# Testing and Evaluation

## Methods of Evaluation

The workplace assistant augmented reality system was evaluated using the images of the markers at diﬀerent angles, distances, lighting and scenarios to ensure that no bias was developed towards the tested markers. The system was tested both quantitatively and qualitatively. The two components which were tested were the augmented reality and the recommendation system.

## Augmented Reality Quantitative Testing and Evaluation

Ideally, the system would be tested by evaluating the precision and recall of the traditional computer vision and deep learning techniques utilised. However, Vuforia does not allow its users full access to its API. One can also assume that since Vuforia is a commonly used commercial library, therefore it is most certainly making use of eﬃcient and accurate models for training and testing. However, as seen in [28], quantity testing on the AR application was also achieved by testing its competency in recognising and augmenting the markers used in diﬀerent scenarios, that enable the application to be tested to its limits. The evaluations used were colour variance, distance variance, rotation variance and occlusion variance. The tests were performed using a Xiaomi Mi A2 smartphone. Tests were done on the model target and image target separately. Image targets amount up to 10 whilst model targets amount up to 11 which includes the coﬀee machine. Hence, in total combining the image and model targets, one test is repeated 21 times. Vuforia also provides a 5-star rating to every image target which represents the detection quality. The higher the rating is the better the image target is detected.

### Colour Variance

For colour detection the images were converted to grayscale and each image was successfully detected via the AR application. Firstly, as presented in [15], this is because Vuforia uses natural features such as edges and contours to detect images both in the traditional computer vision techniques and the pose estimation deep learning techniques. Secondly, the images when analysed by the library are analysed in grayscale to be able to generalise rather than form a bias towards images or objects with a speciﬁc colour.

### Distance Variance

The distance variance tests the maximum range the smartphone can be away from the object or image being recognised. Each image and model target were tested and the mean variance, standard deviation, and mean standard error were calculated as shown in table 3. As observed the model targets were on average more detectable than the image targets. This further emphasises the strength of deep learning techniques in comparison to traditional computer vision approaches. The average maximum distance test also shows how stable the AR application is in recognising an object from far away. Figures 35 and 36 found in the appendix show results highlighting the diﬀerence between each image and model target, respectively.

### Orientation Variance

The maximum orientation test tests the maximum orientation angles of the image or model target the AR app can recognise. The results in Figures 37 and 38 show that the AR app can handle any type of orientation. Therefore, the application is capable of generalising between orientations since natural features were used in both modern and tradition computer vision approaches.

### Occlusion Variance

The occlusion variance tested the AR application’s capability in recognising an object while it is being occluded. This is necessary as the more occlusion the AR can handle the more applicable it is in real life scenarios; hence the application can recognise the targets while an object or a person is partially occluding them. Each image was tested by occluding them by 25%, 50%, and 75% as shown in Figure 39. In contrary to the procedure used in [28] in evaluating occlusion, a diﬀerent form of occlusion was tested. The procedure used within this project aimed to mimic real life occurrences with regards to an occluded object. As in reality the markers (the door signs in this case) may be obstructed horizontally by another person in view. Table 5 shows that the AR application was capable of recognising on average more than 50% of an obstructed object both using model and image targets. Therefore, the results show that the application is more applicable and eﬃcient to use in real life scenarios, where noise in data can be of an issue due to the environment one is using the application in. Figures 40 and 41 provide in detail the maximum occlusion an object could take to be detected eﬃciently by the augmented reality.

## Recommendation System Quantitative Testing and Evaluation

The main challenge for this project was choosing the correct collaborative ﬁltering algorithm. According to [10], an SVD++ algorithm was the one that gave the best results when applied to the MovieLens dataset. The datasets built in this FYP project were similar in structure to the MovieLens dataset. Therefore, a SVD++ algorithm was applied on the three datasets separately due to the fact that each task is not related to another and the algorithm is compared to other algorithms on the RMSE and MAE values.

The ﬁrst step of evaluation was evaluating and analysing the data which the SVD++ algorithm was going to be applied to. Three distribution barcharts per task type were plotted. Figures 42, 43, and 44 show how ratings were distributed. One can notice that the distribution of the ratings is similarly balanced, thus showing no bias towards one speciﬁc rating. The second form of plotted distributions were of distribution of number of ratings per task as shown in ﬁgures 45, 46, and 47. As previously explained, for a delivery, a visit or an interview there are multiple tasks depending on the locations of interest. The plots clearly show that some tasks received more ratings than others, thus highlighting the users’ preference for speciﬁc locations they were interested in. The third and last form of data distribution analysis is that of the distribution of a delivery, an interview or a visit per user as shown in ﬁgures 48, 49, and 50 respectively. As each distribution shows an individual user has an equivalent amount of ratings made, thus having no bias towards one speciﬁc user’s preference.

The second step of evaluation was baseline comparison of SVD++ along with other algorithms. The algorithms were compared with each other on Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values. RMSE compares a predicted value with the actual vale and measures the amount of error, whilst MAE is the absolute distance between points in a scatter plot. The other algorithms compared with the SVD++ were the ones in table 6. The SVD++ and SVD are matrix factorization based algorithms, the SlopeOne and CoClustering are collaborative ﬁltering algorithms, whilst the KNN Baseline, KNN Means and KNN Basic are diﬀerent types of classiﬁcation and regression methods which use neighbours.

A 5-fold cross-validation was carried out to compare the algorithms’ accuracy. The cross validation was carried out on the whole the three separate datasets (Visit,Interview and Delivery Tasks). On average the SVDpp performed the best across the three visitor’s tasks as shown in Figures 54 and 55. The SVDpp achieved the lowest RMSE and MAE values on average. It measured the least amount of average magnitude errors during prediction. Therefore, the algorithm was the most accurate. On average the KNN based algorithms, the collaborative ﬁltering algorithms, and the SVD Vanilla algorithm were close in accuracy as shown in Figures 53, 51, and 52. However, this could be due to the fact that they were not applied to larger-scales of datasets, were the diﬀerence in accuracy might be magniﬁed.

Finally, the hyperparameters for SVD++ is tuned using the grid search method. The ‘GridSearchCV()’ method provided by Surprise library calculates the RMSE and MAE values for every combination of hyperparameters using in this case a 5-fold cross validated dataset. Finally, it returns with a set of correct parameters that improve the model’s accuracy. Table 7 provides the hyperparameters to get the best RMSE and MAE values for each dataset SVD++ was applied to.

## Qualitative Testing and Evaluation

The questionnaire was used in order to collect qualitative data from users with regards to the system’s usability and applicability. The questionnaire followed a similar form of The System Usability (SUS) framework. Most questions could be answered in a 5-star rating frame, where one star meant ”Very Poor” of rating whilst 5 stars meant ”Excellent”, some were yes or no questions and others were short answer questions simply to acquire more detail on what the users liked or disliked. The application was tested to its limits outlining its successful features and its unsuccessful ones via video streaming to avoid any bias. In total 30 users participated half of them were employees working within the company WAAR was applied to, whilst the other half were not employees within the company (Figure 56). Each user was given 15 questions in total to answer as shown in table 9.

Questions 1 to 6 intended to obtain a general opinion of how well the application managed to accomplish its intended task. Questions 7 and 8 were about the overall performance of the recommendation system applied within the AR application. Questions 9 to 15 were about the overall performance of the augmented reality’s side of the application, its tracking and recognition capabilities. The last question was for the users to leave any recommendations of how the app may be improved, although it was optional most users left some recommendations. Overall, the users really appreciated the application’s performance and usability. They were impressed on average with its eﬃciency and naturalness in its functionality.

The main diﬀerence between the visitors and the employees was that most employees found occlusion and lighting did not aﬀect much the AR app’s performance whilst on the other hand a good number of visitors did seem to ﬁnd the latter as a problem within the application. Most users complained about the user interface. They felt that it could be improved in terms of the UI-theme by giving more appealing colours and making some instructions more visible to read, whilst also providing some form of explainable A.I. The visitors mostly found that the 3D holographic map sketch was confusing and felt it could be more informative and explainable. The second thing which bothered the users was that they could not completely understand how the recommendation worked within the application in terms of the back-end side. This may be improved by providing an instruction phase where the users are directed step by step by through the application and where explanations would be provided on how certain elements of the application functioned. However, most users could see the applicability of the application in the context it was used in and how it may beneﬁt the interns and the visitors when ﬁrst stepping foot within the workplace.

## Conclusion

This chapter discussed in detail evaluations on the features implemented within the application. The qualitative and quantitative evaluations provided show overall the application’s performance from a technical point of view and from a usability perspective. The next chapter will discuss the future work and improvements whilst also concluding the report.