A Comparison of Single and Multi-View IR image-based AR Glasses Pose Estimation Approaches
Appendix: Evaluation on individual glasses

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

1 RESULTS

We additionally evaluate the GlassPose method regarding translation and orientation error on the defined metrics on the four different glasses of the HMDPose dataset individually. For our tables, we use acronyms for the four different AR glasses contained in the dataset. We refer to the Everysight Raptor as EVS, Hololens 1 as HOLO, North Focal Generation 1 as NORTH and the Mini Augmented Vision glasses as MAV. ALL refers to all glasses combined.

1.1 Orientation results

Table 1 shows the error of our GlassPose method on the defined metrics for the orientation on all axis individually and on average for all AR glasses models and all three view combinations.
We first evaluate results for each number of views individually. For the single view approach, the MAE and RMSE values per axis are not affected by the glasses type, leading to similar resulting errors. The MAE for the EVS and MAV glasses is between 0.47° and 0.54°. HOLO and NORTH are the biggest and smallest glasses of the dataset, where the error for the MAE is between 0.51° and 0.60° and thus slightly higher than for EVS and MAV. This also observable for RMSE. For the EVS and MAV glasses, the RMSE is between 0.71° and 0.84°, which ranges from 0.78° to 0.97° for HOLO and NORTH. In contrast to the similar value on all axes for MAE and RMSE, the BMAE on the roll is higher than on the other axes for all glasses models. It ranges from 2.46° for MAV to 3.86° for EVS. The other axes are between 1.18° for the pitch for the MAV to 2.36° for the pitch for NORTH. For ALL, the errors are generally slightly higher for all axes individually and combined. One exception to this is the BMAE, especially on the roll and yaw, which can be more than twice as high compared to the individual glasses. This is due to the challenge to learn from and estimate accurate extreme poses for different glasses models with varying appearance. The per view pattern of the errors for two and three image approaches are similar to the one image approach, resulting in higher errors for NORTH and HOLO on the MAE and RMSE. The BMAE is equally higher on the roll compared to the other axes. The estimation errors are also higher when trained on all glasses combined.
When conducting an evaluation by comparing patterns per view, we notice that all view combinations result in comparable errors with minimal differences. Still, in case of a combination of all glasses, the three view approach delivers better results on the BMAE, showing that using three images images improve the estimation performance for extreme poses.

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

		1				2				3			
Glass type	Metric	Roll	Pitch	Yaw	Avg	Roll	Pitch	Yaw	Avg	Roll	Pitch	Yaw	Avg
EVS	MAE	**0.50**	**0.49**	0.54	**0.51**	0.53	0.53	**0.50**	0.52	0.55	0.50	0.51	0.52
	RMSE	**0.80**	**0.78**	0.84	**0.81**	0.82	0.83	**0.80**	0.82	0.86	0.80	0.84	0.83
	BMAE	3.86	1.93	1.58	2.46	**3.20**	**1.72**	1.56	**2.16**	4.21	**1.72**	**1.46**	2.46
MAV	MAE	0.52	**0.49**	0.47	**0.49**	0.53	0.51	0.45	0.50	**0.51**	0.55	**0.44**	0.50
	RMSE	**0.81**	**0.71**	0.79	**0.77**	0.85	0.76	0.75	0.79	0.84	0.82	**0.72**	0.79
	BMAE	2.46	1.18	1.49	1.71	**2.20**	**1.15**	**1.35**	**1.57**	2.29	**1.43**	1.37	1.70
HOLO	MAE	**0.60**	**0.50**	**0.51**	**0.54**	**0.60**	0.52	**0.51**	**0.54**	0.63	0.53	0.52	0.56
	RMSE	0.89	**0.78**	0.89	0.86	**0.88**	0.83	**0.77**	**0.83**	0.93	0.83	0.82	0.86
	BMAE	2.61	1.96	1.49	2.02	**2.27**	**1.70**	**1.31**	**1.76**	2.51	2.27	1.51	2.10
NORTH	MAE	0.57	**0.54**	0.54	**0.55**	0.57	0.55	0.53	**0.55**	**0.56**	0.56	**0.52**	**0.55**
	RMSE	0.91	**0.97**	0.88	0.92	0.92	0.98	**0.83**	**0.91**	**0.90**	1.01	0.84	0.92
	BMAE	2.77	**2.36**	**2.09**	**2.41**	2.89	2.91	2.36	2.72	**2.66**	3.35	2.51	2.84
ALL	MAE	**0.65**	0.60	**0.60**	**0.62**	0.69	0.60	0.61	0.63	0.67	0.60	0.61	0.63
	RMSE	**1.02**	**0.97**	1.08	**1.02**	1.23	0.99	1.17	1.13	1.06	0.98	**1.03**	**1.02**
	BMAE	4.64	1.76	5.71	4.03	5.40	1.92	5.55	4.29	**4.44**	**1.69**	**5.54**	**3.89**

Table 1: Orientation results of the GlassPose approach for the three view combinations and four glasses models on the given error metrics in degrees. The Everysight Raptor is referenced as EVS, Hololens 1 as HOLO, North Focal Generation 1 as NORTH and the Mini Augmented Vision glasses as MAV. ALL stands for all glasses combined. The roll, pitch, yaw and the average of all three axis are given on the defined metrics. The lowest value on the same axis per row is highlighted.

Glass type	1				2				3			
	x	y	z	L_2	x	y	z	L_2	x	y	z	L_2
EVS	**2.14**	2.49	2.52	4.75	2.39	2.26	2.23	4.56	2.35	**2.14**	**2.22**	**4.50**
MAV	2.42	2.43	2.19	4.71	2.49	2.35	2.21	4.74	**2.11**	**2.00**	**1.95**	**4.03**
HOLO	2.45	3.02	**2.34**	5.25	2.34	**2.00**	2.47	4.55	**2.28**	2.39	2.43	**4.71**
NORTH	2.62	**2.04**	**2.12**	4.53	2.45	2.06	2.18	**4.47**	**2.29**	2.25	2.17	4.52
ALL	2.67	3.22	2.56	5.65	**2.63**	**2.37**	**2.28**	**4.89**	2.85	2.73	2.54	5.43

Table 2: Results for the positional, Euclidean error of the GlassPose approach in millimeters. The Everysight Raptor is referenced as EVS, Hololens 1 as HOLO, North Focal Generation 1 as NORTH and the Mini Augmented Vision glasses as MAV. ALL stands for all glasses combined. The lowest value on the same axis per row is highlighted.

1.2 Translation results

Table 2 shows the positional L_2 error of the GlassPose method on all individual axis and in total for all AR glasses and all three view combinations. Generally, a low Euclidean error can be observed for the position estimation, resulting in single digit millimeter errors. The different view types result in generally comparable results. An error between 1.95mm and 3.02mm is observable among all glasses and all individual axes. The L_2 error of the estimated 3D points are all in the range between 4.03mm and 5.25mm.
For the individual glasses models, a lower error of the x-axis, which represents the depth, can be observed for the three image approach. In case of combining all glasses models, the multi-view approach utilizing the left and right images of the dataset shows the lowest error on all axes and combined. We can generally see improvements in estimation performance from the single-view approach to the multi-view approaches.

2 Discussion

We observe similar results for orientation estimation among all number of views. Sufficient information contained in one cropped image might be the reason for this. NORTH and HOLO perform worse than the other two types of glasses. In case of NORTH, this is due to less information contained in the image because of the small size of the glasses. For HOLO, we see lots of reflections because of the built-in sensors, which are more visible in infrared. This has a negative effect as the glasses appear differently depending on the orientation. The estimation performance decreases when we combine all glasses, stemming from the combination of the images with different features as all glasses models in the HMDPose dataset differ in their appearance. The position error generally improves with more views, especially in the x-axis for three images, which constitutes to the depth. This shows that the network benefits from multiple views. This can be observed for the two image approach regarding the combination of all glasses. The direct view from the central image used for the three image approach decreases the estimation accuracy when all glasses are combined. The different appearances of the glasses with less positional information from the central image have a negative affect on the accuracy when combined with side view images of the glasses. Thus, the neural network trained on the left and right image performs better if all glasses are mixed.
In general, we see improvements in regards to translation when more views are added. We can observe this for orientation for extreme poses in case of combining all glasses. However, even for setups with efficiency requirements like cost and easy installation, the usage of one camera only can result in comparable results.