Towards Field-Of-View Prediction for Augmented Reality Applications on Mobile Devices

Na Wang
George Mason University
Fairfax, VA
nwang4@masonlive.gmu.edu

Haoliang Wang
Stefano Petrangeli
Viswanathan Swaminathan
Adobe Research
San Jose, CA
{hawang,petrange,vishy}@adobe.com

Fei Li
Songqing Chen
George Mason University
Fairfax, VA
{fli4,sqchen}@gmu.edu

ABSTRACT

By allowing people to manipulate digital content placed in the real world, Augmented Reality (AR) provides immersive and enriched experiences in a variety of domains. Despite its increasing popularity, providing a seamless AR experience under bandwidth fluctuations is still a challenge, since delivering these experiences at photorealistic quality with minimal latency requires high bandwidth. Streaming approaches have already been proposed to solve this problem, but they require accurate prediction of the Field-Of-View of the user to only stream those regions of scene that are most likely to be watched by the user. To solve this prediction problem, we study in this paper the watching behavior of users exploring different types of AR scenes via mobile devices. To this end, we introduce the *ACE Dataset*, the first dataset collecting movement data of 50 users exploring 5 different AR scenes. We also propose a four-feature taxonomy for AR scene design, which allows categorizing different types of AR scenes in a methodical way, and supporting further research in this domain. Motivated by the ACE dataset analysis results, we develop a novel user visual attention prediction algorithm that jointly utilizes information of users' historical movements and digital objects positions in the AR scene. The evaluation on the ACE Dataset show the proposed approach outperforms baseline approaches under prediction horizons of variable lengths, and can therefore be beneficial to the AR ecosystem in terms of bandwidth reduction and improved quality of users' experience.

CCS CONCEPTS

• **Information systems → Multimedia streaming**; • **Computing methodologies → Mixed / augmented reality**; **Machine learning algorithms**; • **Human-centered computing → User studies**.

KEYWORDS

Augmented Reality, AR Scene, Delivery, Prediction, Field-Of-View

MMVE'20, June 8, 2020, Istanbul, Turkey
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7947-2/20/06…$15.00
https://doi.org/10.1145/3386293.3397114

ACM Reference Format:
Na Wang, Haoliang Wang, Stefano Petrangeli, Viswanathan Swaminathan, Fei Li, and Songqing Chen. 2020. Towards Field-Of-View Prediction for Augmented Reality Applications on Mobile Devices. In *International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE'20), June 8, 2020, Istanbul, Turkey.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3386293.3397114

1 INTRODUCTION

Augmented Reality (AR) transforms the physical environment around people into a digital interface. The blending of digital content into the real world provides a great level of interactivity and immersiveness in a variety of domains including health-care, education, entertainment and so on. With the introduction of AR development toolkits (e.g., ARKit by Apple and ARCore by Google), highly-detailed persistent AR experiences can now be accessible from commodity smartphones and tablets. It is reported there will be one billion AR users by 2020 [11]. Despite the increasing popularity of AR, an important bottleneck towards its widespread adoption is the large bandwidth requirement to deliver AR content. Photorealistic AR experiences require an enormous amount of data, such as high-res mesh and texture information, to make every AR object look realistic when placed in the real world. Downloading the experiences in high quality would result in high latency and therefore sub-optimal user experience.

A similar problem has been investigated in the context of 360-degree video delivery, with Field-Of-View (FOV)-dependent streaming approaches [2, 12] being proposed to solve the problem. This approach takes advantage of the fact that users' FOV is limited and only a subset of the whole 360-degree video can be consumed at a single point of time. By streaming different regions of the 360-degree video at different qualities based on the current and predicted users' FOV, FOV-dependent approaches can reduce not only the startup latency, but also the amount of data to be transmitted, without significant impact on the quality of the video content consumed by the user. A similar idea can be applied to AR content delivery as well [9], such that only the AR objects currently or likely to be in the FOV are delivered at higher quality. However, the application of FOV-based approaches to AR applications comes with its unique challenges. First, users can walk around in the AR scenes, resulting in both translational and rotational movements, as opposed to only rotational movements in the context of 360-degree videos. Second, interactivity is an important part of an AR experience. Various *triggers*, linking AR objects to user/system events and/or other objects, are introduced into AR scenes. Hence, the triggers and the interactivity associated with AR experiences may have a significant impact on users' exploration behaviors.

To address these issues, we introduce the first dataset collecting movement traces of users exploring AR scenes, the *AR Content Exploration (ACE)* dataset, which contains 6-DOF movement data of 50 users watching 5 distinct AR scenes. As an effort to facilitate our and future data collection studies, we also propose an AR scene taxonomy that considers the intrinsic characteristics of AR applications. The taxonomy allows designing AR scenes with different fundamental characteristics. Moreover, motivated by the preliminary analysis results on the ACE dataset, we develop a visual attention prediction approach that jointly utilizes information about users' historical movements and digital objects positions in the AR scene. The evaluation on the ACE dataset shows the proposed approach significantly outperforms baseline approaches in terms of prediction accuracy, from dead-reckoning to linear regression approaches. The improvement holds under variable-length prediction horizons with minimal computational overhead.

The rest of the paper is organized as follows. Sec. 2 reviews the existing works on head movement datasets of users watching 360-degree videos and FOV prediction approaches. Sec. 3 details the proposed AR scenes taxonomy, the data collection methodology, and analysis of the dataset. In Sec. 4 we present the prediction approach for AR scenes delivery, followed by the evaluation in Sec. 5, and discussion in Sec. 6. Sec. 7 concludes the paper.

2 RELATED WORK

Several works have introduced datasets of head movements of users watching 360-degree videos. In [13], 18 videos from 5 genres are watched by 48 participants.To encourage participants to focus on the scene content, participants are asked questions about virtual objects after watching each scene. This methodology has been proven effective and employed by subsequent studies. The videos classification method in [1] captures intrinsic properties of the video content, which collects head movement data of 32 users watching 4 categories of 360-degree videos. The result shows moving objects have an significant impact on users' viewport patterns. A more advanced video taxonomy is proposed in [7], which categorizes 360-degree videos based on both moving objects and camera motion.

An important usage of users' movement dataset is to predict users' movement in the future so as to improve bandwidth efficiency and the quality of users' experience. For example, the approach proposed by Corbillon et al.[3] combines 360-degree video tiles and FOV prediction by only requesting the video tiles overlapping with the predict FOV at the highest quality. This greatly help reducing the amount of bandwidth needed to stream 360-degree videos. Similar techniques include offset projections [15]. All these approaches require an accurate knowledge of the user's future FOV.

For the user's FOV prediction, existing solutions mainly utilizes two types of information: users' historical trajectory and video (or scene) content information. Users' historical trajectory is generally described as the movement in 3-DOF in the context of 360-degree videos studies. On the other hand, saliency maps provide information about the probability that a certain region of the video may attract human visual attention. These two types of information can be used separately or jointly in the FOV prediction problem. The Dead-reckoning method [6], for example, makes a future FOV prediction only based on users' historical trajectory. Bao et al. [2] use linear regression to predict FOV center locations in the future, with prediction horizons ranging from 100 ms to 500 ms. The study [4] proposes a fixation prediction network, based on LSTM networks, which leverages both historical FOV locations and video content features to predict the future FOV trajectory.

3 ACE DATASET

3.1 Creation of ACE Dataset

In this section, we present the details of the data collection study we performed on AR Content Exploration (ACE). Since the user study can only accommodate a limited number of AR scenes, we first introduce a taxonomy of AR scenes, which we use to design five representative AR scenes. Then, we present the methodology of our data collection, including the app developed for data collection, the process of user study, and the dataset structure.

3.1.1 Taxonomy of AR Scenes. The AR scenes content strongly influences users' exploration patterns. In order to study such patterns, it is necessary to create as many diverse AR scenes as possible. For this reason, we extract a set of intrinsic characteristics of common AR scenes and propose an AR scene taxonomy to guide the AR scene designs in our experiments, with the expectation that scenes belonging to the same category should result in similar users' exploration patterns. We also expect this taxonomy to support further research in this domain, by providing a set of guidelines for the design of additional AR scenes.

To this end, we explore the impact of multiple AR characteristics on users' movement and define four fundamental features of an AR scene. First, the presence of moving virtual objects is firstly introduced, as some studies have demonstrated how moving targets in the scene guide users' visual attention [8]. Second, the layout style of the digital objects in the scene is expected to play an important role as it may determine a coarse predefined path for the users' movement. Third, AR prototyping applications provide various triggers so that the digital objects in the scene can respond to users and/or system events. Such an interaction consists of the trigger events and responses associated with the objects. Lastly, the complexity of the AR scene is determined by the number of digital objects within it. Generally, the more objects in the scene, the more complex the scene is. In turn, the complexity of the AR scene directly influences the users' movements.

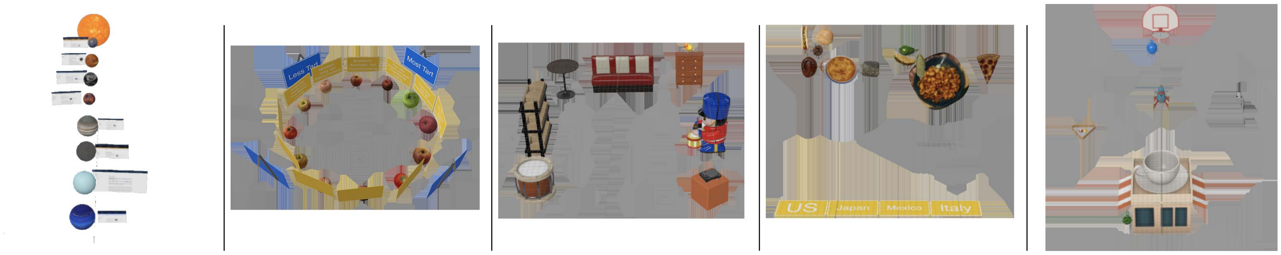
As a result, we propose a four-feature taxonomy for AR scenes, consisting of: (1) the number of digital objects, (2) the number of digital moving objects (taking on values from {zero, single, multiple}), (3) the layout style of digital objects (including linear, circular, multi-row, stratified, and random), and (4) triggers (described later in the section). Based on the above taxonomy, we create five representative AR scenes shown in Figure 1, with scene characteristics reported in Table 1. The scenes were created using the Reality Composer by Apple while models for digital objects in the scenes are either built-in resources of the app or downloaded from Sketchfab[1].

For all scenes, the right-handed coordinate system is used for the 6-DOF user movement, with z-axis pointing towards the direction

[1] https://sketchfab.com accessed Mar. 27, 2020

Table 1: Description of the AR scenes used in our study

	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5
Theme	Solar System	Apple Sweetness	Toy Room	Global Food	Fiction
# of objects	17	28	9	16	10
# of moving objects	zero	zero	single	multiple	single
Layout style	linear	circular	multi-row	stratified	random
Trigger type	action	action	distance	action	time



Scene 1: The Solar System Scene 2: Apple sweetness Scene 3: The toy room Scene 4: Global Food Scene 5: Fiction

Figure 1: AR scenes

of the viewer, and the user is initially positioned at the origin point of the scene. As this study represents the first attempt to model the exploration patterns for AR applications, we are only concerned with the visual stimuli coming from visual action triggers. The scene setting is placed in the real-world environment in a fixed relationship in terms of size, location and scale of digital objects. No manipulation by the user is allowed in the experiment, except interactions via the provided triggers. Three types of triggers employed in the scenes are described as follows. For Scene 1, action triggers are added to all planets. Once the user taps a planet, the description image appears by its side. The image stays in the scene unless the user taps on it. Similarly, action triggers are also introduced in Scene 2 and Scene 4. In Scene 3, a distance trigger is added to the 3D bird model. Once the user is in the proximity of the model, the bird is activated to fly around. The limit of proximity can be specified by the scene creator. The third type of trigger, time trigger, is instead used in Scene 5. With this trigger, two objects in Scene 5 (basketball hoop and balloon) rise from the ground after a specific period of time passes from the beginning of the experience.

3.1.2 Data Collection Methodology. We develop an iOS application that allows participants to explore AR scenes, and records their movements in 6-DOF on the basis of ARKit and RealityKit. The data is recorded to a Firebase database[2] in real time. The app can run on an iPad or iPhone. No head-mounted devices are needed in this case. To use the app, the participant is first prompted to input the assigned subject ID and then to click the *start* button to begin. During the experiment, the actual camera feed from the device is also recorded for later analysis and verification. The interaction between the participant and virtual objects is recorded too.

The experiments are carried out in a university lab (about 27 square meters), which is for the most part empty. In the experiment, each participant is involved in one session lasting between 10 and 15 minutes. Each data collection session consists of three parts: training session, watching session and final survey. The training session is designed for participants to learn the experiment flow and practice using the app. In our experiments, the participants are instructed to stare straight at the iPad screen, and move the iPad and their body together. In the watching session, each participant watches the five designed AR scenes, consecutively. During the experiment, participants can walk around to explore and interact with the virtual objects. When the exploration is considered finished for a specific scene, the participant is asked to report the number of 3D objects present in the scene. This follows the tradition of visual attention studies in 360-degree videos [13] and guarantees the participant is paying attention to the scenes content. Finally, the participant is required to answer a questionnaire, concerning the user's demographic information, previous experience with AR technologies, and the experience with the experiment.

ACE dataset has been publicly released at *https://cs.gmu.edu/~sqchen/open-access/ACE-Dataset.tgz*. It contains the following directories: *AR Scenes*, which contains all five AR scenes; *Questionnaires*, including all subjects' survey responses; *Traces* which contains the FOV trace of all subjects. Each trace has the same structure: *(subjectID, sceneID, timestamp, loc_x, loc_y, loc_z, pitch, yaw, roll)*, where the first two values indicate which AR scene is watched by which participant. The rest describes the participant's position and orientation in 6-DOF. Table 2 shows the demographics info of the participants, along with their experience with AR technologies.

3.2 ACE Dataset Analysis

We aim in this section to provide a high-level analysis of the ACE dataset. Particularly, we are interested in determining whether common features can be identified for different scenes exploration. This is an important pre-requisite for the development of effective user's movement prediction algorithms in AR applications.

A clustering algorithm [10] is used on the ACE dataset to identify possible movement patterns. However, since participants are encouraged to freely explore the scene without time limit, the duration of the watching session varies from person to person. Therefore, we first re-sample the movement data for each AR scene so that

[2] https://firebase.google.com

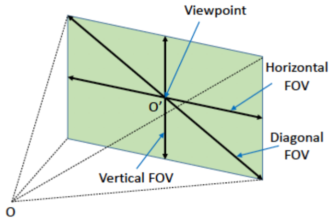


Figure 2: Illustration of FOV space[2]

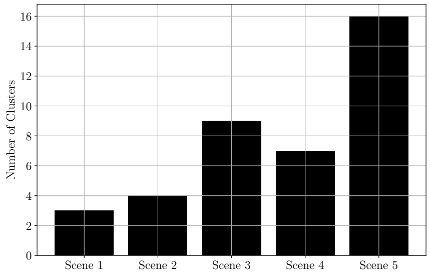


Figure 3: # of Clusters per Scene

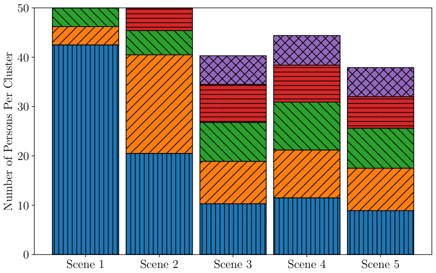


Figure 4: # of Subjects for Top-5 Clusters

Table 2: Participants demographics

Gender		Age		AR Experience	
		18-20 :	6	Never :	2
Female :	19	20-24 :	19	1-5 times :	14
Male :	31	25-30 :	18	6-10 times :	31
		>30 :	7	>10 times :	3

the duration of the watching session per scene remains the same among all participants. For each scene, the maximum watching session duration is used as the standard length. For those shorter sessions, a linear interpolation is applied to construct new data points in 6-DOF within the range of the existing movement data.

After data pre-processing, we cluster traces for which the *FOV space*, which is described as a squared pyramid as shown in Figure 2, overlaps more than 80% for about 60% of the whole watching session. We choose to cluster traces on the basis of the actual FOV instead of the 6-DOF movement traces because two users with different movement traces may look at the same objects. For the cluster analysis results, Figure 3 shows the number of identified clusters for each AR scene, while Figure 4 shows the average number of participants in the 5 most populated clusters for each AR scene. The fewer the number of clusters, the more similarity participants share in the exploration of the AR scene.

As shown in Figure 3, a large number of clusters for Scene 5 entails that the scene does not have predominant features with enough saliency to attract users. In Scene 5, 7 virtual objects are located randomly in the scene, and no semantic relationships exist among the 3D objects (see Figure 1Scene 5:). Therefore, participants demonstrate different watching patterns. Moreover, we find that the time trigger does not influence participants in the same way as the other two triggers (e.e., action- and distance-based). Indeed, without explicit direction, few participants would notice the rising objects (basketball hoop and balloon).

In contrast, the numbers of clusters for both Scene 1 and Scene 2 are small, as shown in Figure 3. In Scene 1, all planets are placed in a straight line. When the participant stands at the origin point, the straight line of planets appears exactly in front of the participant. Moreover, the natural semantic relationship of the planets' objects attracts the participants to move forward along the planets' line to watch objects and description images activated by users' actions. To study the relationship between the user and digital objects, we retrieve the object access sequence as did in the study by Zhou et al on the area of interest in 3D scenes [16]. The sequence is retrieved in time order by aligning the trace data with the recorded videos. If all planets in Scene 1 are labeled as following: Sun (0), Mercury (1), Venus (2), Earth (3), Mars (4), Jupiter (5), Saturn (6), Uranus (7), and Neptune (8), and associated 2D descriptions are numbered 1' – 8', the object access sequences for most participants is {8/8', 7/7', 6/6', 5/5', 4/4', 3/3', 2/2', 1/1'}, i.e., explore each planet from furthest to closest to the Sun. Similarly, in Scene 2, most participants walk along the circle of 3D apples to examine the apple objects. In this case, the behaviors differ in the movement direction, i.e., clockwise or counterclockwise.

Scene 3 and Scene 4 provide intermediate similarity in exploration patterns compared to the other scenes (Figure 3). Objects in Scene 3 are placed in multiple rows, so the watching behaviors are more dispersed than the first two scenes. The participant may choose different ways to finish the scene exploration. Moreover, the flying bird activated by the distance trigger further diversifies the possible movements. Similar results can be seen for Scene 4. By comparing object access sequences retrieved from the traces, it is observed that traces in the same cluster lead to highly similar object access sequences. This observation is the basis of our approach to predict objects to be likely watched by the user in the near future.

4 FIELD-OF-VIEW PREDICTION

Motivated by results presented in Section 3.2, we present a new prediction approach based on the users' object access sequence pattern. Being able to correctly predict this sequence could allow to pre-load objects that are most likely to be watched by the user in the near future.

We first extract the virtual objects in the users' FOV space (Figure 2) for each timestamp of the watching session, for each user trace. Virtual objects in the scene are labeled with integers in the ascending order. The object access extraction result for each timestamp is potentially a set of objects. In this work, however, we only consider the closest one to the user in FOV as actually being watched by the user, since the interaction with a specific object forces the user to focus on the object. Inspired by how the Hamming distance quantifies the distance between two strings of equal length [5], we define the distance between two object access sequences as the number of symbols which are different at the same position in both sequences.

Next, we apply agglomerative hierarchical clustering to partition a set of object access sequences into a set of clusters [14]. We use the traces from 30 participants as the training set and the remaining 20 traces for test. Based on the computed Hamming distance of every pair of object access sequences, we group sequences into a hierarchical tree. The process is repeated until a single cluster is reached. We use the results shown in Figure 3 to set the input parameter for each scene in advance. For each cluster, we choose as representative the sequence with the minimum average Hamming distance to others within the cluster. The result of this step is, for each scene, a set of representative sequences.

During testing, we dynamically select the most matched patterns from the resulting set found during training, which entails selecting the sequence with the minimum Hamming distance to the user's current sequence. Once the next element of object access sequence is updated but not matched with the choice, a new selection process is performed for the updated user sequence.

5 EVALUATION

We compare the proposed prediction algorithm with other three existing approaches: (i) No prediction (NP), which uses the current FOV as the future FOV, (ii) Dead Reckoning (DR) [6], which uses the user's velocity to predict the future FOV, (iii) Linear regression (LR) model, which predicts the future FOV based on the past trajectory. We evaluate the four solutions on the ACE dataset we collected, in terms of prediction accuracy and prediction time cost, with variable prediction horizons for all scenes. To compare the prediction accuracy of four approaches, we need to convert the FOV prediction results for NP, DR, and LR into object access sequences. As for our method, only the closest object in users' FOV is considered. If no object is present in the predicted FOV space, the symbol X is added to the sequence.

The results of experiments are presented in Figure 5(a) to Figure 5(e). We can observe that all approaches demonstrate high prediction accuracy when the prediction horizon is shorter than 500 ms. However, the prediction accuracy for NP, DR, LR decreases quickly as the prediction horizon increases, implying the non-linearity of FOV trajectories. In contrast, our proposed solution based on the object access sequences patterns outperforms all three baselines by a large margin when the prediction horizon is longer than 500 ms.

For the first two scenes, the accuracy of our proposed approach maintains a high value. This can be attributed to the very limited number of object access sequence patterns for the first two scenes. The results are in accordance with our previous analysis in Section 3.2. For the solar system scene, most participants explore the scene by walking along the planets straight line (see Figure 1Scene 1:). Similarly, for Scene 2, two clusters, representing participants exploring the scene in clockwise or counterclockwise direction, are sufficient to cover most participants.

Scene 3 and Scene 4 demonstrate intermediate similarity in exploration patterns, compared to the other three scenes (figure 3). Because of the limited number of clusters, the prediction accuracy of our approach for both Scene 3 and 4 is still better than the three baselines. Even so, the accuracy for these two scenes is lower than that of Scene 1 and 2, mainly because of more complex scenes layout and more diverse exploration patterns.

For Scene 5, the prediction accuracy of our approach is still better than that of the other solutions, even though the digital objects are placed randomly, and participants demonstrate very different movement patterns. The underlying reason for these results is the employment of the object access sequences pattern for the cluster analysis, instead of the explicit FOV. Indeed, two users with different movement traces may look at the same objects even if their FOVs do not overlap at all.

We also investigate the prediction time cost of four approaches, for Scene 3. Figure 5(f) reports the prediction time of all approaches, for a specific participant under prediction horizons of varying length. The time cost of both NP and DR solutions are low and stable. The NP solution does not perform any computation, so the time cost is always zero. The time cost of DR solutions is low and stable. The other two solutions are similar, varying from 4.5 ms to 7.9 ms. We also calculate the average prediction time cost across all scenes and all traces, shown in Table 3. As the table shows, the time cost of our solutions is very close to DR, and thus can be considered negligible.

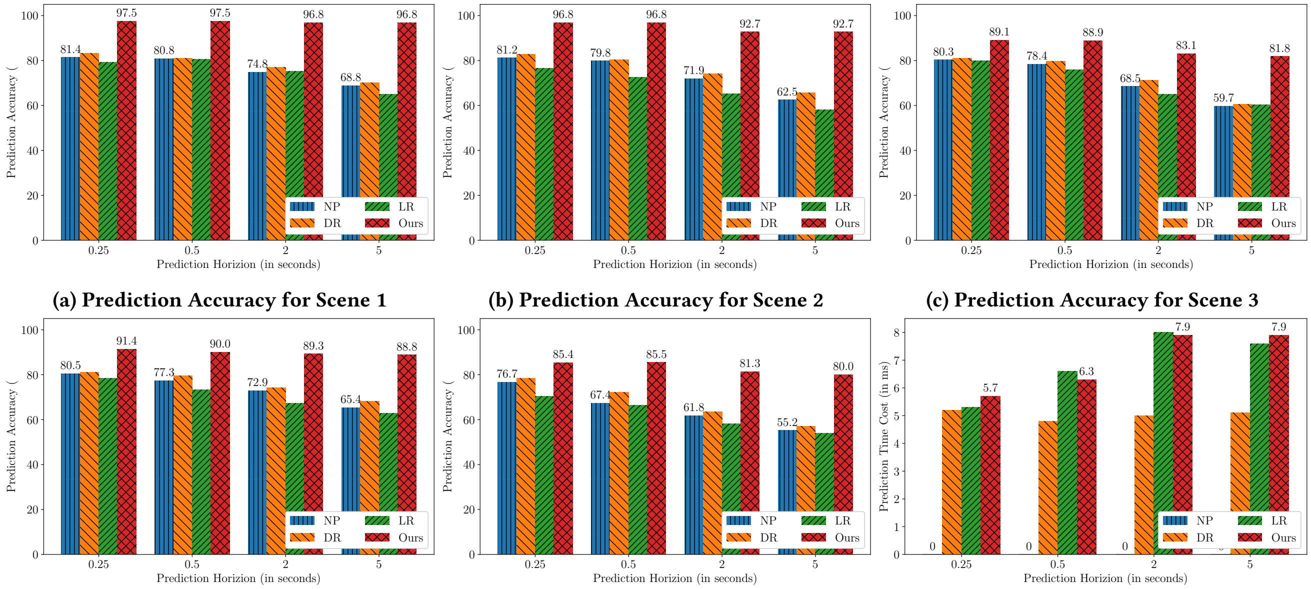
Table 3: Comparison of average prediction time (ms)

	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5
NP	0	0	0	0	0
DR	4.7	4.9	5.0	4.8	4.4
LR	4.9	5.1	5.4	5.2	4.7
OURS	4.5	4.5	5.3	5.5	4.6

6 DISCUSSION

As a first attempt to investigate the potential of FOV prediction for AR applications, our proposed approach jointly utilizes users' trajectory and AR scene content information. Inspired by the fact that participants may follow specific object access patterns (although they could move freely during the data collection study), the prediction based on such pattern is a promising solution to effectively predict the scene content that is likely to be accessed in the future. In the experiment, since users have to move and interact with objects in a limited space, the complexity of the scenes is limited in the current dataset. So the usefulness of the FOV prediction may be weakened in such small AR scenes. However, as suggested in [9], streaming of AR scenes is necessary to ensure good user experience as AR scenes may contain many objects that are often large in size (tens to hundreds of MB per object), especially with increased photo-realism. The time it takes to cache each object will likely to exceed the duration of the viewing session. Hence, it is necessary to have FOV prediction to determine the right priority and therefore decide on the best object and quality level to prefetch.

As an initial step, our work has several limitations and can be improved in following directions. First, the Hamming distance is employed here to measure the distance between two objects access sequences. Extra test is required to avoid mis-clustering. Therefore new distance measurement such as Jaccard index may be used to replace it in the future work. Second, the normalization on users' movement traces can impact the length of the prediction horizon, because it stretched the time users spent on partial or all actions while they were exploring the scenes. The performance of the proposed approach may thus degrade with increase of the prediction horizon. In the future, we expect to improve the prediction under longer horizons with higher precision.



(a) Prediction Accuracy for Scene 1

(b) Prediction Accuracy for Scene 2

(c) Prediction Accuracy for Scene 3

(d) Prediction Accuracy for Scene 4

(e) Prediction Accuracy for Scene 5

(f) Prediction Time Cost for Scene 3

Figure 5: Prediction Accuracy and Time under Variable-Length Prediction Horizons

7 CONCLUSION

In this work, we propose a taxonomy to classify AR scenes based on the number of moving objects and virtual objects, scene layout and type of triggers. Following the taxonomy, five AR scenes are designed and a 50-user study is conducted to collect 6-DOF movement data while subjects explores the scenes. The resulting *ACE* dataset represents the first publicly available dataset on AR scene exploration. Moreover, we have proposed and implemented a new approach that utilizes both users' movement trace and AR scene information to predict scene content to be likely viewed in the near future, and therefore potentially improve the AR scene delivery. In our solution, users' movement trace data are translated into objects access sequences. The common users' object access patterns are then extracted to predict new user's movement in the same scene. We evaluate the approach with three baseline algorithms on the ACE dataset. The results show that our proposed method significantly improves the prediction accuracy even under long prediction horizons, with negligible computing costs.

8 ACKNOWLEDGEMENTS

We appreciate the constructive comments from reviewers and our shepherd. This work is supported in part by the NSF grant CNS-1524462 and gift funding from Adobe Research.

REFERENCES

[1] Mathias Almquist, Viktor Almquist, Vengatanathan Krishnamoorthi, Niklas Carlsson, and Derek Eager. 2018. The prefetch aggressiveness tradeoff in 360 video streaming. In *Proceedings of the 9th ACM Multimedia Systems Conference*. 258–269.

[2] Yanan Bao, Huasen Wu, Tianxiao Zhang, Albara Ah Ramli, and Xin Liu. 2016. Shooting a moving target: Motion-prediction-based transmission for 360-degree videos. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 1161–1170.

[3] Xavier Corbillon, Alisa Devlic, Gwendal Simon, and Jacob Chakareski. 2017. Optimal set of 360-degree videos for viewport-adaptive streaming. In *Proceedings of the 25th ACM international conference on Multimedia*. 943–951.

[4] Ching-Ling Fan, Jean Lee, Wen-Chih Lo, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 2017. Fixation prediction for 360 video streaming in head-mounted virtual reality. In *Proceedings of the 27th Workshop on Network and Operating Systems Support for Digital Audio and Video*. 67–72.

[5] Richard W Hamming. 1950. Error detecting and error correcting codes. *The Bell system technical journal* 29, 2 (1950), 147–160.

[6] Aditya Mavlankar and Bernd Girod. 2010. Video streaming with interactive pan/tilt/zoom. In *High-Quality Visual Experience*. Springer, 431–455.

[7] Afshin Taghavi Nasrabadi, Aliehsan Samiei, Anahita Mahzari, Ryan P McMahan, Ravi Prakash, Mylène CQ Farias, and Marcelo M Carvalho. 2019. A taxonomy and dataset for 360° videos. In *Proceedings of the 10th ACM Multimedia Systems Conference*. 273–278.

[8] Cagri Ozcinar and Aljosa Smolic. 2018. Visual attention in omnidirectional video for virtual reality applications. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–6.

[9] Stefano Petrangeli, Gwendal Simon, Haoliang Wang, and Vishy Swaminathan. 2019. Dynamic Adaptive Streaming for Augmented Reality Applications. In *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, 56–567.

[10] Silvia Rossi, Francesca De Simone, Pascal Frossard, and Laura Toni. 2019. Spherical clustering of users navigating 360 content. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4020–4024.

[11] AR Statistics. 2018. *6 VR & AR Statistics: Shaping the Future of Augmented Reality with Data*. https://www.newgenapps.com/blog/6-vr-and-ar-statistics-shaping-the-future-of-augmented-reality-with-data

[12] Liyang Sun, Fanyi Duanmu, Yong Liu, Yao Wang, Yinghua Ye, Hang Shi, and David Dai. 2018. Multi-path multi-tier 360-degree video streaming in 5G networks. In *Proceedings of the 9th ACM Multimedia Systems Conference*. 162–173.

[13] Chenglei Wu, Zhihao Tan, Zhi Wang, and Shiqiang Yang. 2017. A dataset for exploring user behaviors in VR spherical video streaming. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. 193–198.

[14] Ying Zhao, George Karypis, and Usama Fayyad. 2005. Hierarchical clustering algorithms for document datasets. *Data mining and knowledge discovery* 10, 2 (2005), 141–168.

[15] Chao Zhou, Zhenhua Li, and Yao Liu. 2017. A measurement study of oculus 360 degree video streaming. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. 27–37.

[16] Zhong Zhou, Ke Chen, and Jingchang Zhang. 2015. Efficient 3-D scene prefetching from learning user access patterns. *IEEE Transactions on Multimedia* 17, 7 (2015), 1081–1095.