A Comparison of Single and Multi-View IR image-based AR Glasses Pose Estimation Approaches

Ahmet Firintepe*
BMW Group Research, New Technology, Innovations
TU Kaiserslautern

Alain Pagani†
German Research Center for Artificial Intelligence (DFKI)

Didier Stricker‡
German Research Center for Artificial Intelligence (DFKI)
TU Kaiserslautern

ABSTRACT

In this paper, we present a study on single and multi-view image-based AR glasses pose estimation with two novel methods. The first approach is named GlassPose and is a VGG-based network. The second approach GlassPoseRN is based on ResNet18. We train and evaluate the two custom developed glasses pose estimation networks with one, two and three input images on the HMDPose dataset. We achieve errors as low as 0.10° and 0.90mm on average on all axes for orientation and translation. For both networks, we observe minimal improvements in position estimation with more input views.

Index Terms: Computing methodologies—Artificial intelligence—Computer vision—Tracking; Computing methodologies—Machine learning—Machine learning approaches—Neural networks

1 INTRODUCTION

After constant progress in industry and research over decades, Augmented Reality (AR) is currently on its way into our daily lives. AR applications and AR-dedicated hardware have become part of most smartphones. Further, many companies have started to commercialize AR glasses. Already available glasses like Microsoft Hololens have shown the capabilities of AR thanks to extensive usage of sensors. A variety of built-in sensors enable highly precise tracking, which is crucial for seamless and accurate display of AR content.
The deployment of AR glasses for car drivers and passengers inside the car enables a multitude of use cases enhancing the driving experience and safety, such as AR navigation in front of the eyes or information display about the car status to cite a few. In the car context, tracking based on built-in cameras inside the glasses is difficult, as the car interior and the dynamic outside world is visible. In this specific case, cameras deployed inside the car for tracking are unavoidable. Tracking AR glasses inside a car comes with the challenge to ensure the functionality in adverse and changing lighting conditions. This can be provided through infrared (IR) cameras, capturing images that are less affected by changes in lighting. Despite the mentioned advantage of IR images, little research has been conducted for IR-based object pose estimation.
RGB image-based single view object pose estimation based on Deep Learning has been the focus of many research works in Computer Vision. However, the influence of the number of input images between single-view and multi-view Deep Learning-based object pose estimation approaches has not been analyzed yet.
To address this, we conduct a thorough evaluation of single and multi-view object pose estimation approaches in IR using the HMDPose dataset [1] for our benchmark (Figure 1). This dataset provides IR images of three different views. We utilize the center image for the single view approach, the two outer images for the stereo image approach, and all three images for the triple image approach. For the evaluation, we introduce "GlassPose" and "GlassPoseRN", two novel glasses pose estimation deep neural networks. We train and evaluate both networks on three different view combinations.

Figure 1: Overview of the three different view cases compared on GlassPose and ResNet18 for AR glasses pose estimation.

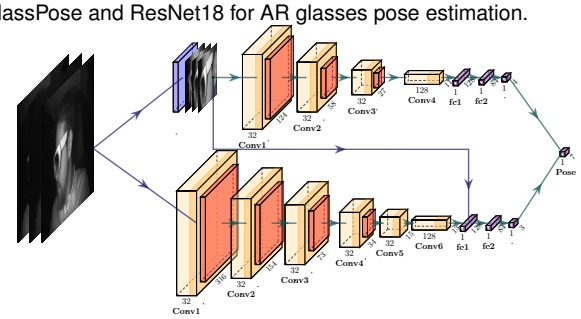

Figure 2: Our full pipeline to regress the 6DoF glasses pose. The blue square represents the automated glasses cropping. The upper part of the network performs orientation estimation, the lower part estimates the translation. The translation part contains two more Convolutional-layer, handling the higher image resolutions.

2 AR GLASSES POSE ESTIMATION

We conduct our single vs. multi-view comparison of AR glasses pose estimation on the recently published HMDPose dataset [1]. HMDPose is a large-scale data glasses dataset, consisting of around 3 million 1280×752 pixel images. It contains IR images from 3 different perspectives of 4 different AR glasses, worn by 14 subjects. Our network GlassPose (Figure 2) is inspired by the state-of-the-art of IR head pose estimation [2]. For the orientation part of the network, we first perform glasses cropping by building upon an existing DNN-based face detector [5] and automatically adjust the output to glasses area and height with a resolution of 128×54 pixel. The translation part of the network takes the full images with resolution 320×188 as input, which requires a deeper network. The orientation part of the architecture first contains two convolutional-layers with a filter size of 5×5, both followed by a max pooling layer. For the translation, we add one more layer with the same specification. For orientation, there are two more convolutional-layer with a filter size of 3×3 filter, where the first one is followed by a max pooling layer. For translation, we again add one convolutional-layer without

*e-mail: Ahmet.Firintepe@bmwgroup.com
†e-mail: Alain.Pagani@dfki.de
‡e-mail: Didier.Stricker@dfki.de

max pooling. Three fully connected layers finalize both branches of the network. During training, the first two fully connected layers use dropout as regularization (σ = 0.5). The output of both networks is then concatenated. The 2D bounding box coordinates of the orientation part are further used in the translation part to enhance translation prediction accuracy by appending the normalized bounding box coordinates to the first fully connected layer of the translation estimation part.

In addition, we benchmark GlassPoseRN, which is based on a ResNet18 [3] backbone. We add three fully connected layers with the dimensions 256, 64 and 7 to regress the orientation and translation. We use the full images with resolution 320×188 as input.

We train both networks with an Adam optimizer and the initial learning rate $\alpha = 0.0001$. Our training, validation and test split is 94/3/3. The split is based on the widely used 98/1/1 split for Big Data. We slightly adjust the split given the high framerate of 60FPS in the HMDPose dataset. We shuffle the data before splitting. We use the ReLU activation function for the hidden layers and deploy linear activation for the output layers. We utilize the Euclidean distance for translation and orientation based on the loss introduced by Kedall et al. [4]. Accordingly, our loss function is defined as follows:

$$loss := \beta \left|\left|t - \widetilde{t}\right|\right|_2 + \left|\left|q - \frac{\widetilde{q}}{||\widetilde{q}||}\right|\right|_2 \quad (1)$$

q and $\widetilde{q}$ describe the ground truth and the estimated quaternion, respectively. t and $\widetilde{t}$ are defined equally for translation. We normalize the predicted quaternion and compute the Euclidean distance to the ground truth quaternion. We regress unit quaternions on the positive w scale to obtain unambiguous estimations for the orientation. In addition, we compute the Euclidean distance for the translation. The translation is weighted accordingly through the scaling factor β to achieve a similar scaling to the orientation before being added to the orientation loss. We set this to 0.5 to achieve similar scaling levels. We train both networks until convergence. The GlassPose networks train for 180 epochs with a batch size of 128, the GlassPoseRN networks for 120 epochs with a batch size of 64.

3 Evaluation

3.1 Evaluation metrics

We define three metrics for benchmarking. The first two metrics are the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE), the latter to penalize large errors in the estimation larger. We use the Balanced Mean Angular Error (BMAE) as our third metric. It takes the unbalanced amount of different head orientations into account by defining different ranges:

$$BMAE := \frac{d}{k} \sum_{i=1} \phi_{i,i+d}, i \in d\mathbb{N} \cap [0,k], \quad (2)$$

$\phi_{i,i+d}$ defines the average angular error. For our evaluation, we set the section size d to 5 degrees and the range size k to 180 degrees. For the position error, we utilize the L_2 loss. We calculate each metric on each individual axis and on all axes combined.

3.2 Results

Table 1 shows the orientation error of our GlassPose and GlassPoseRN methods on the defined metrics on all axis individually and on average for all three view combinations. GlassPose shows results between 0.60° and 0.69° for the MAE with negligible change between the number of views. The RMSE is higher for stereo, resulting in 1.13° on average compared to 1.02° for the other view variants. Similarly, the BMAE for the stereo model is 4.29°, while being 4.03° and 3.89° for the one and three view model, respectively. GlassPoseRN performs significantly better with even less differences between the number of views. The MAE and the RMSE are 0.10° and 0.19° for all different view combinations on average, respectively. The BMAE is minimally higher on the three view variant with 0.24° against 0.18° for the other view combinations on all axes combined. Table 2 shows the positional L_2 error of the GlassPose and GlassPoseRN methods on all individual axis and in total for all three view combinations. The position estimation minimally improves by considering more than one view. For GlassPose, the overall L_2 error is 5.65mm for one view compared to 4.89mm and 5.43mm for two and three views, respectively. GlassPoseRN results in generally lower error, where the L_2 error drops from 1.09mm for one view to 0.94mm and 0.90mm for two and three views.

# Views	Metric	GlassPose Roll	Pitch	Yaw	Avg	GlassPoseRN Roll	Pitch	Yaw	Avg
1	MAE	**0.65**	0.60	**0.60**	**0.62**	0.07	**0.10**	0.14	0.10
	RMSE	**1.02**	**0.97**	1.08	**1.02**	0.10	**0.24**	0.24	0.19
	BMAE	4.64	1.76	5.71	4.03	0.20	**0.11**	**0.24**	**0.18**
2	MAE	0.69	0.60	0.61	0.63	0.07	**0.10**	**0.13**	0.10
	RMSE	1.23	0.99	1.17	1.13	0.10	**0.24**	**0.22**	0.19
	BMAE	5.40	1.92	5.55	4.29	**0.18**	0.12	**0.24**	**0.18**
3	MAE	0.67	0.60	0.61	0.63	**0.06**	0.11	0.14	0.10
	RMSE	1.06	0.98	**1.03**	**1.02**	**0.09**	0.25	**0.22**	0.19
	BMAE	**4.44**	**1.69**	**5.54**	**3.89**	0.29	0.12	0.32	0.24

Table 1: Orientation results of the GlassPose and GlassPoseRN approaches by view combinations on the given error metrics for the roll, pitch, yaw and the average in degrees. The lowest values per view for each approach are highlighted.

# Views	GlassPose x	y	z	L_2	GlassPoseRN x	y	z	L_2
1	2.67	3.22	2.56	5.65	0.66	0.54	0.42	1.09
2	**2.63**	**2.37**	**2.28**	**4.89**	0.59	**0.46**	**0.35**	0.94
3	2.85	2.73	2.54	5.43	**0.49**	0.50	**0.35**	**0.90**

Table 2: Results for the positional, Euclidean error of the GlassPose and GlassPoseRN approaches in millimeters. The lowest values per view for each approach are highlighted.

Our experiments show little difference for orientation estimation with increasing number of views. For position, we observe some decrease in error with more views. In general, we recommend using one image for a cost-efficient setup as it results in comparably similar error. If the main goal is a minimal pose estimation error where the setup cost is not the focus, more images can bring some improvements. More results on GlassPose networks trained on individual glasses are available in the supplementary material.

4 Conclusion

In this paper, we compared single and multi-view variants of two different AR glasses pose estimation methods on the HMDPose dataset. We benchmarked our custom developed CNNs GlassPose and GlassPoseRN in three different forms, estimating the pose with one, two and three input images. We achieve errors as low as 0.10° and 0.90mm on average on all axes for the orientation and translation. For both networks, we observe minimal improvements in position estimation with more input views. Future work will consist of RNN and depth-based AR glasses tracking approaches.

References

[1] A. Firintepe, A. Pagani, and D. Stricker. HMDPose: A large-scale trinocular IR Augmented Reality Glasses Pose Dataset. In *26th ACM Symposium on Virtual Reality Software and Technology*. ACM, 2020.

[2] A. Firintepe, M. Selim, A. Pagani, and D. Stricker. The More, the Merrier? A Study on In-Car IR-based Head Pose Estimation. In *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[4] A. Kendall, M. Grimes, and R. Cipolla. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2938–2946, 2015.

[5] Opencv. opencv/opencv. `https://github.com/opencv/opencv/tree/master/samples/dnn/face_detector`.