences in which the users' were asked to keep their heads still, we next compared our method with the state of the art on the more challenging and more practically relevant problem of estimating and tracking gaze direction when natural head movement accompanied the movement of the eyes. As previously, we summarize the key results in Tables 3 and 4. In line with our expectations both the proposed method and that of Funes Mora and Odobez performed less well in this less constrained setup. This is witnessed by the increase in the gaze direction error. However, importantly, it can be readily observed that the aforesaid error increase is rather different across the two methods. For example, looking at the results obtained using kNN regression, it can be seen that the error of the proposed method increased from the previous value of 7.7° to 8.9° i.e. for approximately 15%. In contrast, the error of the estimates achieved by the algorithm of Funes Mora and Odobez increased from 9.9° to 16.3° which corresponds to a far greater proportional error increase of approximately 65%. Comparing the average errors of the methods directly shows that the

Table 1: Gaze direction estimate errors obtained using k-nearest neighbour regression on the EYEDIAP subset of video sequences in which the users were instructed to keep their head still and alter their gaze by means of eye movement only.

	Left eye	Right eye	Mean
Proposed method	8.8°	6.5°	7.7°
Previous state of the art	10.2°	9.6°	9.9°

Table 2: Gaze direction estimate errors obtained using adaptive linear regression on the EYEDIAP subset of video sequences in which the users were instructed to keep their head still and alter their gaze by means of eye movement only.

	Left eye	Right eye	Mean
Proposed method	7.5°	6.9°	7.2°
Previous state of the art	9.7°	10.5°	10.1°

[1] The database can be downloaded from http://www.idiap.ch/dataset/eyediap.

Table 3: Gaze direction estimate errors obtained using k-nearest neighbour regression on the EYEDIAP subset of video sequences in which the users were allowed to move their head naturally while following a target displayed on the screen.

	Left eye	Right eye	Mean
Proposed method	9.0°	8.9°	8.9°
Previous state of the art	18.0°	14.6°	16.3°

Table 4: Gaze direction estimate errors obtained using adaptive linear regression on the EYEDIAP subset of video sequences in which the users were allowed to move their head naturally while following a target displayed on the screen.

	Left eye	Right eye	Mean
Proposed method	9.8°	9.5°	9.6°
Previous state of the art	15.6°	14.2°	14.9°

average error of the proposed method is nearly half that of the previous state of the art (8.9° compared with 16.3°) using simple kNN regression. Somewhat smaller but still major improvement of 36% is observed with the use of the adaptive linear regression.

4 Summary and conclusions

In this paper we described a novel algorithm for gaze direction estimation from RGB-D data acquired using an affordable, consumer market device such as Microsoft Kinect. The method we introduced comprises two key stages. In the first stage accurate head pose estimation is achieved by fitting a person specific morphable model of the face to depth data. Our approach achieves high accuracy and high speed through a carefully engineered fitting objective function which comprises a combination of terms which excel either in accurate or fast point cloud matching. The contribution of these terms is then adaptively adjusted during the iterative process of model fitting i.e. model parameter estimation. Following the fitting of an accurate 3D face model, pose normalization is done by re-rendering the model from the frontal view. In the second stage of the proposed method, appearance based eye features are extracted from the synthetic image and used to train a regressor. The proposed algorithm was evaluated on the standard benchmark database EYEDIAP on which it is shown to outperform significantly the current state of the art, reducing the error in the gaze direction estimate by more than a third.

5 Acknowledgements

The authors would like to thank Prof. Denis Laurendeau for sharing his thoughts and experience through numerous discussions related to the work described in the present paper, as well as Auto21 for their financial support.

References

[Arandjelović and Cipolla, 2006] O. Arandjelović and R. Cipolla. Face set classification using maximally probable mutual modes. *In Proc. IAPR International Conference on Pattern Recognition*, pages 511–514, 2006.

[Arandjelović *et al.*, 2010] O. Arandjelović, R. I. Hammoud, and R. Cipolla. Thermal and reflectance based personal identification methodology in challenging variable illuminations. *Pattern Recognition*, 43(5):1801–1813, 2010.

[Arandjelović, 2012] O. Arandjelović. Colour invariants under a non-linear photometric camera model and their application to face recognition from video. *Pattern Recognition*, 45(7):2499–2509, 2012.

[Arandjelović, 2013] O. Arandjelović. Discriminative k-means clustering. *In Proc. IEEE International Joint Conference on Neural Networks*, pages 2374–2380, 2013.

[Arandjelović, 2014] O. Arandjelović. Discriminative extended canonical correlation analysis for pattern set matching. *Machine Learning*, 94(3):353–370, 2014.

[Ba and Odobez, 2009] S. O. Ba and J. M. Odobez. Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, 39(1):16–33, 2009.

[Blanz and Vetter, 1999] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. *In Proc. Conference on Computer Graphics*, pages 187–194, 1999.

[Bouaziz *et al.*, 2013] S. Bouaziz, A. Tagliasacchi, and M. Pauly. Sparse iterative closest point. *Computer Graphics Forum,*, 32:113–123, 2013.

[Corcoran *et al.*, 2012] P. M. Corcoran, F. Nanu, S. Petrescu, and P. Bigioi. Real-time eye gaze tracking for gaming design and consumer electronics systems. *IEEE Transactions on Consumer Electronics*, 58(2):347–355, 2012.

[Coutinho and Morimoto, 2012] F. L. Coutinho and C. H. Morimoto. Improving head movement tolerance of cross-ratio based eye trackers. *International Journal of Computer Vision*, 2012.

[Craw *et al.*, 1999] I. Craw, N. P. Costen, T. Kato, and S. Akamatsu. How should we represent faces for automatic recognition? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:725–736, 1999.

[Funes Mora and Odobez, 2012] K. A. Funes Mora and J. Odobez. Gaze estimation from multimodal Kinect data. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 25–30, 2012.

[Ghiass *et al.*, 2013] R. S. Ghiass, O. Arandjelović, A. Bendada, and X. Maldague. Illumination-invariant face recognition from a single image across extreme pose using a dual dimension AAM ensemble in the thermal infrared spectrum. *In Proc. IEEE International Joint Conference on Neural Networks*, pages 2781–2790, 2013.

[Ghiass *et al.*, 2015] R. S. Ghiass, O. Arandjelović, and D. Laurendeau. Highly accurate and fully automatic head pose estimation from a low quality consumer-level RGB-D sensor. *In Proc. ACM Conference on Multimedia*, pages 25–34, 2015.

[Gross *et al.*, 2000] R. Gross, J. Yang, and A. Waibel. Growing Gaussian mixture models for pose invariant face recognition. *In Proc. IAPR International Conference on Pattern Recognition*, 1:1088–1091, 2000.

[Hansen and Ji, 2010] D. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):478–500, 2010.

[Horsley *et al.*, 2014] M. Horsley, M. Eliot, B. A. Knight, and R. Reilly. *Current trends in eye tracking research.* Springer, 2014.

[Huber and Ronchetti, 2009] P. J. Huber and E. M. Ronchetti. *Robust Statistics.* John Wiley & Sons,, 2nd edition, 2009.

[Hung and Yin, 2010] R. M. Hung and L. Yin. Pointing with the eyes: Gaze estimation using a static/active camera system and 3D iris disk model. *In Proc. International Conference on Multimedia and Expo*, pages 280–285, 2010.

[Khan and Ahmadb, 2004] S. S. Khan and A. Ahmadb. Cluster center initialization algorithm for k-means clustering. *Pattern Recognition Letters (PRL)*, 25(11):1293–1302, 2004.

[Lee *et al.*, 2003] K. Lee, J. Ho, M. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1:313–320, 2003.

[Lu *et al.*, 2011a] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. A head pose-free approach for appearance-based gaze estimation. *In Proc. British Machine Vision Conference*, 2011.

[Lu *et al.*, 2011b] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Inferring human gaze from appearance via adaptive linear regression. *In Proc. IEEE International Conference on Computer Vision*, pages 153–160, 2011.

[Lui and Beveridge, 2008] Y. M. Lui and J. R. Beveridge. Grassmann registration manifolds for face recognition. *In Proc. European Conference on Computer Vision*, 2:44–57, 2008.

[Model and Eizenman, 2010] D. Model and M. Eizenman. User-calibration-free remote gaze estimation system. *In Proc. Symposium on Eye Tracking Research & Applications*, pages 29–36, 2010.

[Nagamatsu *et al.*, 2010] T. Nagamatsu, R. Sugano, Y. Iwamoto, J. Kamahara, and N. Tanaka. User-calibration-free gaze tracking with estimation of the horizontal angles between the visual and the optical axes of both eyes. *In Proc. Symposium on Eye Tracking Research & Applications*, pages 251–254, 2010.

[Orozco *et al.*, 2009] J. Orozco, F. X. Roca, and J. Gonzalez. Real time gaze tracking with appearance based models. *Machine Vision and Applications*, 20(6):353–364, 2009.

[Romdhani and Vetter, 2003] S. Romdhani and T. Vetter. Efficient, robust and accurate fitting of a 3D morphable model. *In Proc. IEEE International Conference on Computer Vision*, pages 59–66, 2003.

[Romdhani *et al.*, 2002] S. Romdhani, V. Blanz, and T. Vetter. Face identification by fitting a 3D morphable model using linear shape and texture error functions. *In Proc. European Conference on Computer Vision*, pages 3–19, 2002.

[Sheela and Vijaya, 2011] S. V. Sheela and P. A. Vijaya. Mapping functions in gaze tracking. *International Journal of Computer Applications*, 3(26):36–42, 2011.

[Sigut and Sidha, 2011] J. Sigut and S. Sidha. Iris center corneal reflection method for gaze tracking using visible light. *IEEE Transactions on Biomedical Engineering*, 58(2):411–419, 2011.

[Sugano *et al.*, 2012] Y. Sugano, Y. Matsushita, and Y. Sato. Appearance-based gaze estimation using visual saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[Sun and Yin, 2008] Y. Sun and L. Yin. Automatic pose estimation of 3D facial models. *In Proc. IAPR International Conference on Pattern Recognition*, pages 1–4, 2008.

[Taba, 2012] I. Taba. Improving eye-gaze tracking accuracy through personalized calibration of a user's aspherical corneal model. Master's thesis, University of British Columbia, 2012.

[Tan *et al.*, 2002] K.-H. Tan, D. J. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. *IEEE Workshop on Applications of Computer Vision*, pages 191–195, 2002.

[Wang *et al.*, 2012] R. Wang, S. Shan, X. Chen, Q. Dai, and W. Gao. Manifold-manifold distance and its application to face recognition with image sets. *IEEE Transactions on Image Processing*, 21(10):4466–4479, 2012.

[Wollner and Arandjelović, 2011] P. Wollner and O. Arandjelović. Freehand 3D scanning in a mobile environment using video. *In Proc. IEEE International Conference on Computer Vision Workshops*, pages 445–452, 2011.

[Yamazoe *et al.*, 2008] H. Yamazoe, A. Utsum, T. Yonezawa, and S. Abe. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. *In Proc. Eye Tracking Research & Application Symposium*, 2008.

[Yang *et al.*, 2012] X.-H. Yang, J.-D. Sun, J. Liu, X.-C. Li, C.-X. Yang, and W. Liu. A remote gaze tracking system using gray-distribution-based video processing. *Journal of Biomedical Engineering: Applications, Basis & Communications*, 24(3):217–227, 2012.

[Yuan *et al.*, 2011] X. Yuan, Q. Zhao, D. Tu, and H. Shao. A novel approach to estimate gaze direction in eye gaze HCI system. *In Proc. International Conference on Intelligent Human-Machine Systems and Cybernetics*, 1:41–75, 2011.