Expert Sample Consensus Applied To Camera Localization for AR Sports Spectators

Shazia Gul, Lewis Baker, Ryan Boult, Steven Mills and Stefanie Zollmann
Department of Computer Science
University of Otago, New Zealand
shaziagul@postgrad.otago.ac.nz

***Abstract*—Augmented Reality (AR) applications require robust and accurate localization and tracking of the user or the user's device. This is important to allow for seamless integration of 3D digital content into the real 3D environment. Robust localization is still a challenge in large-scale dynamic environments such as large sports venues. The purpose of our research is to investigate localization methods that are suitable for dynamic large-scale environments. For this purpose, we explored and evaluated the performance of state-of-the-art methods such as the 6D Camera Localization via 3D Surface Regression (DSAC++) and Expert Sample Consensus (ESAC) method.**

To investigate the feasibility of these methods, we trained DSAC++ and ESAC using a large-scale stadium image dataset captured when the stadium was empty and accessible for capture. We then used the trained systems for analyzing their robustness and accuracy for a set of different camera sequences in a stadium environment with different levels of crowdedness. Through our experiments, we found that both DSAC++ and ESAC produce acceptable results in an empty stadium. However, the experiments show that the DSAC++ localization robustness decreases for the semi-crowded dataset (75%) and completely struggles (0%) to localize the camera when the environment has a lot of dynamic elements such as a crowd. Our experiments show that ESAC trained on an empty stadium with four experts performs already better on the crowded dataset (localization success rate 87.5%) and further improves when trained with ten experts (93.75%). Our results indicate that ESAC trained on an empty environment can be used for localization in a large-scale dynamic environment.

I. Introduction

Augmented Reality (AR) describes interfaces that embed digital content into the field of view of users. On-site sports spectating has the potential to benefit from this technology as it can help to increase user engagement and spectators are already used to video footage overlaid with graphical elements from broadcast [1]. However, embedding game-related graphical elements into a live view of an on-site spectator poses additional challenges compared to embedding such content into video footage for broadcasting during a postproduction pipeline. Broadcast cameras are often calibrated with time-consuming methods [2] and often make certain assumptions about where the cameras are placed within the stadium environment. Localizing and tracking an AR capable device is a challenging task and difficult to solve because spectators can be located in almost all areas and the mobile device's position is not always static. Even though users are often seated and limited in their movements. Therefore, it is more than challenging to implement mobile AR systems in such environments.

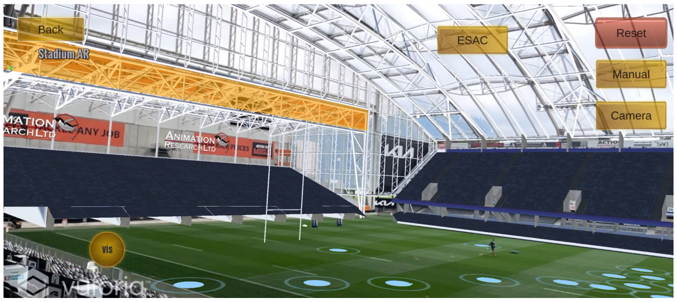

Fig. 1. Using the results of the Expert Sample Consensus for Augmented Reality visualization in a stadium overlaying a stadium model.

The main challenge is that for on-site sports spectating we cannot rely on expensive setups with pre-calibrated cameras, as we can only use the on-site spectators devices (often mobile phones). While commercial AR applications are already reliable for small-scale application scenarios such as within a living room to place the furniture[1], these solutions are often not applicaple for large-scale environments.

For dynamic large-scale environments applications remain limited with one of the key factors being the challenges around localizing as well as tracking the user. Dynamic scene elements might disturb the feature detection often used for localization and tracking, and in large crowded environments users might not be able to acquire a wide baseline as often required for state-of-the-art localization and tracking [3]. In addition, 3D data such as point clouds required for some localization methods [4] often can only be captured before an event as it requires a good coverage of the complete venue. This creates a discrepancy between the 3D model used for localization and the current appearance and structure of the stadium environment, making a successful localization more difficult.

Our research aims to address these challenges and exploring the feasibility of state-of-the-art localization methods in a stadium environment for AR sports spectating. In particular, we focus on how robust localization methods are with regards to the factor of crowdedness. Our main goal is to provide precise localization in a large dynamic stadium environment.

978-1-6654-0645-1/21/$31.00 ©2021 IEEE.

[1] IKEA Place : https://apps.apple.com/us/app/ikea-place/id1279244498

Our initial experiments identified Expert Sample Consensus (ESAC) [5] as a promising candidate for localizing images in a large-scale dynamic stadium environment. ESAC is based on a machine learning approach similar to a Mixture of Experts (MoE) [6]. MoE uses a combination of different 'experts', and each expert is specialized in a specific domain. A supplementary gating network is then responsible for deciding the relevancy of a given input and decides which experts are responsible for a given input. The final prediction is subjective based on the selected experts' outcome. In order to evaluate the performance of ESAC we performed different experiments to measure how well the ESAC method works on stadium images. The different methods were applied to a range of images, which are divided into empty, semi-dynamic and completely dynamic stadium environments, in order to evaluate their performance.

Our work aims to contribute to the foundations of realistic AR experience for large outdoor environments (such as sports stadiums) and providing a sense of real and interactive AR experiences to on-site sports sites for mobile phones and head-worn AR displays. Our contributions include demonstrating the feasibility of ESAC in dynamic large-scale sports venues using mobile phone footage from a sports spectator perspective and investigating the factor of crowdedness for AR localization. In addition, we created a dataset for providing ground truth in a large-scale stadium environment.

II. Background

Previous work identified challenges in AR research [7] [8], one of the key areas being accurate tracking and localization. In particular, for using AR on-site for live sports events these challenges increase due to the dynamic and large-scale environment. AR is a technology that aims to integrate virtual objects into the real world in real-time. An AR interface should perform seamless incorporation of virtual objects with the real environment. The virtual objects should appear to be fixed in real-world space or attached to real-world objects. This requires the continuous computation of user's location. Often this achieved by using a combination of localization and tracking [9]. Localization thereby often refers to the initial computation of the user's point of view and viewing direction in a global coordinate system. In our sports spectating use case, this is coordinate system of a 3D model of the stadium environment. Such a 3D model can then be used to place 3D content in the virtual environment for authoring using game engines such as Unity[2]. Localizing the user with respect to the world coordinate system of a 3D model allows us to then place virtual 3D content at the same position in the real world by overlaying this onto the video image (e.g. in video-see-through AR) [10]. In order to do this successfully, the localization step needs to be accurately performed. Once a global spatial relationship is established via the localization step, a continuous tracking step will be performed. While there are several solutions that have been proposed to solve localization and tracking in small scale environment [9] and indoor environments [7] [11], localization and tracking for AR in large dynamic areas such as sports is still challenging. Feigl et al. recently evaluated the performance of state-of-the-art commercial AR SDKs. Their results indicate that AR systems do not perform well in these larger environment [12]. In addition, a lot of AR solutions use fiducial registration markers, which restrict an application to a specific position or area.

Some of the early methods for localization and tacking for AR have been investigated in controlled environments or with costly infrastructure [13] [14] [15]. For instance, Livingston et al. investigated magnetic trackers [14]. These systems rely on numerous technologies and the installation of multiple cameras [15]. Ribo et al. used an approach based on 3D interaction devices for motion tracking and camera localization in a specific location with the installation of room-mounted cameras [13]. These existing methods are based on complex infrastructure and special hardware; such as multiple cameras, depth sensors, or a bespoke tracking system. Such an infrastructure is cost-intensive and depth sensors might not work at a certain distance.

As we have seen, none of the currently available methods tackle the localization challenge in a dynamic large-scale scene using AR on mobile devices. In this work, we aim to solve some of these challenges and to demonstrate an effective and reliable method for localizing camera images for providing unconstrained AR visualization on smartphones for larger dynamic event sites. The proposed methodology focuses on localization in large environments (i.e. stadiums), but also prepares for integrating this with a continuous tracking using state-of-the-art methods such as ARKit[3] or ARCore.[4]

III. Localization Methods for AR Spectating

We identified a number of methods as possible candidates for localizing users within a stadium environment by reviewing the state-of-the-art. These include ARKit, ARCore, and SLAM as well as two machine learning based approaches. ARKit and ARCore are promising as they provide an integrated tracking technology, merging information from cameras and sensors. However, they rely on fiducial image targets or sensor input such as GPS which is not always accessible in a stadium environment or might not be accurate enough. Simultaneous Localization And Mapping (SLAM) [16], which is one of the most known methods for tracking and localization in unknown environments is also not suitable to use in a large environment. In particular in the sports spectating scenario, the user is often relatively stationary while sitting and standing in a particular location. Thus SLAM [16] initialization can be challenging as it requires a certain baseline for localization [3]. Due to the above mentioned challenges, we further explored machine learning based methods; such as 6D Camera Localization via 3D Surface Regression [17] (DSAC++) and Expert Sample

[2] https://unity.com

[3] https://developer.apple.com/augmented-reality/

[4] https://developers.google.com/ar

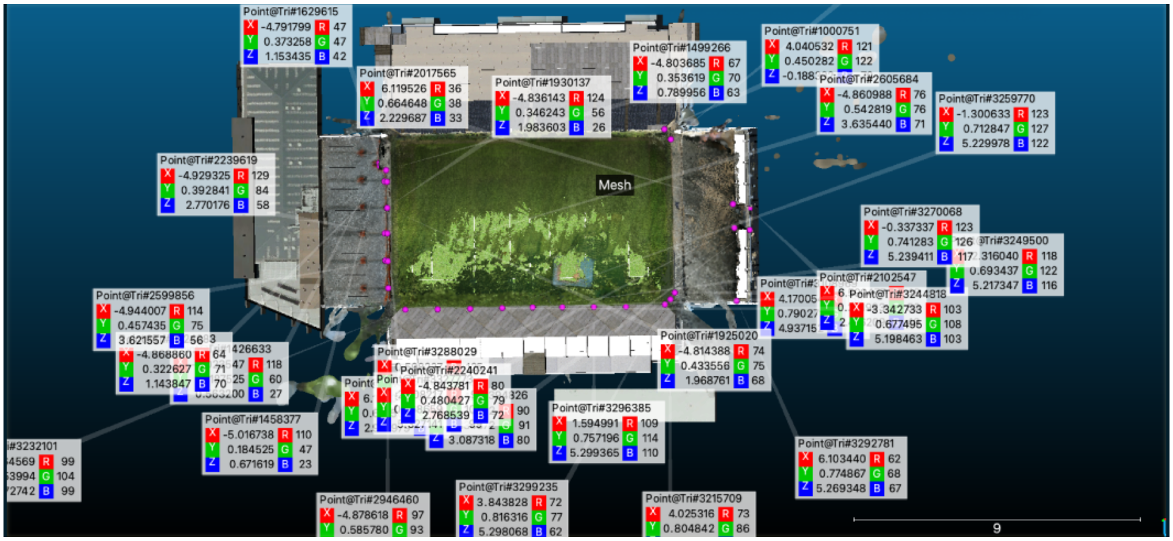
Point@Tri#1629615 X -4.791799 R 47 Y 0.373258 G 47 Z 1.153435 B 42
Point@Tri#2017565 X 6.119526 R 36 Y 0.664648 G 38 Z 2.229687 B 33
Point@Tri#1930137 X -4.836143 R 124 Y 0.346243 G 56 Z 1.983603 B 26
Point@Tri#1499266 X -4.803685 R 67 Y 0.353619 G 70 Z 0.789956 B 63
Point@Tri#1000751 X 4.040532 R 121 Y 0.450282 G 122 Z -0.188
Point@Tri#2605684 X -4.860988 R 76 Y 0.542819 G 76 Z 3.635440 B 71
Point@Tri#3259770 X -1.300633 R 123 Y 0.712847 G 127 Z 5.229978 B 122
Point@Tri#2239619 X -4.929325 R 129 Y 0.392841 G 84 Z 2.770176 B 58
Mesh
Point@Tri#3270068 X -0.337337 R 123 Y 0.741283 G 126 Z 5.239411 B 117
Tri#3249500 316040 R 118 0.693437 G 122 Z 5.217347 B 116
Point@Tri#2102547
Point@Tri#3244818 X -3.342733 R 103 Y 0.677495 G 108 Z 5.198463 B 103
Point@Tri# X 4.17005 Y 0.79027 Z 4.93715
Point@Tri#1925020 X -4.814388 R 74 Y 0.433556 G 75 Z 1.968761 B 68
Point@Tri#2599856 X -4.944007 R 114 Y 0.457435 G 75 Z 3.621557 B 56
X -4.868860 R 64 Y 0.322627 G 71 Z 1.143847 B 70
426633 47 R 118 25 G 60 B 27
Point@Tri#3288029
Point@Tri#2240241 X -4.843781 R 80 Y 0.480427 G 79 Z 2.768539 B 72
Z 3.087318 B 80
Point@Tri#3296385 X 1.594991 R 109 Y 0.757196 G 114 Z 5.299365 B 110
Point@Tri#1458377 X -5.016738 R 110 Y 0.184525 G 47 Z 0.671619 B 23
#3232101 4569 R 99 3994 G 104 2742 B 99
Point@Tri#2946460 X -4.878618 R 97 Y 0.585780 G 93
Point@Tri#3299235 X 3.843828 R 72 Y 0.816316 G 77 Z 5.298068 B 62
Point@Tri#3215709 X 4.025316 R 73 Y 0.804842 G 86
Point@Tri#3292781 X 6.103440 R 62 Y 0.774867 G 68 Z 5.269348 B 67
9

Fig. 2. Top view of the stadium structure-from-motion 3D model with the 3D reference points used for calculating the reprojection errors.

Consensus Applied to Camera Re-Localization (ESAC) [5]. DSAC++ and ESAC are scene coordinate regression methods that use neural networks for training with a precaptured environment. ESAC refers to an ensemble of experts and is trained using a 3D representation of the environment and camera images that provide pose information. During localization ESAC works on a single image input and fits an absolute 6D camera pose to it. Thereby, it uses a strategy based on a mixture of experts which decompose large scenes into a smaller scene to train the expert for each scene. ESAC distributes the model input among all the trained experts. Training ESAC with a single expert is equivalent to DSAC++ [17] and both of the methods have been shown to work successfully in large mostly static scenes (such as urban scenes showing different buildings).

In preliminary tests with our dataset of an empty stadium environment, we identified DSAC++ and ESAC as promising candidates for solving the localization problem in a stadium environment as we were able to compute camera poses with a sufficient quality (Figure 1). However, it was unclear how well these methods would perform when being trained on an empty environment but tested with dynamic environments. By answering this question, we are able to also determine the feasibility of these methods for AR sports spectating as large image datasets can only be captured before a game and suitable methods still need to perform well with a high amount of crowdedness.

IV. Experiment

We evaluate the performance of the DSAC++ and ESAC methods on our stadium dataset. The dataset contains images that have been captured in an empty stadium, in a semi-crowded stadium and a crowded stadium during a live Rugby game. In this experiment we analyzed how the camera localization methods perform under different levels of crowdedness. As it is difficult to capture ground-truth data for camera poses of mobile phone video footage, we use the reprojection error as a measurement to analyzing robustness and accuracy.

TABLE I
Average Training Time for each method.

Method	Avg Time (h)
DSAC++ (One expert)	22 hrs, 17 minute
ESAC (Four experts)	76 hrs, 9 minutes
ESAC (Ten experts)	181 hrs, 58 minutes

A. System

For all training and testing, we used a desktop computer with an Intel® Core™ i9-9900KF CPU @ 3.60GHz with 8 hyper-threaded cores, a GeForce RTX 2080 graphic card, and 32 GB of RAM. The computer runs Ubuntu version 18.04.5 LTS (64-bit). While this system setup is not suitable for mobile AR, it allows benchmarking on a large number of images. The system also runs a server application to provide pose data based on a single image input. Mobile AR applications can connect to the server to access pose data on-site, but details of this client-server system are beyond the scope of this paper.

B. Training

We train three different systems, 1) DSAC++ (one expert), 2) ESAC (four experts) and 3) ESAC (ten experts). The goal is to analyze if the number of experts has an impact on the performance for different levels of crowdedness. For training the different networks, we used the same images dataset captured in the empty stadium. The training data set was captured as part of the ARSpectator project [3], [10]. The data set was captured with Canon camera model (Canon EOS 650D), and consists of 895 stadium RGB images rescaled to a resolution of 720 × 480 pixels in different locations. The camera has been calibrated using a checkerboard and Zhang's algorithm [18]. We computed reference camera pose data using

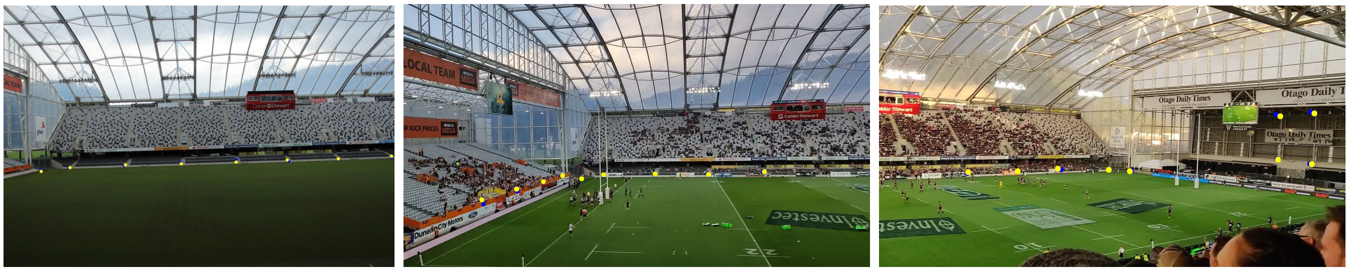

Fig. 3. Annotated sample image from the empty, semi-crowded, and crowded test dataset. Yellow points show the 2D annotations and blue points show the projected 2D points.

Colmap[5] when computing the SfM model. We use the SfM model with reference camera poses as input to training the three systems.

1) DSAC++ (One expert): [17] Firstly, we performed the end-to-end training of the rugby data-set with one expert. The gating network is trained with 80k iterations and only one expert with 160k iterations and then end-to-end training is performed using 30k iterations. The total training duration for one expert was 22 hrs, 17 minutes (Table I).

2) ESAC (Four Experts): The training of four experts is performed on using 1000k iterations for each expert and the refining process of each expert is performed for additional 1000k iterations. The gating network is trained individually with 100k iterations and the end-to-end training for 50k iterations was performed. The complete training process took 76 hrs, 9 minutes to finish (Table I).

3) ESAC (Ten Experts): We performed end-to-end training of the rugby data-set with ten experts and each expert is trained with 1000k iterations. In addition, each expert is refined with further 1000k iterations. Furthermore, we trained a gating network with 100k iterations and then we perform the trained of the ensemble of experts end-to-end for 50k iterations. The training duration for computation was 7 days, 13 hours, 58 minutes (181 hrs, 58 min) (Table I).

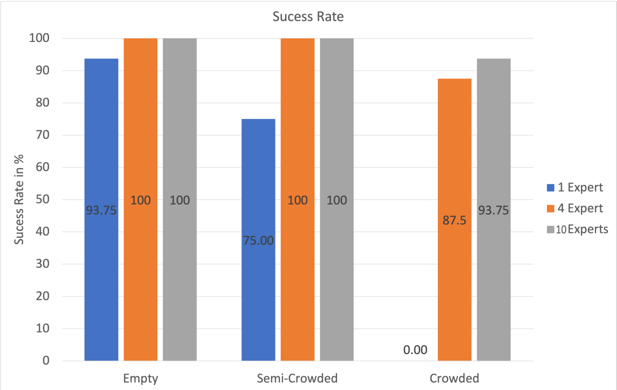

Fig. 4. Success rate in percent for the empty, semi-crowded and the crowded stadium image dataset using 1, 4 and 10 experts. Localization success is defined as having less than 1.5% reprojection error.

TABLE II
AVERAGE COMPUTATION TIME PER IMAGE FOR EACH DATASET.

	Empty Stadium	Semi-crowded	Crowded
Method	Avg Time (s)	Avg Time (s)	Avg Time (s)
DSAC++ (1 expert)	0.07	0.17	0.62
ESAC (4 experts)	0.11	0.13	0.67
ESAC (10 experts)	0.15	0.25	1.18

TABLE III
MEAN REPROJECTION ERRORS AND STANDARD DEVIATIONS FOR EACH DATASET USING ONE, FOUR AND TEN EXPERTS.

	Empty Stadium	Semi-crowded	Crowded
Method	Mean (Std)	Mean (Std)	Mean (Std)
DSAC++ (1 expert)	13.77(9.28)	22.38 (10.37)	623.34(294.87)
ESAC (4 experts)	6.57 (1.38)	10.77 (4.88)	16.92(14.45)
ESAC (10 experts)	6.33 (1.5)	9.14 (5.89)	11.63 (8.81)

We can see that the number of experts has an impact on training times. While training a system with one expert takes under a day, ten experts may take about a week, a time factor that needs to be taken into account for the preparation time applying this at different venues.

C. Testing

We analyze the accuracy of the different networks using different image datasets reflecting different levels of crowdedness: empty, semi-crowded, and a crowded stadium. We visualize all experimental results in the 3D environment (Figure 6) and compute robustness and accuracy based on the reprojection errors for each image.

1) Datasets: We captured three image datasets in different stadium environments. The first dataset captured an empty stadium similar to the images used for training the systems (Figure 3 Left). These images were captured as single photographs using the Canon EOS 650D. The second dataset was captured in video mode with a mobile phone (OnePlus6 with 1920 × 1080 pixel resolution) in a semi-crowded stadium after a rugby game. The semi-crowded dataset covers a small number of spectators in the stands and players on the field (Figure 3 Middle). The third dataset was also captured with the OnePlus6 in video mode and captured a crowded stadium during a rugby game. The mobile sequences were recorded during different times in the stadium and with different lighting conditions. Data was captured both during and after the game,

[5] https://colmap.github.io

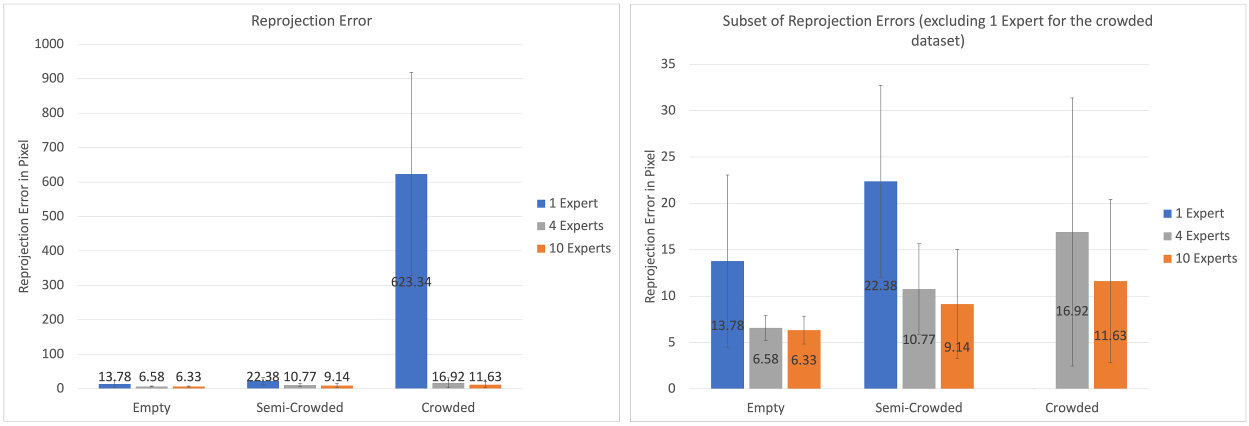

Fig. 5. Mean reprojection error in pixels for the empty, semi-crowded and the crowded stadium image dataset with one standard deviation error bars. Left: Complete results. Right: Selection of results excluding data for 1 expert for the crowded dataset for better comparison.

providing semi-crowded and crowded dynamic contents such as spectators and players in the stadium. For each of of the three datasets we manually selected 16 images with a variation of viewing angles, illumination, and appearance. All cameras were calibrated beforehand using Zhang's algorithm [18].

2) Localization: For each image in the three datasets, we use the three different localization methods to compute camera poses. DSAC++ and ESAC take a single RGB image as input and output a translation vector and quaternion of each image. We convert this into a 4x4 pose matrix that can then be used for further analysis and visualization. To visualize the results, we use a 3D SfM model of the stadium in combination with the pose matrix to overlay the 3D data onto the camera images(Figure 1).

3) Robustness and Accuracy Computation: To calculate the robustness and accuracy of the method, we compute the reprojection error for each image. In order to compute the reprojection error we need ground truth. As this is difficult to obtain for mobile phone sequences that replicate the behavior of a stationary user within a stadium, we manually create ground truth by selecting 3D reference points P_r from the SfM 3D model (SfM model) and manually selecting the corresponding 2D points p_a in each of the camera images. We then reproject the selected 3D reference points into the camera image p_r by using the camera's intrinsic calibration and computed pose information. We then calculate the reprojection error as the Euclidean distance between the reprojected 3D references points p_r and the corresponding 2D references points p_a. We also compute the localization success by taking a threshold of 1.5% reprojection error in vertical resolution. All results that show a higher reprojection error than 1.5% are considered to be not successfully localized. We compute the average reprojection error and success rate for each method individually on crowded, semi-crowded, and empty stadium images. Figure 2 shows the 3D model and 3D points picked in the different locations, and Figure 3 shows the annotated 2D points in the crowded, semi-crowded and empty test data.

V. Results

We performed testing of our datasets with DSAC++, ESAC (four experts), and ESAC (ten experts) and compare the results of the experiments. Our results are shown in Figure 6 and Table III, measuring the mean reprojection error and standard deviation for each localization method for each dataset. The results show that the robustness and accuracy of DSAC++ (One expert) decreases in the more crowded environment (mean=22.38 and success rate 75% for the semi-crowded, and a mean=623.34 pixel, success rate 0% for the crowded data). However, ESAC trained with four (mean=16.92 pixel) and ten experts (mean = 11.63) performs better in crowded stadium environments. Comparatively, the ESAC method with a large number of experts has a smaller reprojection error and higher success rate as shown in Figure 4. We also compute the average computation times of DSAC++ and ESAC (Table II) which ranges from 0.15 seconds for DSAC++ and 0.25 seconds for ESAC with four experts to 1.18 seconds for ESAC with ten experts for the crowded images. These timings can be considered as sufficient as they would still allow for relatively immediate localization.

VI. Conclusion and Future Work

In this work, we conducted feasibility testing of different state-of-the-art methods for AR localization in a large dynamic stadium environment. In particular, we trained three different systems (DSAC++ and ESAC with four and ten experts) and investigated the robustness and accuracy for different levels of crowdedness in an empty, a semi-crowded, and a crowded stadium environment on camera images and mobile phone input. For this purpose, we calculate reprojection errors and the success rate for each dataset. In addition, we provide qualitative results by visualizing the results and overlaying the 3D point cloud model onto camera images using the computed pose data (Figure 6). Comparing all three methods, we conclude that ESAC trained with ten experts has lower mean projection error and higher success rates compared to one and four experts. However, this comes with a higher

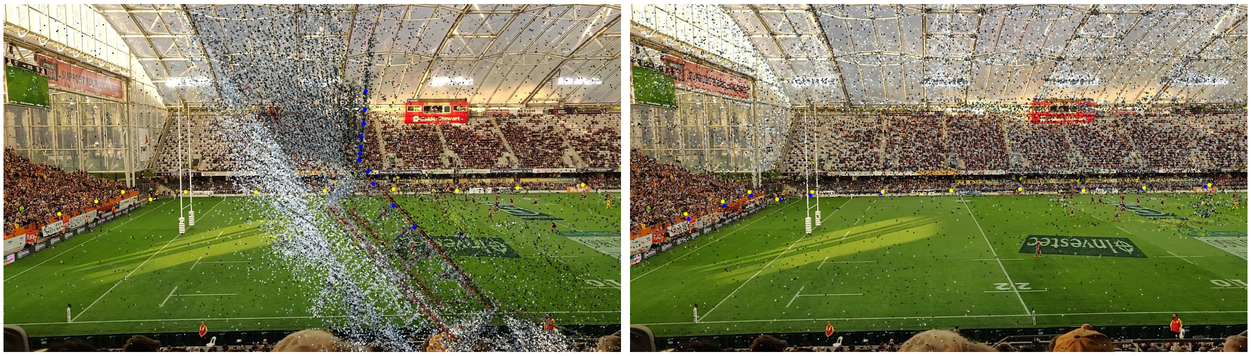

Fig. 6. Visualization of 3D point cloud overlaid onto a camera image (from the crowded dataset) using the computed camera pose. Left: Visualization using camera pose computed by one expert. This image is considered to be not correctly localized with an average reprojection error above 500pixel. Right: Visualization using camera pose computed by ten experts. This image is considered to be correctly localized with an average reprojection error of 8.9 pixel.

cost of training time as well as computation time. Training a network for over a week can be challenging, a compromise might be to use a system with four experts that has a slightly lower accuracy, but still shows an acceptable robustness. The system with four experts provides average accuracies around 17 pixels (below 1% for the vertical resolution). Depending on the content to be shown this can still be sufficient for a good AR user experience. Similarly, computation times of around 70ms for ten experts can be reduced to around 40ms by using four experts. Conclusively, DSAC++ struggles to localize in a crowded environment due to the fact that only one uses one expert for training therefore, for a large dynamic environment, it completely struggles to localize. However, on the other hand, ESAC can be trained with a large number of experts which in turn divides the large environment into several small environments therefore it works better in a large environment.

For future work, we plan to run an on-site feasibility study with the trained system. For this purpose, we are planing to fully integrate our trained networks in a client-server infrastructure. Our AR application can then directly connect and localize itself. Combining this with real-time tracking will allow us to investigate the potential of AR spectating applications more in detail.

Acknowledgements

This project is supported by an MBIE Endeavour Smart Ideas grant. We thank Animation Research Ltd, Forsyth Barr Stadium, the Highlanders, Otago Rugby (ORFU), and Opta-Perform for their support of this research. In addition, we thank Mike Denham and Craig Tidey from the School of Surveying for their support in surveying the stadium and Wei Hong Lo for support with the AR visualization.

References

[1] W. Hong Lo, S. Zollmann, and H. Regenbrecht, "Who kicked the ball? situated visualization in on-site sports spectating," in *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2021, pp. 496–497. 1

[2] T. Moeslund, G. Thomas, and A. Hilton, *Computer Vision in Sports*, 2013. 1

[3] L. Baker, J. Ventura, S. Zollmann, S. Mills, and T. Langlotz, "Splat: Spherical localization and tracking in large spaces," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2020, pp. 809–817. 1, 2, 3

[4] J. Ventura and T. Höllerer, "Wide-area scene mapping for mobile visual tracking," in *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2012, pp. 3–12. 1

[5] E. Brachmann and C. Rother, "Expert sample consensus applied to camera re-localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7525–7534. 2, 3

[6] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991. 2

[7] E. Marchand, H. Uchiyama, and F. Spindler, "Pose estimation for augmented reality: a hands-on survey," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 12, pp. 2633–2651, 2015. 2

[8] D. Yan and H. Hu, "Application of augmented reality and robotic technology in broadcasting: a survey," *Robotics*, vol. 6, no. 3, p. 18, 2017. 2

[9] G. Klein and D. Murray, "Parallel tracking and mapping on a camera phone," in *Proc. Eigth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'09)*, Orlando, October 2009. 2

[10] S. Zollmann, T. Langlotz, M. Loos, W. H. Lo, and L. Baker, "Arspectator: Exploring augmented reality for sport events," in *SIGGRAPH Asia 2019 Technical Briefs*, 2019, pp. 75–78. 2, 3

[11] J. Linowes and K. Babilinski, *Augmented Reality for Developers: Build practical augmented reality applications with Unity, ARCore, ARKit, and Vuforia*. Packt Publishing Ltd, 2017. 2

[12] T. Feigl, A. Porada, S. Steiner, C. Löffler, C. Mutschler, and M. Philippsen, "Localization limitations of arcore, arkit, and hololens in dynamic large-scale industry environments." in *VISIGRAPP (1: GRAPP)*, 2020, pp. 307–318. 2

[13] M. Ribo, A. Pinz, and A. L. Fuhrmann, "A new optical tracking system for virtual and augmented reality applications," in *IMTC 2001. Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference. Rediscovering Measurement in the Age of Informatics (Cat. No. 01CH 37188)*, vol. 3. IEEE, 2001, pp. 1932–1936. 2

[14] M. A. Livingston and A. State, "Magnetic tracker calibration for improved augmented reality registration," *Presence: Teleoperators & Virtual Environments*, vol. 6, no. 5, pp. 532–546, 1997. 2

[15] G. Welch and E. Foxlin, "Motion tracking: No silver bullet, but a respectable arsenal," *IEEE Computer graphics and Applications*, vol. 22, no. 6, pp. 24–38, 2002. 2

[16] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual slam algorithms: a survey from 2010 to 2016," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, no. 1, p. 16, 2017. 2

[17] E. Brachmann and C. Rother, "Learning less is more-6d camera localization via 3d surface regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4654–4662. 2, 3, 4

[18] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, p. 1330–1334, Nov. 2000. [Online]. Available: https://doi.org/10.1109/34.888718 3, 5