International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2017)
R. Lindeman, G. Bruder, and D. Iwai (Editors)

Reference Framework on vSRT-method Benchmarking for MAR

R. Ichikari[1], T. Kurata[1], K. Makita[2], T. Taketomi[3], H. Uchiyama[4], T. Kondo[5], S. Mori[6] and F. Shibata[7]

[1]National Institute of Advanced Industrial Science and Technology (AIST), Japan
[2]CANON INC., Japan
[3]Nara Institute of Science and Technology, Japan
[4]Kyushu University, Japan
[5]The Open University of Japan, Japan
[6]Keio University, Japan
[7]Ritsumeikan University, Japan

Abstract

This paper presents a reference framework on benchmarking of vision-based spatial registration and tracking (vSRT) methods for Mixed and Augmented Reality (MAR). This framework can provide typical benchmarking processes, benchmark indicators, and trial set elements that are necessary to successfully identify, define, design, select, and apply benchmarking of vSRT methods for MAR. In addition, we summarize findings from benchmarking activities for sharing how to organize and conduct on-site and off-site competition.

CCS Concepts

•***Human-centered computing*** → *Mixed / augmented reality;*

1. Introduction

In the development of mixed and augmented reality (MAR) applications, spatial registration and tracking methods, especially visionbased methods, represent some of the most important technologies. In this context, research and development of vision-based spatial registration and tracking (vSRT) is flourishing and many algorithms have been proposed every year. Therefore, standardized frameworks on benchmarking of vSRT methods have become necessary in order to foster objective evaluation and comparison of diverse vSRT methods. Such frameworks would also facilitate much fairer competition among small and major companies and institutes involved in MAR technologies, applications, and services.

Based on this background, the TrakMark working group (WG) was established in 2009 as a subsidiary organization of the Special Interest Group on Mixed Reality (SIG-MR) in the Virtual Reality Society of Japan (VRSJ) to standardize benchmark schemes. Table 1 shows a history of TrakMark WG activities [TK*09] [SIKU*10] [TMK11] [MOK14] [TMK12] [TMK13] [TMP14] [TMC15] [TMW]. In accordance with such grass-roots activities including on- and off-site comparisons of vSRT methods and MAR systems [LBM*09] [KMP*13] [WGB13], which are often held as contests [PITC], three core components as shown in Fig. 1 are identified and defined for creating a reference framework.

- **Benchmarking Processes**, which include how to produce benchmarking outcomes using benchmark indicators and trial sets, as well as how to share benchmarking outcomes.
- **Benchmark Indicators**, which quantify the performance of vSRT methods in MAR by considering not only the characteristics of vSRT methods in MAR such as reliability and temporal characteristics, but also fair comparison.
- **Trial Set Elements**, which consist of datasets and physical object instances for providing each benchmarking attempt with the same condition.

On-site benchmarking methods are used to conduct benchmarking on the spot while capturing images of physical objects with working MAR systems. Because human factors such as a contestant's limited preparation time inevitably affect on-site benchmarking result, stakeholders often lose focus in terms of what they are supposed to evaluate. In addition, time and cost constraints exist for competition organizers as benchmarking service providers. Therefore, simplifying the benchmarking framework is often necessary for practical operation of on-site tracking competitions. However, the pros and cons shall be considered with the findings given in Section 5.

By contrast, off-site benchmarking methods are used to conduct benchmarking with target images in datasets prepared in advance. Compared to on-site benchmarking methods, both contestants and competition organizers have more time to prepare and conduct benchmarking. However, organizers must make additional effort to alleviate issues related to fine-tuning the benchmarking process/methods.

Typical processes of on- and off-site benchmarking are extracted

DOI: 10.2312/egve.20171363

from grass-roots activities, and they are schematically described in Section 2 by referring to ISO/IEC 29155 series, especially ISO/IEC 29155-1 [I291].

Each benchmark indicator and trial set element is also extracted from outcomes and discussions in the grass-roots activities. Section 3 describes three major types of benchmark indicators: reliability, temporality, and variety. Section 4 describes reference elements in a trial set, which contains dataset elements and physical object instances.

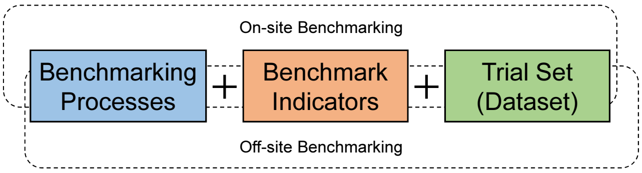

Figure 1: *Core components of on-site and off-site benchmarking framework.*

2. Benchmarking processes

This section outlines benchmarking processes and related components necessary to produce and share benchmarking outcomes. Fig. 2 illustrates the basic benchmarking process flow. Although the flow may differ for each specific benchmarking process, it generally consists of process, target, input, output, outcome, and organized storage, described in detail as follows:

- **Process**, which consists of one or more micro processes:
  - to develop and gather vSRT methods and MAR systems,
  - to prepare and conduct benchmarking,
  - to provide and maintain benchmarking instruments and repositories,
  - and to share benchmarking results.
- **Target**, which is a vSRT method or MAR system used as a benchmarking target.
- **Input**, which consists of MAR systems that use vSRT methods, trial sets, and physical objects.
- **Output/Outcome**, which includes:
  - benchmarking instruments such as benchmark indicators and tools, methods, or guides,
  - benchmarking results such as benchmarks, intermediate results, and reports,
  - and benchmarking surveys.
- **Organized storage**, which may be a benchmarking repository or other external repositories.

Note that trial sets are classified in the Input process as previously described. However, they are also regarded as important outcomes of benchmarking activities. Various stakeholders are involved in benchmarking vSRT methods for MAR. Fig. 3 illustrates a typical example of the correspondence between stakeholders and their roles and activities. Based on roles, stakeholders can logically be classified into the following groups:

- **Benchmark provider**, who surveys benchmarking results, creates and gathers datasets, maintains benchmarking repositories, and provides benchmark surveys;
- **Benchmarking service provider**, who develops and provides benchmarking instruments, conducts benchmarking at the request of technology users, and submits benchmarking results to a benchmarking repository;
- **Technology developer**, who develops vSRT methods, MAR systems, or MAR services;
- **Technology supplier**, who supplies vSRT methods, MAR systems, or MAR services that technology developers have developed;
- **Technology user**, who utilizes the outcomes of benchmarking.

Of course, various role-sharing schemes can be used in practice. Any person or organization may fulfill one or more roles. For example, benchmark providers can also be benchmarking service providers. By contrast, one role may be fulfilled by several persons or organizations. For example, the competition organizers together with contestants often fulfill the role of benchmarking service provider in conducting benchmarking.

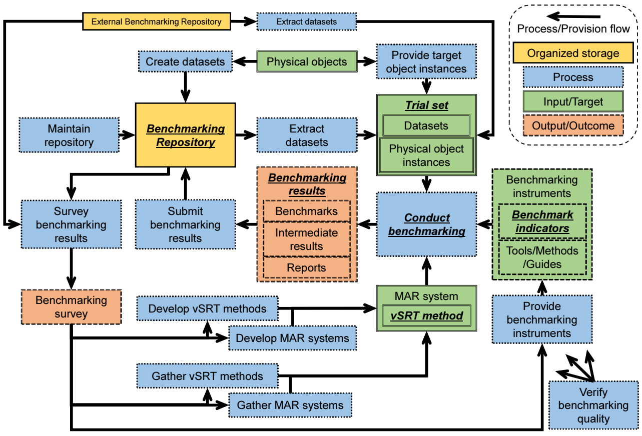

Figure 2: *Basic benchmarking process flow.*

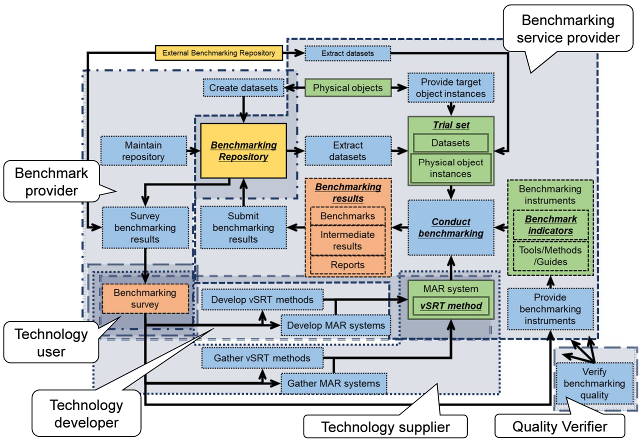

Figure 3: *Example of stakeholders and their roles.*

Table 1: *History of TrakMark WG activities: Breakthroughs and technical findings.*

Year	Event	TrakMark WG activities	■Breakthrough ●Technical findings
2009	ISMAR2009	TrakMark Proposal was disclosed.	■The process of evaluation at the 1st phase was proposed. ●Categorization of tracking methods is needed. ●Criteria for evaluation is needed. ●Dataset is needed.
2010	ISMAR2010	The first International TrakMark workshop was conducted.	■First version of evaluation criteria in TrakMark was proposed. ●Dataset and evaluation methods designed by Metaio was proposed (closed ground truth, how to set threshold values) ●HIT Lab's Potential (categorization of dataset, simulation platform, how to conduct a test)
2011	ISMAR2011	The 2nd International Workshop was conducted.	■Benchmarking results with TrakMark dataset were presented (by Hayashi et al, and Antoine Petit et al) ●Depth data is needed for benchmarking tracking methods using depth sensor.
2012	ICPR2012	The 3rd International Workshop (TrakMark2012) was conducted.	■Image sequences captured by a RGB-D sensor were presented (by Daniel Cremers from Technische Universitat Munchen (TUM)) ■Methods of creating ground truth or reference data of camera position and orientation were proposed. -A method using motion capture system (by Daniel Cremers et al) -A method using visual markers (by Evgeniy Minaev et al) -A method using CG models (by Koji Makita et al, Sara Martull et al) ●Meta-data with Dataset may be effective. -Forward-backward data / Moving object -Region of Interest ●Dataset-adjustment for specific tracking methods (e.g. methods using line detection, methods using visual markers). (by Keisuke Hirose et al, Yukiko Kubo et al) ●Relation between specifications of hardware devices and tracking results. ■Benchmarking method independent of specifications of hardware devices with bootable USB of Linux system (by Yoshinari KAMEDA).
2013	ISMAR2013	Joint Workshop on Tracking Methods & Applications and TrakMark was conducted.	●Cost of set up needed before using AR/MR application. ■Evaluation using PEVO (Projection Errors of Virtual Objects). (by Masayuki Hayashi et al)
2014	ISMAR2014	TrakMark panel session was conducted.	■Current status of TrakMark dataset was presented. (25 datasets to public、380 Blu-ray disks have been distributed)
2015	ISMAR2015	Tracking competition was conducted. (TrakMark was one of supporters of the competition)	■Level 3 submissions (most difficult level) for off-site competition were evaluated by the ISMAR2015 Tracking Competition committee. Evaluation criteria in TrakMark based on PEVO(*) was applied for the evaluation. ●Not only PEVO(*) but also completeness of a trial is important. (*PEVO: Projection Error of Virtual Objects)

Table 2: *Benchmark indicators for off-site and on-site benchmarking.*

	Off-site	On-site
Reliability	• 3DEVO • PEVO • Reprojection error of image features • Position and posture errors of a camera	• 3DEVO • PEVO • Reprojection error of image features • Position and posture errors of a camera • Completeness of a trial
Temporality	• Throughput • Latency	• Throughput • Latency • Time for trial completion
Variety	• Number of datasets • Variety on properties of datasets	• Number of trials • Variety on properties of trials

3. Benchmarking indicators

This section outlines three major types of benchmark indicators (reliability, temporality, and variety), which should be considered for fair comparison of vSRT methods in MAR. Table 2 shows representative benchmark indicators for off- and on-site benchmarking.

3.1. Reliability indicators

The following four indicators on reliability are for both off- and on-site benchmarking.

- **3D error of a virtual object (3DEVO)**, which is the difference between the estimated position of a virtual object and the ground truth. 3DEVO is one of the most direct and intuitive indicator for vSRT methods for MAR, as one of the most principal functions of MAR systems is to allign virtual objects in the 3D space based on the results obtained by the target vSRT method.
- **Projection error of a virtual object (PEVO)**, which is also one of the most direct and intuitive indicators for vSRT methods for

MAR, as one of the most important functions of MAR systems is to render virtual objects based on the results obtained by the target vSRT method. Assuming the simplest case in which a virtual point is projected as a virtual object to an estimated image plane, the distance between the projected and correct point is calculated as a PEVO value (see Fig. 4). The PEVO value may be measured in degrees or pixels. The angular distance measure can provide a uniform measure in a screen space, whereas the pixel number varies depending on the position in a screen space.

- **Re-projection error of an image feature**, which is the distance between a detected image feature in an image plane and the re-projection to the image plane with the 3D coordinates of the image feature that are recovered based on the target vSRT method. Assuming the simplest case in which the image feature is a feature point, the re-projection error may be the distance between the detected feature point and the re-projected point, and may be measured in degrees or pixels as with PEVO.
- **Position and posture errors of a camera**, which is the difference between the estimated position and posture of a camera and the ground truth.

In addition to the aforementioned reliability indicators, completeness of a trial should be employed, especially for on-site benchmarking. This is because in many on-site competitions, many MAR systems cannot help but to stop performing spatial registration and tracking in the middle of trial.

- **Completeness of a trial**, which involves evaluating the extent of a trial completion. It is regarded as the robustness of the target vSRT method.

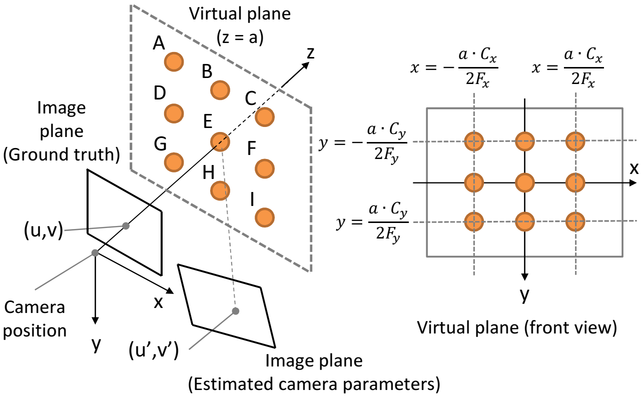
Virtual plane
(z = a)
A
B
C
D
E
F
G
H
I
z
Image plane (Ground truth)
(u,v)
Camera position
x
y
(u',v')
Image plane (Estimated camera parameters)
$x = -\frac{a \cdot C_x}{2F_x}$
$x = \frac{a \cdot C_x}{2F_x}$
$y = -\frac{a \cdot C_y}{2F_y}$
$y = \frac{a \cdot C_y}{2F_y}$
x
y
Virtual plane (front view)

Figure 4: *PEVO with 3D coordinates of virtual objects.*

3.2. Temporality indicators

The following two indicators on temporality (see Fig. 5) are basically suitable for off-site benchmarking.

- **Throughput**, which is the rate at which a target image is processed through a target vSRT method or target MAR system during a specific period. It is often called frame rate.
- **Latency**, which is the time delay inevitably produced by a target vSRT method or target MAR system.
  - ○ For MAR-system benchmarking, the latency may be the length of time from when starting to capture a target image with the system to when rendering a virtual object based on the estimated position and posture of a camera with which the target image was captured.
  - ○ For vSRT-method benchmarking, the latency may be the length of time from when starting to load a target image into the target method for input to when returning the estimated position and posture of a camera with which the target image was captured.

Table 3: *Trial set for benchmarking.*

		Off-site	On-site
Dataset	Contents	• Image sequences • Intrinsic/extrinsic camera parameters • Challenge points • Optional contents • 3D models for the target objects and for virtual objects • Image feature correspondences • Depth image sequences • Self-contained sensor data, etc.	• Challenge points • 3D models for the target objects and for virtual objects
	Metadata	• Scenario • Camera motion type • Camera configuration • Image quality	• Scenario
Physical object instances	Contents	• Physical objects	
	Metadata	• Information on how to find the physical objects	

For on-site benchmarking, the time for trial completion may be used as a temporality indicator because it is easy to measure.

- **Time for trial completion**, which is the length of time from starting a trial to finishing it.

Actually, the time for trial completion is often used in on-site competitions and can represent the overall performance of a target MAR system. However, that it may include other aspects of the MAR system such as image capturing, virtual object rendering, and human factors regarding the operator should be considered.

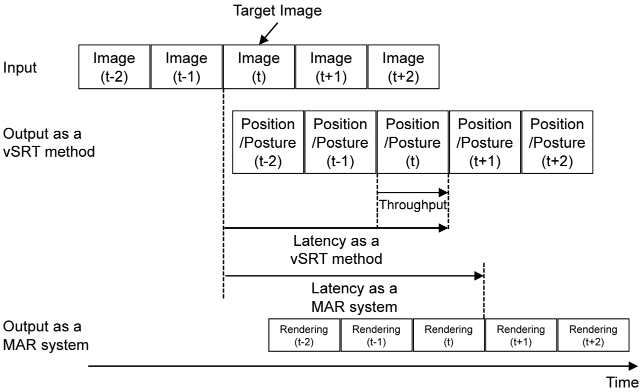

Figure 5: *Temporality indicators: latency and throughput.*

3.3. Variety indicators

This subsection presents two variety indicators required to prevent fine tuning and cheating with some specific datasets for benchmarking. The following two indicators are for off-site benchmarking.

- **Number of trial sets**, which is the number of datasets used to obtain a benchmark indicator.

- **Variety on properties of trial sets**, which is variety on properties of trial sets used to obtain a benchmark indicator. Typical examples of properties are the following:
  - Camera motion types, camera configurations, image quality, and lighting conditions.

  In general, performing fine tuning or cheating with many trial sets is difficult. However, this difficulty may diminish if properties of the datasets used for benchmarking are homogeneous.

  The following two indicators are for on-site benchmarking.
- **Number of trials**, which is the number of trials attempted in an on-site benchmarking.
- **Variety on properties of trials**, which is variety on properties of trials in an on-site benchmarking.

4. Trial set for benchmark

This section identifies the reference elements in a trial set for benchmarking. Table 3 shows representative elements in a trial set for on- and off-site benchmarking.

4.1. Dataset for off-site benchmarking

- **Contents**
  - Image sequences, which are target image sequences.
  - Intrinsic/extrinsic camera parameters, which are the ground truth of intrinsic and extrinsic parameters for one or more video cameras to capture the image sequences.
  - Optional contents
    - 3D models for both target and virtual objects, which are 3D model data for the target objects in image sequences and for virtual objects overlaid in benchmarking.
    - Image feature correspondences, which is the ground truth of image feature correspondences through the image sequences.
    - Depth image sequences, which are image sequences that each pixel value is measured with depth sensors using, for example, an active stereo method.
    - Self-contained sensor data, which are sensor data measured by self-contained sensors such as an accelerometer, gyro sensor, magnetometer, and barometer.
- **Metadata**
  - Scenario, which describes an application scenario of MAR, such as an indoor/outdoor navigation, tabletop MAR interface, or industrial application.
  - Camera motion type, which describes camera motion such as translation only, rotation only, walking motion, handheld motion, or vehicle motion.
  - Camera configuration, which describes the camera configuration such as white balance, shutter type (global or rolling), and shutter speed.
  - Image quality, which describes the target-image quality such as image resolution, defocusing, and motion blur.

The metadata can be used as a source of variety indicators and can also correspond to properties of each dataset for ease of dataset retrieval.

4.2. Dataset for on-site benchmarking

- **Contents**
  - **Challenge points**, which are points for rating each MAR system with a 3DEVO or a PEVO measurement. The positions of each challenge point are estimated and visualized by the target MAR system for rating.

4.3. Physical objects for off-site benchmarking

- **Contents**
  - **Physical objects**, which must be easily available or deliverable physical objects that can be captured as images by contestants. Such physical objects are necessary for contestants in developing and adjusting their vSRT method or MAR system before conducting off-site benchmarking with target image sequences in which the physical objects are observed.
- **Metadata**
  - **Information on how to find the physical objects**

In actual benchmarking activities, paper models and toy bricks can be employed as easily available or deliverable physical objects, as in [TMC15] [COS].

4.4. Physical objects for on-site benchmarking

- **Contents**
  - **Physical objects**, which are physical objects for benchmarking preparation and actual benchmarking. For preparation use, physical objects must be easily available or deliverable as in off-site benchmarking. However, for actual benchmarking, physical objects may be observable only during trials, as in [TMC15] [VTC14]

5. Findings from the ISMAR 2015 tracking competition

5.1. Off-site competition

Benchmark indicator

PEVO is one of the most direct and intuitive indicators for vision-based geometric registration and tracking methods for MAR. PEVO worked well because it is sensitive to errors of position and orientation of the camera. Actually, some cases existed in which PEVO was large even if re-projection errors of image features was small. The organizers had asked each contestant to submit projective camera matrices for data with lens distortions. However, some contestants submitted the normalized camera matrices that consisted of position and orientation and that did not contain the intrinsic parameters of the camera. In addition, some other contestants submitted the projective camera matrices. This made it difficult for organizers to fairly and equally evaluate each contestant.

Scalability

Dealing with various submissions is becoming problematic with the tremendous increase in the number of contestants. Ensuring that each contestant thoroughly understand the rules and regulations is also difficult. That the organizers provide software for evaluation as well as datasets and that each contestant submits only certain parts, such as final trajectories, are appropriate actions to take.

Cheating and fine tuning

Judges should evaluate each method by themselves, for instance, by having each contestant submit a binary (executable) program. Otherwise, checking for cheating (such as using future information to conduct global optimization using the entire data) would be difficult. To alleviate fine-tuning issues, providing contestants with various types of datasets should be effective.

5.2. On-site competition

Effects of Simplification

Simplifying the benchmarking process is often necessary for practical operation of tracking competitions. However, the following pros and cons should be considered.

[Equality and Simplification]

In the competition, the condition for each contestant was not strictly equal. Each contestant was supposed to indicate (with a mark on a textured paper attached on a partition wall) the specified challenge point, which was a virtual point determined as the simplest virtual object for accuracy evaluation and estimated by each MAR system (see Fig. 6). To save time for competition operation, the organizers did not change the paper for subsequent contestants after being marked by previous contestants. Therefore, a paper having a mark became a part of the environment for subsequent contestants. This might not have been fair, as the appearance of the paper had changed due to the mark, even if the mark was decidedly small.

[Measurement and Simplification]

Ideally, the measurement for accuracy evaluation should be based on the distance between the 3D coordinates of the challenge point given by the organizers and those of the challenge point estimated by the contestant's MAR system in the 3D space. However, in the aforementioned competition, it was based on the 2D coordinates on the textured paper. This was because measuring the distance of arbitrary points in the air in the 3D space was quite difficult. More than three challenge points on a plane are necessary if the organizers want to evaluate strictly the 3D position and orientation of the camera with these 2D coordinates.

Benchmark Indicator

In the on-site competition, the organizers employed the following strategy for evaluation:

(1) Compare the number of challenge points each contestant found.

If (1) is the same for multiple contestants,

(2) Compare the mean distance, which corresponds to 3DEVO

If (2) is also the same for multiple contestants,

(3) Compare the time for trial completion.

As a result, only (1) was determined to be important because only one contestant completed the trial. In terms of a comprehensive evaluation, utilizing other benchmark indicators such as temporality indicators including frame rate and latency would be worthwhile.

Measurement process

From the organizer's point of view, on-site competitions should be held at various locales. Therefore, using the same or a standardized tool in the preparation phase for measuring environments and ground truth and in order to produce correspondence between the real-world coordinate system and the local one is preferable. In the trial phase of this competition, the 3DEVO was measured manually using a ruler. The measurement was easily realized because a challenge point should be located on a plane. However, if it is not on a plane, a better means of measuring the 3DEVO should be considered.

From the contestant's point of view, two choices exist for acquiring and constructing environmental 3D models in each MAR system by each contestant: before or during a trial. In this competition, online acquisition (or visual SLAM) was required. Considering various scenarios for MAR system usage, encompassing modeling before a trial in competition is preferable.

Human factor

Because in the aforementioned competition the user of the MAR system was supposed produce marks on the paper attached to the partition wall, when the camera panned closer to the paper, image registration and tracking often became unstable (see Fig. 6). As expected, one of the major techniques used to obtain high scores in the competition was to devise and master a method of moving the camera.

Even though detailed rules and regulations were documented, they were not well understood by the contestants because of the very busy schedule of the competition and the rules and regulations became problematic during the tracking competition. Visual aids such as presentation slides and handouts would help to reacquaint the rules and regulations. Step-by-step contestant acknowledgement and agreement of rules and regulations during each test would also be helpful.

Difficulty level design

The difficulty of each trial of online registration and tracking strongly depends on the combinations and positions of objects aligned in the competitive environment. As a result, adjusting the difficulty to a more moderate level is difficult. However, multiple challenge points alleviated the problem to some extent. In this competition, millimetric accuracy had some meaning for the first challenge point. As previously mentioned, only one contestant completed the trial. For subsequent challenge points involving 10 to 20 m movement, millimetric evaluation was not important. Instead, how to maintain stable tracking was the chief concern.

Dissemination

By showing a screen capture of each MAR system during each trial on a large screen, the organizers were able to ensure the audience enjoyed the competition. This is also effective at make the competition open and public.

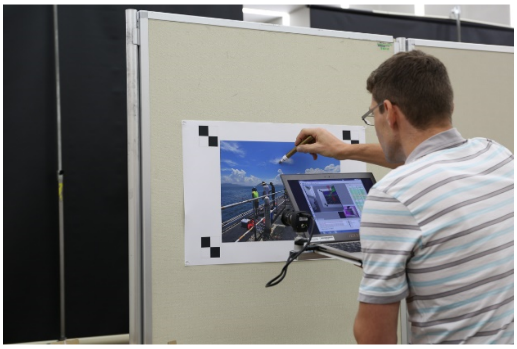

Figure 6: *One of the contestants marking on the textured paper.*

6. Conclusions

This study presented a reference framework on benchmarking of vSRT methods for MAR and summarized findings from benchmarking activities. Although we did not identify or define specific benchmarking frameworks, specific processes, specific formulas of benchmarking indicators, or a specific format of trial sets for benchmarking, our framework can be used by the following individuals:

- A benchmarking service provider, a benchmark provider, or a benchmarking competition organization who wants to align their benchmarking activities including open/closed competitions and self-benchmarking to be consistent with this document;
- A technology developer or supplier who wants to estimate and evaluate performance of a vSRT method for MAR appropriately with a benchmarking service provider, a benchmark provider, or a benchmarking competition organization who aligns their benchmarking activities to be consistent with this document in order;
- A technology user who wants to obtain the benchmarking results based on a benchmarking activity, which is consistent with this document, to compare vSRT methods for MAR in terms of performance.

As in Table 4, we have been making a check sheet to systematically summarize how each benchmarking activity such as an on-site or off-site competition is designed. The check sheet is also assumed to facilitate our confirming the structure of target benchmarking activities. One of future works is to evaluate how the check sheet works in terms of summarization, confirmation, quality control, and so on.

Currently, this framework has been standardized in ISO/IEC JTC 1/SC 24/WG 9 [IW185]. We hope those who are interested in it will contribute to standardization activities.

Moreover, we believe this framework should become the baseline to standardize spatial registration and tracking methods that utilize not only a video camera but combine a video camera with other sensors such as stereo cameras, depth cameras, inertial sensors, infrastructure-based positioning technologies with Internet of Things, or global navigation satellite systems.

Finally, by analogically utilizing this framework, the benchmarking of pedestrian dead-reckoning is also discussed for the standardization [PBS] [PCW17] These efforts will make this framework not only more general but also more adaptive.

			Check	Item	Remarks
Process flow	Process		[ ]	Develop vSRT methods and/or MAR systems:	
			[ ]	Gather vSRT methods and/or MAR systems:	
			[ ]	Prepare and conduct benchmarking:	
			[ ]	Provide and maintain benchmarking instruments:	
			[ ]	Provide and maintain benchmarking repositories:	
			[ ]	Share benchmarking results:	
			T/I/O/S		
	Target (T)/ Input (I)/ Output (O)/ Organized storage (S)		[ ]	vSRT method:	
			[ ]	MAR system:	
			[ ]	Trial sets and physical objects:	
			[ ]	Benchmarking instruments:	
			[ ]	Benchmarking results:	
			[ ]	Benchmarking surveys:	
			[ ]	Benchmarking repository:	
			[ ]	External repositories:	
			Check		
Indicator formura	Reliability		[ ]	3DEVO:	
			[ ]	PEVO:	
			[ ]	Reprojection error of image features:	
			[ ]	Position and posture errors of a camera:	
			[ ]	Completeness of a trial:	
	Temporality		[ ]	Throughput:	
			[ ]	Latency:	
			[ ]	Time for trial completion:	
	Variety		[ ]	Number of datasets/trials:	
			[ ]	Variety on properties of datasets/trials:	
			Check		
Trial set format	Dataset	Contents	[ ]	Image sequences:	
			[ ]	Intrinsic/extrinsic camera parameters:	
			[ ]	Challenge points:	
			[ ]	Optional contents:	
		Metadata	[ ]	Scenario:	
			[ ]	Camera motion type:	
			[ ]	Camera configuration:	
			[ ]	Image quality:	
	Physical object instances	Contents	[ ]	Physical objects:	
		Metadata	[ ]	How to find the physical objects:	

Table 4: *Check sheet for summarization of benchmarking activities.*

References

[COS] The City of Sights: An Augmented Reality Stage Set. http://studierstube.icg.tugraz.at/handheld_ar/cityofsights.php. 5

[I2191] ISO/IEC 29155-1, Systems and software engineering - Information technology project performance benchmarking framework - Part 1: Concepts and definitions. . 2

[IW185] ISO/IEC CD 18520, Information technology-Computer graphics, image processing and environmental representation-Benchmarking framework of vision-based spatial registration and tracking methods for MAR. 7

[KMP*13] KURZ D., MEIER P., PLOPSKI A., KLINKER G.: An outdoor ground truth evaluation dataset for sensor-aided visual handheld camera localization. In *Proc. IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR2013)* (2013), pp. 263-264. 1

[LBM*09] LIEBERKNECHT S., BENHIMANE S., MEIER P., NAVAB N.: A dataset and evaluation methodology for templatebased tracking algorithms. In *Proc. IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR2009)* (2009), pp. 145-151. 1

[MOK14] MAKITA K., OKUMA T., KURATA T.: Benchmarking indicators for AR/MR camera tracking. In *Proc. The 7th Korea-Japan Workshop on Mixed Reality (KJMR 2014)* (2014). 1

[PBS] PDR Benchmark Standardization Committee:. http://www.facebook.com/pdr.bms/. 7

[PCW17] PDR Challenge in Warehouse Picking in IPIN 2017:. http:/www.ipin2017.org/. 7

[PITC] Previous ISMAR Tracking Competitions:. http://ypcex.naist.jp/trakmark/tracking-competition/?page_id=538. 1

[SIKU*10] SHIBATA F., IKEDA S., KURATA T., UCHIYAMA H.: An intermediate report of TrakMark WG. In *Proc. IEEE and ACM International Sympo-sium on Mixed and Augmented Reality (ISMAR 2010)* (2010), pp. 233–236. 1

[TK*09] TAMURA H., KATO H.: Proposal of International Voluntary Activities on Establishing Benchmark Test Schemes for AR/MR Geometric registration and Tracking Methods. In *Proc. IEEE and ACM International Sympo-sium on Mixed and Augmented Reality (ISMAR2009)* (2009), pp. 298–302. 1

[TMC15] Results of ISMAR 2015 Tracking Competition:. http://ypcex.naist.jp/trakmark/tracking-competition/?page_id=1042. 1, 5

[TMK11] Publications in the 2nd international TrakMark workshop (TrakMark 2011):. http://www.trakmark.net/workshop11/index.html. 1

[TMK12] Publications in the 3rd international trakmark workshop (trakmark 2012):. http://studierstube.icg.tugraz.at/ISMARTrackingWorkshop/1. 1

[TMK13] Publications in IEEE ISMAR 2013 Joint Workshop on Tracking Methods & Applications and TrakMark:. http://www.trakmark.net/workshop12/index.html. 1

[TMP14] Publications in TrakMark Panel Session (How to Benchmark AR/MR Geometric Registration and Tracking Methods) in IEEE ISMAR 2014:. http://www.trakmark.net/ismar14/. 1

[TMW] Trakmark web site:. http://www.trakmark.net/. 1

[VTC14] Volkswagen Tracking Challenge in ISMAR 2014:. http://www.tracking-challenge.de/content/tc/content/en/history/challenge_2014/scenarios.html. 5

[WGB13] WILLIAMS S., GREEN R., BILLINGHURST M.: Transform flow: A mobile augmented reality visualization and evaluation toolkit. In *Proc. IVCNZ 2013* (2013). 1