TopoTag: A Robust and Scalable Topological Fiducial Marker System

Guoxing Yu, Yongtao Hu, and Jingwen Dai, *Member, IEEE*

Abstract—Fiducial markers have been playing an important role in augmented reality (AR), robot navigation, and general applications where the relative pose between a camera and an object is required. Here we introduce TopoTag, a robust and scalable topological fiducial marker system, which supports reliable and accurate pose estimation from a single image. TopoTag uses topological and geometrical information in marker detection to achieve higher robustness. Topological information is extensively used for 2D marker detection, and further corresponding geometrical information for ID decoding. Robust 3D pose estimation is achieved by taking advantage of all TopoTag vertices. Without sacrificing bits for higher recall and precision like previous systems, TopoTag can use full bits for ID encoding. TopoTag supports tens of thousands unique IDs and easily extends to millions of unique tags resulting in massive scalability. We collected a large test dataset including in total 169,713 images for evaluation, involving in-plane and out-of-plane rotation, image blur, different distances, and various backgrounds, etc. Experiments on the dataset and real indoor and outdoor scene tests with a rolling shutter camera both show that TopoTag significantly outperforms previous fiducial marker systems in terms of various metrics, including detection accuracy, vertex jitter, pose jitter and accuracy, etc. In addition, TopoTag supports occlusion as long as the main tag topological structure is maintained and allows for flexible shape design where users can customize internal and external marker shapes. Code for our marker design/generation, marker detection, and dataset are available at http://herohuyongtao.github.io/research/publications/topo-tag/.

Index Terms—Fiducial marker, monocular pose estimation, topological information, marker design, ID decoding

1 Introduction

In this paper, we introduce TopoTag, a new fiducial marker and detection algorithm that is more robust and accurate than current fiducial marker systems. Fiducial markers are artificial objects (typically paired with a detection algorithm) designed to be easily detected in an image from a variety of perspectives. They are widely used for augmented reality and robotics applications because they enable localization and landmark detection in featureless environments [1]. Previous work on fiducial markers mainly focus on one or more of the following areas: (1) improving detection accuracy via specialized tag design [2], [3], [4], [5]; (2) reducing pose estimation error via precise vertex estimation [6] or introducing more feature points [7]; (3) increasing unique identities [8], [9], [10], [11]; (4) improving robustness under occlusion [7], [12] and other use cases [13], [14], [15], [16] and (5) speed-up [1], [6], [17], [18].

TopoTag utilizes topological information in tag design to improve robustness, which achieves perfect detection accuracy on the large dataset we collected and on datasets from others. We show that all tag bits can be used to encode identities without sacrificing detection accuracy, thus achieving rich identification and massive scalability. In addition, TopoTag offers more feature point correspondences for better pose estimation. Results show that TopoTag achieves the best performance in vertex jitter, pose error and pose jitter. TopoTag also supports occlusion and noise, to some extent, if the main tag topological structure is maintained and supports flexible shape design where users can customize internal and external marker shapes. Fig. 1 shows three TopoTag markers.

We collected a large dataset including 169,713 images with TopoTag and several state-of-the-art tag systems. A robot arm is used to make sure each tag has the same trajectory for consistent comparison. The rich modalities of the dataset include in-plane and out-of-plane rotations, image blur, different distances and various backgrounds, etc. which offer a challenging benchmark evaluation.

In summary, the contributions of this paper are: (1) we present TopoTag, a topological-based fiducial marker system and detection algorithm; (2) we demonstrate that TopoTag achieves the best performance in various metrics including detection accuracy, localization jitter and accuracy, etc. while at the same time supports occlusion, noise and flexible shapes; (3) we show that it's possible in tag design to use full bits for ID encoding without sacrificing detection accuracy, thus achieving scalability; and (4) we collect a large dataset of various tags, involving in-plane and out-of-plane rotation, image blur, different distances and various backgrounds, etc.

The remainder of the paper is organized as follows: In Section 2, we discuss related work in different marker patterns. We introduce the TopoTag design and detection algorithm in Sections 3 and 4 respectively. Dataset and

• *The authors are with the Guangdong Virtual Reality Technology Co., Ltd. (aka. Ximmerse), Shenzhen, Guangdong 518000, China. E-mail: {calvin.yu, ythu, dai}@ximmerse.com.*

Manuscript received 21 Aug. 2019; revised 7 Apr. 2020; accepted 11 Apr. 2020. Date of publication 20 Apr. 2020; date of current version 29 July 2021. (Corresponding author: Yongtao Hu.)
Recommended for acceptance by K. Kiyokawa.
Digital Object Identifier no. 10.1109/TVCG.2020.2988466

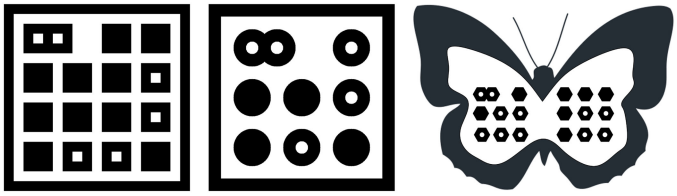


Fig. 1. Three TopoTag markers. TopoTag supports both customized internal and external shapes. Here shows three TopoTags with various internal shapes like squares, circles, hexagons, and different external shapes like square and butterfly.

experimentation are discussed in Section 5. Section 6 is devoted to the conclusions.

2 Related Work

Fig. 2 shows many different fiducial marker systems discussed in this section.

Circular Patterns. Among the earliest work, Gatrell *et al.* [19] propose to use concentric contrasting circle (CCC) for fiducial marker design. It's further enhanced in [8] by adding colors and multiple scales. In [20], [21], dedicated data rings are added to the marker design for rich identification. Sattar *et al.* [22] and Xu *et al.* [23] propose FourierTag with a frequency image as the signature. In RuneTag [7], [24] and Pi-Tag [26], they propose using rings of dots to improve robustness to occlusion and provide more points for pose estimation. CCTag [14], [25] and followed work by Prasad *et al.* [13] use multiple rings to increase robustness to blur and ring width for encoding. Circular patterns, e.g., RuneTag, provide the state-of-the-art for most identities. However, the tracking distance is usually limited due to their requirement of finding enough confident ellipses. In comparison, TopoTag can provide even more identities while at the same time offering much larger tracking range.

Square Patterns. To be easily localized, most fiducial systems are designed to contain a thick square border. Matrix [27], CyberCode [29] and VisualCode [30] are the first and simplest proposals. ARToolkit [28] is well known and widely used in many augmented reality applications. It includes a pattern in their internal region for identification via image correlation. ARTag [2] and ARToolkitPlus [31] improve the recognition technique with a binary coded pattern. In addition, they are designed with an error correction mechanism to increase robustness. BinARyID [9] proposes a method to generate markers that attempt to avoid rotation ambiguities. Schweiger *et al.* [33] propose using SIFT and SURF filters that are specifically designed for SIFT and SURF detectors. Tateno *et al.* [32] propose using nested markers to improve performance under different distances. Several works investigate using multiple fiducial markers in a checkerboard to improve camera calibration [4] and reduce the perspective ambiguity by further adding color [10]. AprilTag [5], [6] is a faster and more robust reimplementation of ARTag. Garrido-Jurado *et al.* [11], [12], [18] propose ArUco using mixed integer programming to generate markers. ChromaTag [1] uses color over AprilTag to improve marker detection speed. Square patterns are most popular among practical applications due to this technique's detection robustness and large tracking range. However, some encoding bits must be reserved to handle rotation ambiguities and incorporate Hamming distance strategy. In contrast, TopoTag can provide much richer identities by encoding full bits while at the same time achieving the state-of-the-art robustness and tracking range. Moreover, unlike square markers using four corner points for pose estimation (which is the minimum number for unambiguous pose estimation [40]), TopoTag offers better pose estimation utilizing all vertices of tag bits. It's worth noting that [10] shows the possibility of reducing rotation ambiguities, increases rich identities by adding color information and achieves better pose accuracy by using more inner corners. However, it still needs to reserve some bits for error detection and correction. In comparison, TopoTag

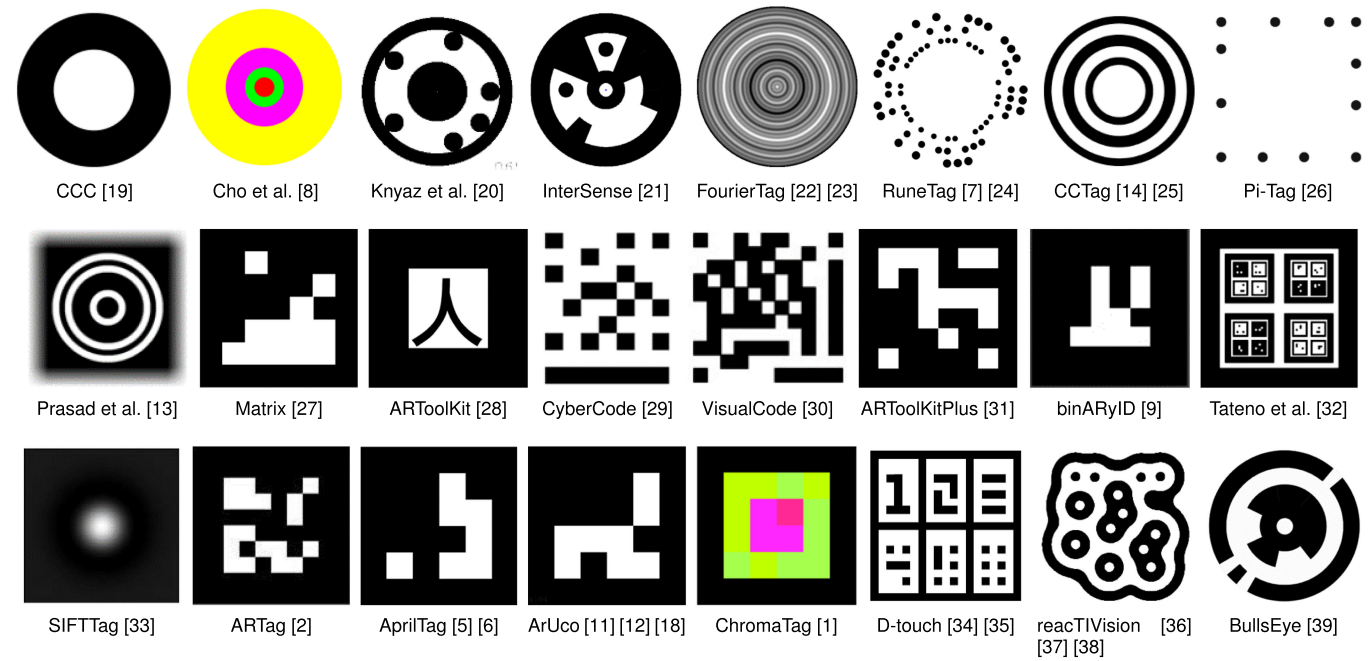


Fig. 2. Existing fiducial marker systems.

offers even richer identities without using color due to the unique baseline node design and can utilize more feature correspondences for better pose estimation.

Topological Patterns. D-touch [34], [35] is the earliest work to use topological patterns in tag design. Marker detection is based on the region adjacency tree information. D-touch employs a single topology for all markers in the set and does not provide a specific method for computing location and orientation. ReacTIVision [36], [37], [38] improves over D-touch and provides unique identities purely with the topological structure by building a left heavy depth sequence of the region adjacency graph. BullsEye [39], which is specially optimized for GPU, consists of a central white dot surrounded by a solid black ring and one or more data rings again surrounded by a solid white ring inside a black ring with three white studs. Topological patterns demonstrate the ability to improve robustness using topological information. However, they (including ReacTIVision and BullsEye) can only recover 2D location and orientation due to the lack of sufficient matched feature points. In comparison, TopoTag offers accurate 3D pose estimation and state-of-the-art robustness at the same time.

Machine Learning. Claus *et al.* [41], [42] use trained classifiers to improve detection in cases of insufficient illumination and blurring caused by fast camera movement. Randomized forests are also used to learn and detect planar objects [43], [44]. Machine learning methods show the potential to detect natural objects. However, in practice, these algorithms do not achieve detection accuracies on par with detection algorithms specifically designed for marker detection [1]. In contrast, TopoTag achieves the state-of-the-art detection accuracy over machine learning and other previous types of patterns.

3 TopoTag Design

TopoTag utilizes topological structure information in tag design. This method has been validated with proven increases in robustness across illumination variation and a reduction in false detection [34]. Existing fiducial marker systems, especially with square patterns, sacrifice tag encoding bits to handle rotation ambiguities during decoding [10]. Additional bits will also be reserved for incorporating Hamming distance strategy in order to improve false positive rejection. Strong robustness with topological design helps by saving tag bits for encoding identities. To avoid rotation ambiguities, TopoTag introduces *baseline node* in its topological structure. The baseline node is specially designed to be different from other nodes in the tag. TopoTag uses a black node with two white children nodes inside as the baseline node and other black nodes, with at most one white child node, as *normal nodes*. Note that, baseline node can be defined with other forms. For example, it can be defined with three or more white children nodes for different needs. Baseline node defines the search starting position of the whole tag, thus avoiding checking rotation ambiguities. All normal nodes are used for identity encoding with 0 denoting no child node and 1 otherwise. The identity encoding for the two markers shown in Fig. 3 is $0000000 = 0$ and $1111111 = 127$ respectively.

For pose estimation, instead of using only four border points in previous square systems [1], [2], [5], [6], [11], [12], [31] which is the minimum number required, TopoTag offers more point correspondences resulting in more accurate pose estimation. Baseline node (more specifically its two children nodes) and all normal nodes are all employed as feature points, thus achieving a better pose estimation.

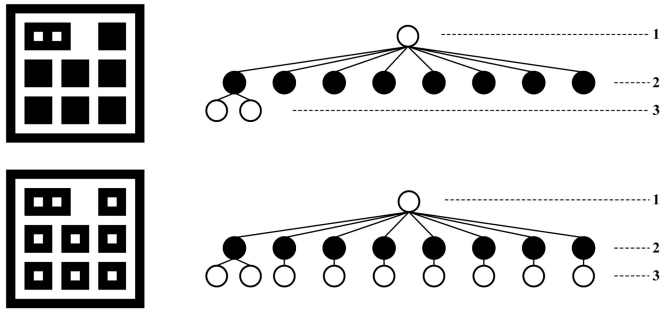


Fig. 3. Topological tree of two TopoTags. Each node in the topological tree denotes one TopoTag connected component (starting from the inner biggest white connected component). Except the two white nodes inside the baseline node, all leaf nodes are used for identify encoding. The identity encodings for these two markers are $0000000 = 0$ and $1111111 = 127$ respectively.

Note that, as TopoTag design is based on topological information, there is no restriction for the shapes used in the tag. Both internal and external shapes can be customized as long as the desired topological structure is preserved. Fig. 1 shows three different design samples of TopoTag. For easy searching and model simplicity, in current TopoTag design, we place all internal nodes uniformly spaced and compacted into a $n \times n$ squared shape.

4 TopoTag Detection

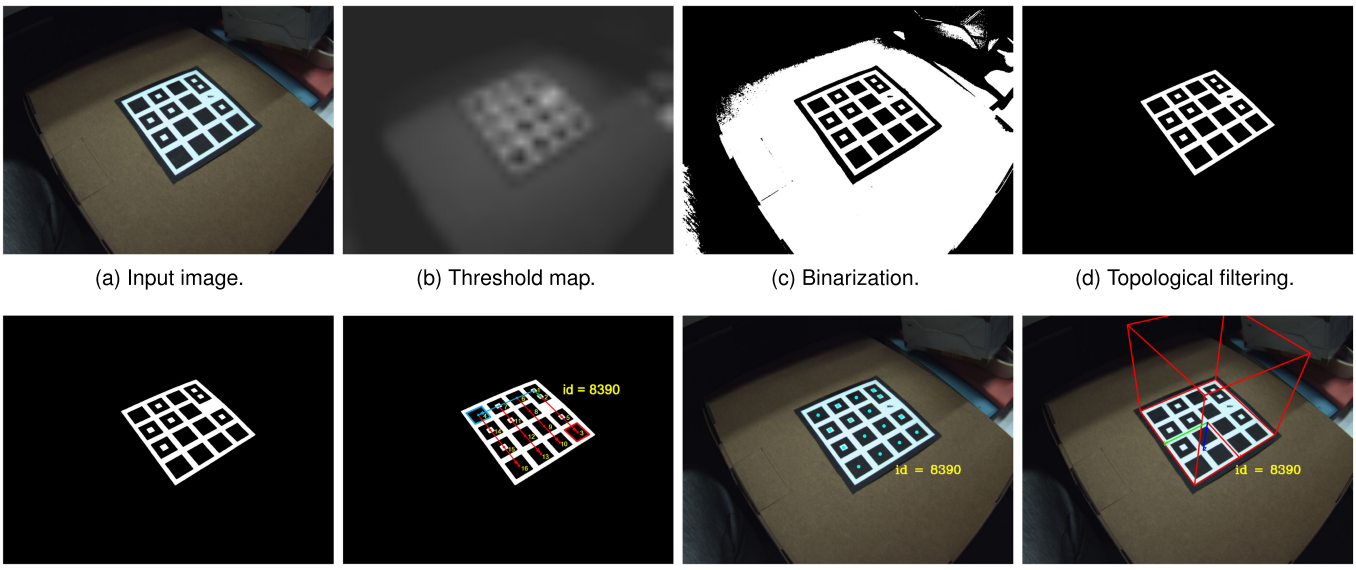
Fig. 4 outlines main steps of TopoTag detection. Topological information is extensively used for 2D marker detection, and further corresponding geometrical information for ID decoding. 3D pose estimation is achieved by taking advantage of all TopoTag vertices.

4.1 2D Marker Detection

Threshold Map Estimation. Similar to the idea of adaptive thresholding, we estimate the threshold for each pixel by analyzing its neighboring pixels. The analysis can be conducted on the original image, however, in order to deal with the image noise and blur in real applications, analyzing a downsampled image (scalar s_1) is more accurate, which also brings speed benefits. Any pixel will be set to α if its value is less than α to remove pixels that are too dark. Average values are computed on a local region (window size w) on the downsampled image. To further handle the image noise, the downsampled average map can be further downsampled (scalar s_2). The final threshold map is achieved by upsampling the downsampled average map by $s_1 \times s_2$ using bilinear interpolation, see Fig. 4b.

Binarization. Binarization is achieved by comparing the input image with the threshold map. A minimum brightness (β) is set to filter regions that are too small (i.e., set to black if pixel value is less than β). See Fig. 4c for an example of binarization result.

Topological Filtering. After the binarization, we build the topological tree of the connected binary regions. To find candidate tags, we search the tree based on two conditions:



(a) Input image. (b) Threshold map. (c) Binarization. (d) Topological filtering.

(e) Error correction. (f) Decoding. (g) Vertex estimation. (h) Pose estimation.

Fig. 4. Main steps of TopoTag detection. (Best viewed in color).

(1) the number of children nodes should be within $[\zeta_{\min} - \tau, \zeta_{\max} + \tau\ ]$, where $\zeta_{\min}$ is the number of nodes for tag ID $= 0$ with all black leaves except the baseline node and $\zeta_{\max}$ for the tag with maximum ID with no black leaves, and τ is the tolerance level allowed; (2) max depth of the tree should be exactly 3. See Fig. 3 for examples of the topological trees for both $\zeta_{\min}$ and $\zeta_{\max}$ cases of 9-bit TopoTags. Fig. 4d shows the result after the topological filtering.

Error Correction. There are possible error nodes within the tag region due to noise or occlusion. Fig. 4d shows an example of one error node close to the baseline node because of one ant sitting on the tag. To correct these error nodes, we first compute the area of the baseline node and then filter out smaller nodes if their areas are less than $\theta_1\%$ of the baseline node area. Fig. 4e shows the result after error correction.

4.2 ID Decoding

To decode ID, we need to determine the node sequence and map it to a binary code string. Take a 16-bit TopoTag as an example, see Fig. 5 and 4f of the sequence where we find for each node of the tag. To start, we first find the baseline node (including p1 and p2) and determine its search direction based on whether there are nodes along the direction with angel tolerance θ_2, i.e., p1 $\rightarrow$ p2. Along the direction, we find the node with the largest distance, i.e., p3. For the remaining nodes, we first find the node with largest angle against the baseline direction p1 $\rightarrow$ p2 and then the largest distance along the direction, i.e., p4. p5 is determined along direction p1 $\rightarrow$ p3, p6 and p7 along p1 $\rightarrow$ p4. The remaining nodes are determined in order and in a similar way. After finding each node, we can simply map each node to 1 or 0 depending on whether it contains a white child node or not and then decode the tag based on the binary code string. For the example shown here and in Fig. 4, the binary code string is 10000011000110, which is decoded with ID $= 8390$. It's worth noting that ID decoding is processed on the images after removing the perspective distortion in which lines will still be lines in images with no distortion to improve the robustness of direction searching.

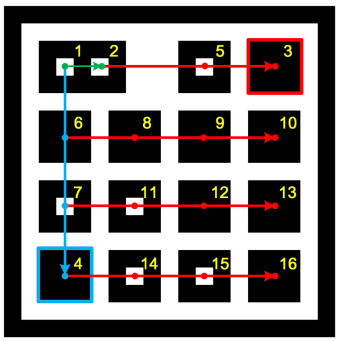


Fig. 5. Vertex decoding order. (Best viewed in color).

4.3 3D Pose Estimation

For each node, we estimate the vertex by computing the centroid on the original image of its supporting region. The supporting region can be the binary mask or its dilated version (with dilate size δ). The centroid can be determined via image moments, i.e., $\{\bar{u}, \bar{v}\} = \{\frac{M_{10}}{M_{00}}, \frac{M_{01}}{M_{00}}\}$.

For pose estimation, the exact correspondence between the 2D image features and the features of the associated model is needed (feature correspondence). At least four points are needed to recover unambiguous pose estimation for planar tags [40]. Unlike most of previous work using only four corner points, all TopoTag vertices of tag bits are used for a better pose estimation. As reported in [45], a larger number of feature correspondences consistently leads to lower error and better pose estimation to noise for various PnP methods. We refer the reader to [46] for a detailed analysis on the stability of homography estimation by 1st-order perturbation theory. For 16-bit tag, 16 vertex correspondences are used, including two baseline white nodes and 14 normal black nodes. 6-DoF pose estimation is achieved by solving the PnP problem and Levenberg-Marquardt algorithms [47], [48] based on these feature correspondences.

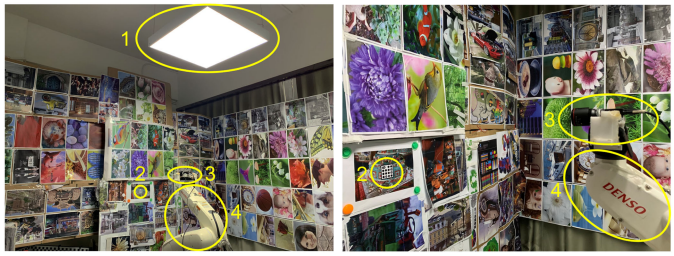


Fig. 6. Dataset collection setup. We collect dataset by putting tags (label #2) in a rich textured background of an indoor environment with fixed lighting (label #1). The camera (label #3) is fixed to a robot arm (label #4) to ensure the same trajectories for different tags.

Fig. 7. Sample images from the dataset. Images are from Seq #1 (with ARToolKit), Seq #2 (with AprilTag 25h9) and Seq #3 (with TopoTag) from left to right respectively.

5 Results and Discussion

Algorithm Setup. Throughout the experiment, we use $s_1 = 4, s_2 = 8, w = 5, \alpha = 45, \beta = 50$ for segmentation, $\tau = 0, \theta_1 = 30, \theta_2 = 0.1$ rad for decoding, $\delta = \max\{2, \lfloor \frac{l}{10} \rfloor\}$ for vertex estimation, where l is the short length of the binary mask region.

All of the experiments have been performed on a typical laptop PC equipped with an Intel Core i7-7700HQ processor (8 cores @2.8 Ghz) and 8 GB of RAM.

5.1 Dataset

The previous work, like [5], [7], mainly focused on evaluating performance on synthetic images. Although some of the work evaluated parts of the performance on more realistic scenes, e.g., ARToolKitPlus [31] evaluates the speed on several handheld devices and AprilTag [5], [6] evaluates false positive on LabelMe [49] dataset which is designed for general object detection and recognition research, there is still no uniform dataset for fiducial marker evaluation. This makes it difficult to reproduce the result and compare with others. More recently, in ChromaTag [1] work, they collected a dataset to compare their work with AprilTag [5], CCTag [14], and RuneTag [7]. However, different tags are placed side-by-side during their dataset collection, thus it is not ideal for comparison, especially when tags viewed from a large angle as different markers will have different distances and facing angles towards the camera.

In this work, we try to fill this gap by collecting a large dataset, including a total of 169,713 images, which include in-plane and out-of-plane rotations, image blur, various distances and cluttered backgrounds, etc. Please refer to the supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TVCG.2020.2988466, for details of our dataset variations. We use an industrial camera with a global shutter that has 1280×960 resolution streaming at 38.8 fps and 98 ° diagonal field of view. The exposure time is fixed at 10 ms. Using relatively long exposure guarantees sufficient brightness of the captured images, which at the same time introduces the image blur phenomenon for more challenging use cases (see the first image in Fig. 7 for an example). The camera is fixed to a robot arm[1] to ensure the same trajectories for different tags. Fig. 6 shows the dataset collection setup. Three sequences will be collected for each tag, and the trajectory for each sequence is shown in Fig. 8. In all the three sequences, the camera keeps facing the front as shown in the first image of Fig. 8. In Seq #1, the camera moves along several lines at a constant speed, with different out-of-plane rotations for each line including 0 ° (i.e., camera faces the tag right ahead), 30 ° and 60 °. In Seq #2, the camera moves back and at the same time rotates in-plane within 0-180 ° at a constant speed back and forth. Note that, as we can only rotate around the end joint of the robot arm and there is an offset between the camera and arm, the camera's trajectory will not be an ideal half circle. In Seq #3, the camera is placed at 10 fixed positions (P1 → P10). Besides 0 °, 30 ° and 60 ° out-of-plane rotations as in Seq #1, we further collect data with 75 ° (P1 and P10). In all three sequences, the background is filled with rich textured images to simulate more complex use scenarios.

We collect the dataset for TopoTag and previous tags including ARToolKit [28], ARToolKitPlus [31], ArUco [12], RuneTag [7], ChromaTag [1] and AprilTag [6]. A 16-bit TopoTag is used throughout the experiment as it provides the most unique identities, see Table 1 for details. And, without loss of generality, the tag comes with square internal and external shapes (see the first image in Fig. 1). For systems with multiple tag families, we collect data for each tag family, including 16h3, 25h7, 36h12 for ArUco and 16h5, 25h7, 25h9, 36h9, 36h11 for AprilTag. For each tag family (including TopoTag), we randomly select one ID for evaluation. In our experiment, we randomly selected ID = 1 for ARToolKit, 262 for ARToolKitPlus, [104, 90, 136] for ArUco's [16h3, 25h7, 36h12], 107 for RuneTag, 0 for ChromaTag, [0, 204, 25, 1314, 343] for AprilTag's [16h5, 25h7, 25h9, 36h9, 36h11] and 278 for TopoTag. Note that, for AprilTag, there are three shared tag families, i.e., 16h5, 25h9 and 36h11, for AprilTag-1 [5] and AprilTag-2 [6], and 25h7, 36h9 only exist in AprilTag-1. In following sections, we will report the best result of AprilTag-1 and AprilTag-2 for these shared tag families if not otherwise specified. For evaluation fairness, outer border sizes of all tags are kept at the same of 5 cm. For each tag, there are ≈ 100,000 images collected, including ≈ 1,000 for Seq #1, ≈ 1,200 for Seq #2 and ≈ 7,800 for Seq #3. Please see Fig. 7 for sample images for each sequence.

It's worth noting that segmentation is crucial for marker detection and pose estimation for all marker systems. Thus, for fair comparison, we fine tune the segmentation parameters for each marker algorithm unless it already uses advanced approaches like adaptive thresholding, line detection, etc. Specifically, we use a threshold of 60 instead of default 100 for ARToolKit, 15 and 2 for `AdaptiveThresholdWindowSize` and `AdaptiveThresWindowSize_range` instead of default -1 and 0 for ArUco. Please refer to the supplementary material, available online, for the performance comparison between their default setups and our finely tuned versions.

1. We use a robot arm from DENSO (VS-6556). Link: https://www.denso-wave.com/en/robot/product/five-six/vs.html

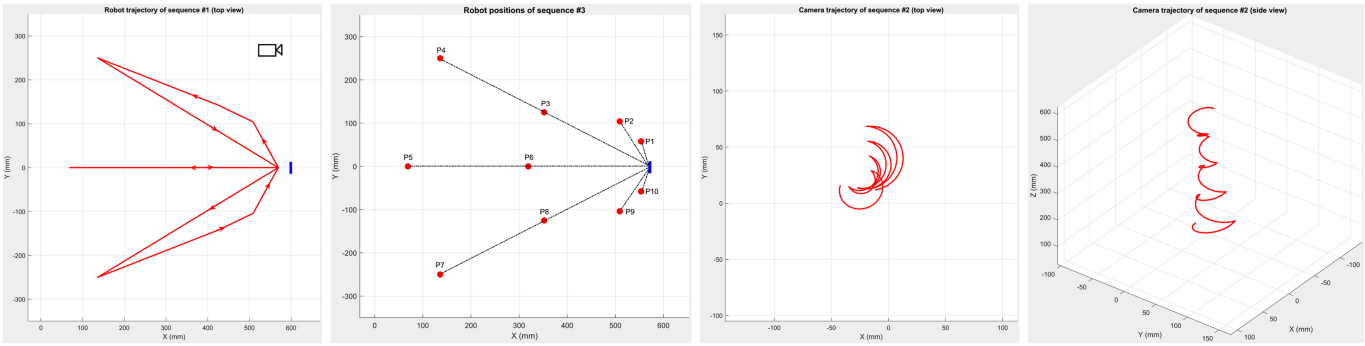


Fig. 8. Robot arm trajectory/points in different sequences (1st image for Seq #1, 2nd image for Seq #3). Camera trajectory is shown for Seq #2 for better visualization (3rd and 4th images). Tag position is shown in blue. (Best viewed in color).

TABLE 1
Dictionary Size Versus Tracking Distance

Tag	Dictionary Size	Min Distance (m)	Max Distance (m)
ARToolKit	10^2	0.047	1.199
ARToolKitPlus	512	0.087	1.154
ArUco (16h3)	250	0.117	1.309
ArUco (25h7)	100	0.117	1.187
ArUco (36h12)	250	0.120	1.199
RuneTag	17,000	0.103	0.221
ChromaTag	30	0.547	0.560
AprilTag (16h5)	30	0.161 → 0.043	1.220 → 0.757
AprilTag (25h7)	242	0.160	1.171
AprilTag (25h9)	35	0.156 → 0.040	1.226 → 0.968
AprilTag (36h9)	5,329	0.163	1.223
AprilTag (36h11)	587	0.163 → 0.042	1.168 → 0.906
TopoTag (3x3)	128	0.029	1.204
TopoTag (4x4)	16,384	0.029	1.055
TopoTag (5x5)	8,388,608	0.029	0.670

For shared tag families of AprilTag-1 and AprilTag-2, results of both versions are reported with format "AprilTag-1 → AprilTag-2".

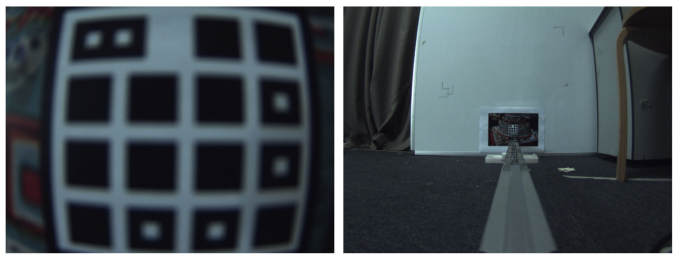


Fig. 9. Images of TopoTag at minimum and maximum tracking distance respectively. Note that, there are partial marker occlusion because of the camera FoV limitation and blur issues especially at the min distance.

5.2 Dictionary Size Versus Tracking Distance

Table 1 shows the comparison of dictionary size versus tracking distance (both min and max) of different tag systems. Generally speaking, more tag bits offer more spaces to encode identities, but sacrifice maximum tracking distance as region for each bit becomes smaller. On the other hand, minimum tracking distance is affected by the marker occlusion because of the camera FoV limitation and blur issue at the small range of a fixed-focus camera. Fig. 9 shows the images of TopoTag at minimum and maximum tracking distance respectively. TopoTag achieves a state-of-the-art minimum tracking distance, which further demonstrates the robustness of TopoTag under partial occlusion and out-of-focus image blur. TopoTag also achieves comparable maximum tracking range when the dictionary size is small (9-bit), while offering a significantly larger tracking range when the dictionary size extends to tens of thousands (16-bit versus RuneTag with the state-of-the-art most identities). In addition, TopoTag offers the scalability of extending the dictionary size to millions with a still acceptable tracking distance (25-bit). Interestingly, AprilTag-2 achieves better minimum tracking range but worse maximum tracking range over AprilTag-1 by a large margin. We suspect that this is due to the tag detection strategy change from gradient computing based in AprilTag-1 to adaptive thresholding based in AprilTag-2 for speedup. It's worth noting that, we also tested markers with different sizes, including 2.5 and 10 cm. The conclusions for both min and max tracking ranges still hold.

5.3 Detection Accuracy

Table 2 summarizes the detection results for TopoTag compared to previous marker systems. Fig. 10 highlights the recall and precision of different captured points on Seq #3. We follow the metrics used in [1]. True positives (TP) are defined as when the tag is correctly detected, including locating the tag and correctly identifying the ID. Correct identification of the tag is determined by having at least 50 percent intersection over union between the detection and the ground truth. False positives (FP) are defined as detections returned by the detection algorithms that do not identify the location and ID correctly. False negatives (FN) are defined as any marker that is not identified correctly. Precision is $\frac{TP}{TP+FP}$ and recall is $\frac{TP}{TP+FN}$.

Table 2 shows that TopoTag performs perfectly on all three sequences, achieving 100 percent across both recall and precision. All tested marker systems, except ChromaTag, work great and achieve $> 99.5\%$ on precision due to their unique false positive rejection techniques. However, most systems except ARToolKit, ARToolKitPlus and ArUco fail to achieve a high recall, i.e., $< 81\%$. Fig. 10 shows that all previous systems degrade on recall or precision or both when markers are viewed from wide angles. This result is probably from large distortion, decreased lightness and

2. 10 tags are provided in ARToolKit package. Theoretically, any pattern can be used for tag design, but the author didn't provide the approach.

TABLE 2
Detection Accuracy (With Run Time)

Tag	Recall (%)	Precision (%)	Time (ms)
ARToolKit	99.990	99.880	5.864
ARToolKitPlus	98.297	100.000	9.314
ArUco (16h3)	100.000	99.910	54.319
ArUco (25h7)	99.009	100.000	53.930
ArUco (36h12)	99.470	100.000	56.001
RuneTag	0.281	100.000	455.832
ChromaTag	9.088	9.190	9.103
AprilTag (16h5)	77.285	99.883	246.762 → 15.114
AprilTag (25h7)	75.711	100.000	244.433
AprilTag (25h9)	80.405	100.000	251.275 → 13.603
AprilTag (36h9)	78.704	100.000	240.694
AprilTag (36h11)	100.000	99.990	241.314 → 13.431
TopoTag	100.000	100.000	33.638

For shared tag families of AprilTag-1 and AprilTag-2, run time of both versions are reported with format "AprilTag-1 → AprilTag-2".

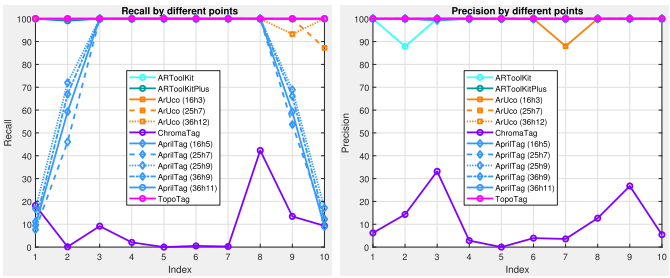


Fig. 10. Recall and precision by different points on Seq #3. (Best viewed in color).

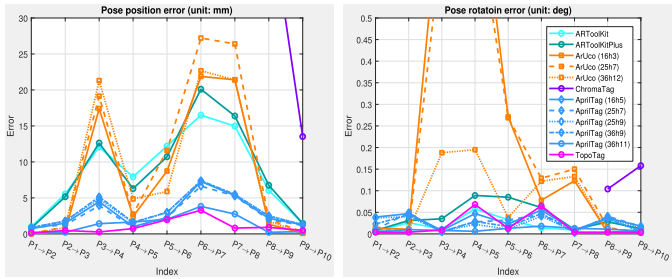


Fig. 11. Pose position (left) and rotation (right) error comparison. We have trimmed the figures for better visualization. Please refer to the supplementary material, available online, for full figures. (Best viewed in color).

TABLE 3
Average and Maximum Pose Errors of Each Tag

Tag	position (mm)		rotation (deg)	
	avg	max	avg	max
ARToolKit	8.639	16.499	0.022	0.058
ARToolKitPlus	8.923	20.101	0.040	0.089
ArUco (16h3)	8.191	21.876	0.248	0.908
ArUco (25h7)	10.049	27.212	0.225	0.765
ArUco (36h12)	8.768	22.663	0.078	0.195
ChromaTag	29.586	45.643	0.131	0.158
AprilTag (16h5)	2.894	7.287	0.031	0.055
AprilTag (25h7)	2.704	6.641	0.026	0.049
AprilTag (25h9)	3.178	7.320	0.024	0.041
AprilTag (36h9)	3.228	7.394	0.024	0.047
AprilTag (36h11)	1.402	3.824	**0.010**	**0.018**
TopoTag	**1.011**	**3.289**	0.019	0.068

Best results are shown in bold and underlined.

blur issues, which distract marker detection. TopoTag, on the other hand, has no obvious degradation on these issues.

ChromaTag performs worse on both recall and precision possibly due to the cluttered colored background and relative low brightness of collected images distracting its detection based on color information. For ablation study, we tried to replace all ChromaTag's background with pure white pixels and keep only the marker region. With such a setup, ChromaTag achieves the same recall (i.e., same number of false negatives) and the number of false positives decreased from 8,962 to 257, which further validates that ChromaTag is sensitive to cluttered background (i.e., detecting false positives).

RuneTag performs the worst with lowest recall $< 0.3\%$ and fails to detect any frame on Seq #3 where it fails to find enough confident ellipses on the images. As also found in ChromaTag work [1], RuneTag requires larger tag sizes for detection, which is the major cause of its lesser performance on our dataset with small marker size in long distance and challenging blur. In our experiment, we found that RuneTag cannot be detected when the marker is smaller than 180×180 pixels.

False Positive Rejection. Since all of the images in our dataset contain valid tags, FP mainly focuses on the background excluding the tag regions. To better evaluate FP, as in [6], we further run the experiment on LabelMe [49] dataset, which consists of 207,883[3] images of natural scenes from a wide variety of indoor and outdoor environments, none of which contain any valid fiducial markers. We run this test for ARToolKit, ARToolKitPlus, ArUco, AprilTag and TopoTag as they achieve top detection accuracy results on our dataset as shown in Table 2. In addition, we further run this test for reacTIVision [36], [37], [38] which only recovers 2D location and orientation by default. There are 49,321 false positives returned by AprilTag (16h5), 9,756 by ARToolKit, 348 by reacTIVision and 146 by ArUco (16h3). In contrast, TopoTag and ARToolKitPlus both have no false positives.

3. This is the latest LabelMe dataset size, which is slightly different the size of 180,829 from that was used in [5], [6].

5.4 Localization Jitter and Accuracy

We evaluate localization jitter (including 6-DoF pose jitter and 2D vertex jitter) and accuracy (i.e., 6-DoF pose accuracy) on Seq #3.

5.4.1 *Pose Error*

We evaluate the accuracy between each point and its adjacent point. The robot's measurements serve as the groundtruth. Since there are in total 10 points in Seq #3, nine accuracy values will be computed. See Fig. 11 for the results of both position and rotation accuracies. Average and maximum pose errors for each tag are listed in Table 3. TopoTag outperforms all previous systems in position error by a large margin (about 28 percent error reduction by average and 14 percent by max compared to the 2nd best) and is comparable with the state-of-the-art on rotation error (< 0.1 degree for both average and max). A further two-sample Kolmogorov-Smirnov test shows that TopoTag significantly outperforms the 2nd best (i.e., AprilTag) in position error with $p = 0.000$.

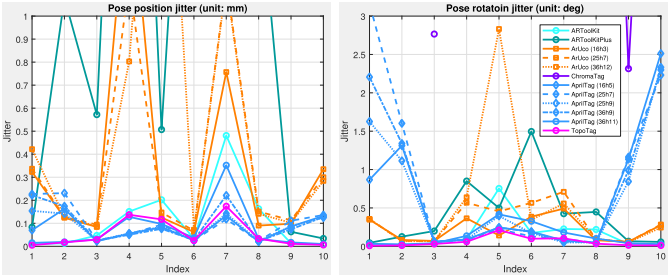


Fig. 12. Pose position (left) and rotation (right) jitter comparison. We trim the figures for better visualization. Please refer to the supplementary material, available online, for full figures. (Best viewed in color).

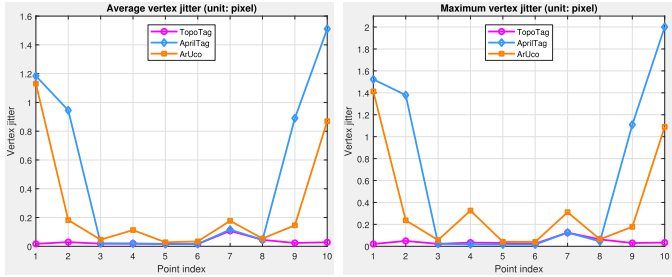


Fig. 13. Average and maximum vertex jitter comparison by different points on Seq #3. (Best viewed in color).

TABLE 4
Average and Maximum Pose Jitters of Each Tag

Tag	position (mm)		rotation (deg)	
	avg	max	avg	max
ARToolKit	0.112	0.481	0.160	0.754
ARToolKitPlus	1.134	3.584	0.421	1.496
ArUco (16h3)	0.363	1.636	0.230	0.491
ArUco (25h7)	0.364	1.155	0.322	0.710
ArUco (36h12)	0.573	2.553	0.526	2.832
ChromaTag	49.880	130.958	8.479	14.616
AprilTag (16h5)	0.079	0.163	0.654	2.512
AprilTag (25h7)	0.104	0.231	0.879	3.160
AprilTag (25h9)	0.087	**0.154**	0.673	2.333
AprilTag (36h9)	0.102	0.222	0.753	2.299
AprilTag (36h11)	0.074	0.352	0.133	0.416
TopoTag	**0.055**	0.173	**0.058**	**0.211**

Best results are shown in bold and underlined.

5.4.2 Pose Jitter

Both position and rotation jitters are evaluated at each point using the standard deviation (STD) metric. See Fig. 12 for the result. Average and maximum jitter for each tag can be seen in Table 4. TopoTag outperforms all previous systems in rotation jitter by a significant margin (about 56 percent average jitter reduction and 49 percent by max compared to the 2nd best). This result is comparable with the state-of-the-art on position jitter ($<$ 0.1 mm for average and $<$ 0.2 mm for max). A further two-sample Kolmogorov-Smirnov test shows that TopoTag significantly outperform the 2nd best (i.e., AprilTag) in rotation jitter with $p = 0.000$.

5.4.3 Vertex Jitter

Vertex jitter measures the noise of the 2D feature point estimation, whose errors will propagate to the estimation of the 6-DoF pose. To evaluate vertex jitter, we compare two of the best previous methods, AprilTag and ArUco. Both AprilTag and ArUco are square markers, which use intersections of quad lines to achieve sub-pixel vertex precision. RUNE-Tag and ChromaTag are not evaluated as they fail to reliably detect all positions in Seq #3, i.e., the number of detected frames for a point is less than 50.[4] Square markers, like ARToolKitPlus and ChromaTag, theoretically will have similar performance as AprilTag and ArUco. ARToolKit is not evaluated as it uses correlation against a database to detect instead of finding fixed corners. All candidate methods are evaluated on markers with 16 bits (i.e., AprilTag's 16h5 and ArUco's 16h3). Similar to pose jitter evaluation, STD metric is used.

Results can be seen in Fig. 13. It is evident that TopoTag performs consistently the best or comparable to the state-of-the-art across all points, especially when the marker angles become greater (e.g., $\geq 60°$) and with more image blur (see P1, P2, P9 and P10). AprilTag performs better than ArUco where marker angles are relatively small ($\leq 30°$, see P3-P8) thanks to its edge refinement, but become worse where the marker has larger angle w.r.t. the camera.

4. ChromaTag fails to reliably detect P2, P4, P5, P6 and P7; and all 10 positions are failed for RUNE-Tag.

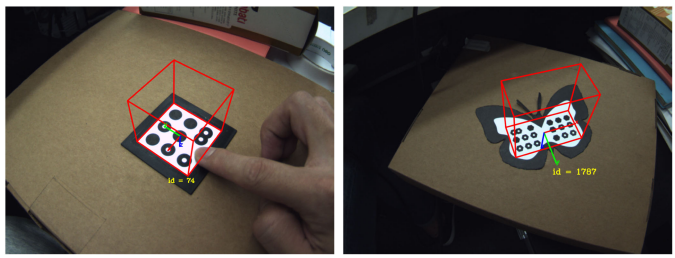


Fig. 14. Detection and pose estimation of two customized TopoTags. (Best viewed in color).

5.5 Speed

5.5.1 Dictionary Computation

Dictionary computation is usually a time-consuming operation due to the specially designed lexicode generation algorithm and Hamming distance strategy required to achieve high detection robustness. Although there is no need to do dictionary computation online normally, it's still meaningful to make this step efficient enough. ArUco takes approximately 8, 20 and 90 minutes respectively for dictionaries of sizes 10, 100 and 1000 [12], while it can take several days to generate 36-bit tags for AprilTag [5]. As TopoTag supports full tag bits for identity encoding, it is extremely fast for dictionary computation as an ID can be directly mapped to the binary code string. In our experiment, it takes only 4.1 seconds to generate dictionary of size 8,388,608 (i.e., TopoTag-5 $\times$ 5).

5.5.2 Tag Detection

The last column of Table 2 shows the running time comparison. TopoTag takes less time than ArUco (38% $\Downarrow$), AprilTag-1 (86% $\Downarrow$) and RuneTag (93% $\Downarrow$). Though ARToolKit, ChromaTag, AprilTag-2 and ARToolKitPlus run faster than

TABLE 5
Pose Estimation of Different Bits and Shapes

Different Bits & Shapes	Pose Accuracy				Pose Jitter			
	position (mm)		rotation (deg)		position (mm)		rotation (deg)	
	avg	max	avg	max	avg	max	avg	max
3x3, circle	1.073	3.299	0.016	0.048	0.069	0.265	0.080	0.205
3x3, square	0.837	2.780	0.022	0.065	0.085	0.383	0.081	0.276
4x4, circle	0.995	2.867	0.017	0.057	0.058	0.192	0.073	0.272
4x4, square	1.011	3.289	0.019	0.068	0.055	0.173	0.058	0.211

TopoTag, they offer significantly less unique identities. See Table 1 for details. For TopoTag, most time is spent on segmentation (68.8 percent), followed by decoding and vertex estimation (29.7 percent). Pose estimation takes the least time (1.5 percent).

It's worth noting that no parallelization is utilized in current TopoTag implementation, which will normally bring further speed-up. To demonstrate the possible applications on mobile, we have implemented the 2D marker detection process in a single pipeline on a Lattice FPGA (LFE5UM-45 with 44k LUTs, 1.9 Mb RAM and without using external DDR), which is decreased to < 100 us achieving $230\times$ speedup.

5.6 Flexible Shape Support

TopoTag supports both customized external and internal shapes as long as the topological structure is maintained. Fig. 1 shows three TopoTags with various internal shapes like square, circle, hexagon and different external shapes including square and butterfly. Fig. 14 shows our algorithm running upon these customized TopoTags.

Experiments show that tags with different shapes have compatible results. Please see Table 5 for detailed comparison. It's worth noting that all these four different TopoTags have 100 percent result on both detection recall and precision which further validates the robustness of the TopoTag system.

Fig. 15. Occlusion test by blocking 40 percent marker area starting from away the baseline node. Makers from left to right are ARToolKit, TopoTag, RuneTag, ArUco, ARToolKitPlus, ChromaTag, AprilTag-1&2, and AprilTag-3 respectively. (Best viewed in color).

TABLE 6
Occlusion Test Result

Tag	top → bottom occlusion				bottom → top occlusion			
	10%	20%	30%	≥ 40%	10%	20%	30%	≥ 40%
ARToolKit	×	×	×	×	×	×	×	×
ARToolKitPlus	×	×	×	×	×	×	×	×
ArUco	×	×	×	×	×	×	×	×
RuneTag	✓	✓	✓	×	✓	✓	✓	×
ChromaTag	×	×	×	×	×	×	×	×
AprilTag-1&2	×	×	×	×	×	×	×	×
AprilTag-3	×	×	×	×	×	×	×	×
TopoTag	✓	×	×	×	✓	×	×	×

"top → bottom occlusion" and "bottom → top occlusion" means occlusion starting from and away from the baseline node side respectively.

5.7 Occlusion Support

TopoTag can handle occlusion as long as topological structure is preserved. The left image of Fig. 14 is an example working under occlusion. Similar to [50], we conduct an occlusion test by blocking different percentages (10%→ 100% with 10 percent step size) of the marker area. As TopoTag is used with a unique baseline node, for fairness, we conduct the occlusion test twice, i.e., one starting from the baseline node side and the other away from it. Fig. 15 shows an example of an occlusion test setup with result in Table 6. We can see that all markers except TopoTag and RuneTag fail all occlusion tests. RuneTag achieves the best occlusion performance with max 30 percent occlusion and TopoTag can work well with up to 10 percent occlusion. Note that, as shown in above results, RuneTag has limitations of low detection rate and narrow tracking range due to its requirement of finding enough confident ellipses.

To handle more severe occlusions, similar to [4], [12], we can use multiple tags in a grid to increase the probability of detecting complete markers and other forms can be also considered. Fig. 16 shows an example of achieving

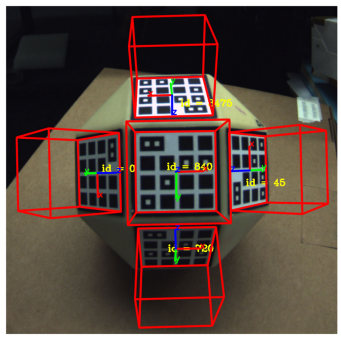
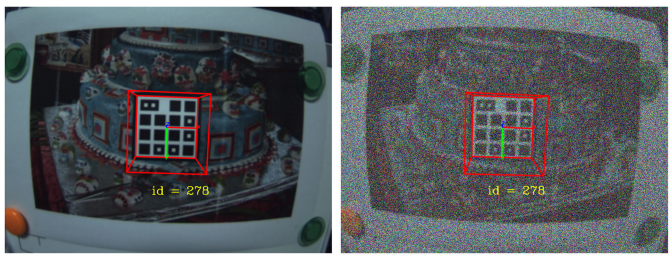


Fig. 16. $360°$-freedom tracking via using 18 TopoTags on a rhombicuboctahedron-shaped object. (Best viewed in color).



(a) Original image. (b) Image after adding noise.

Fig. 17. Example of TopoTag working under severe noise. Original image is from Seq #1. Noise image is obtained by adding Gaussian noise $\sigma = 0.45$ to the original image.

(a) Dark. (b) Bright outdoor. (c) Shadow. (d) Motion blur.

Fig. 18. TopoTag detection in different real scene tests with a rolling shutter camera. (Best viewed in color).

TABLE 7
Average and Maximum Pose Jitters of Each Tag Under Different Real Test Scenarios

Tag	Dark				Bright				Shadow			
	position (mm)		rotation (deg)		position (mm)		rotation (deg)		position (mm)		rotation (deg)	
	avg	max	avg	max	avg	max	avg	max	avg	max	avg	max
ARToolKit	—	—	—	—	—	—	—	—	0.504	0.879	0.093	0.210
ARToolKitPlus	0.103	1.301	0.068	0.200	**0.020**	0.113	0.030	0.110	1.399	3.711	0.469	0.960
ArUco	—	—	—	—	0.056	0.183	0.023	0.077	1.768	7.793	1.199	4.913
RuneTag	—	—	—	—	—	—	—	—	—	—	—	—
ChromaTag	—	—	—	—	—	—	—	—	—	—	—	—
AprilTag-1&2	0.065	0.224	0.040	0.139	0.033	0.118	0.018	0.066	—	—	—	—
AprilTag-3	0.112	0.334	0.060	0.228	0.044	0.462	0.020	0.099	0.103	0.523	**0.027**	**0.099**
TopoTag	**0.038**	**0.105**	**0.030**	**0.113**	0.021	**0.066**	**0.009**	**0.025**	**0.067**	**0.220**	0.039	0.116

"—" means detection failure. Best results are shown in bold and underlined.

360°-freedom tracking using 18 TopoTags on a rhombicuboctahedron-shaped object.

5.8 Noise Handing

TopoTag can handle certain noise due to our specially designed threshold map estimation, topological filtering and error correction. In Fig. 17, we show an example of TopoTag working under severe noise (adding Gaussian noise $\sigma = 0.45$ to the original image) by introducing image smoothing (i.e., Gaussian blur with kernel size $= 5,\ \sigma_x = 5.5,\ \sigma_y = 5.5$) as the pre-processing step.

5.9 Real Scene Test With a Rolling Shutter Camera

Besides the above laboratory testing with a global shutter camera, we further conduct real indoor and outdoor scene tests with a rolling shutter camera which is widely used in mobile phones and other smart devices. Specifically, we use a Logitech C930E webcam with 1280 × 720 resolution at 30 fps and 90 ° diagonal field of view. The experiment is conducted in four different scenarios, including dark, bright outdoor, shadow and motion blur. Fig. 18 shows the test setup with TopoTag detection overlay. For dark, bright outdoor and shadow scenarios, we also evaluate pose jitter and compare it with existing markers including the latest AprilTag-3 [51]. For fairness, for markers with multiple tag families, we randomly select one tag from the tag family with closest and smaller dictionary size compared with used TopoTag-4 x 4. Results are evaluated for a fixed length sequence of 100 frames and results shown in Table 7. We can see that TopoTag, together with ARToolKitPlus and AprilTag-3, performs well in all test scenarios, while all other markers fail in at least one scenarios. TopoTag also achieves the best (9 out of 12) or 2nd best (3 out of 12) performance in terms of position and rotation jitter for all scenarios.

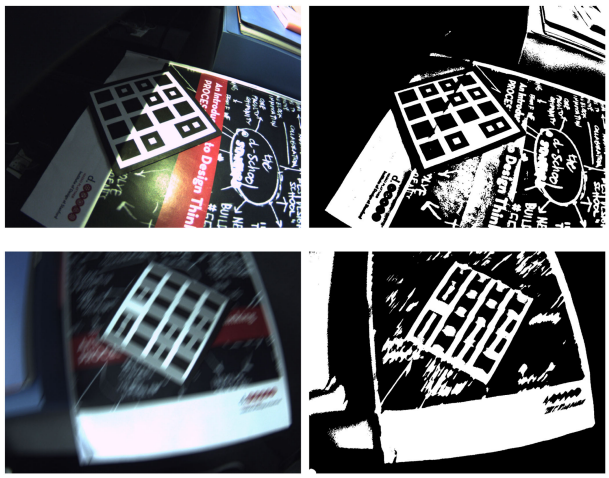


Fig. 19. Failure cases. Top row shows one of the failure cases of dramatic lighting change, and bottom row shows one of the severe motion blur. On right, binarization results are shown for each case respectively.

5.10 Failure Cases

TopoTag can handle lighting change and motion blur better due to our unique threshold map estimation and topological filtering modules, see examples in Fig. 18. However, it will fail to detect the markers where dramatic lighting change or severe motion blur happens over the marker region. Fig. 19 shows two typical failure cases as a result of dramatic lighting change and severe motion blur. Their binarization results show that markers' topological structure is dramatically changed. This change is the root cause of the detection failure.

6 Conclusion

We present TopoTag, a new topological-based fiducial marker and detection algorithm that utilizes topological

information to achieve high robustness, and near-perfect detection accuracy. We show that all tag bits can be used to encode identities without sacrificing detection accuracy, thus achieving rich identification and scalability. TopoTag offers more feature correspondences for better pose estimation. We demonstrate that TopoTag achieves the best performance in various metrics including detection accuracy, localization jitter and accuracy, and at the same time supports occlusion and flexible shapes. We also collected a large dataset of TopoTag and other previous state-of-the-art tags for better evaluation, involving in-plane and out-of-plane rotations, image blur, various distances and cluttered background, etc.

For future research, we will explore novel ID encoding/decoding strategy. We believe that this is key for a better marker system with a goal of strong occlusion resistance and scalability in addition to high detection rate and long distance tracking range.

REFERENCES

[1] J. DeGol, T. Bretl, and D. Hoiem, "ChromaTag: A colored marker and fast detection algorithm," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1472–1481.

[2] M. Fiala, "ARTag, a fiducial marker system using digital techniques," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 590–596.

[3] M. Fiala, "Designing highly reliable fiducial markers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1317–1324, Jul. 2010.

[4] B. Atcheson, F. Heide, and W. Heidrich, "CALTag: High precision fiducial markers for camera calibration," in *Proc. Int. Workshop Vis. Model. Vis.*, 2010, pp. 41–48.

[5] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2011, pp. 3400–3407.

[6] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 4193–4198.

[7] F. Bergamasco, A. Albarelli, E. Rodolà, and A. Torsello, "RUNE-Tag: A high accuracy fiducial marker with strong occlusion resilience," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 113–120.

[8] Y. Cho, J. Lee, and U. Neumann, "A multi-ring color fiducial system and an intensity-invariant detection method for scalable fiducial-tracking augmented reality," in *Proc. Int. Workshop Augmented Reality*, 1998, pp. 147–165.

[9] D. Flohr and J. Fischer, "A lightweight id-based extension for marker tracking systems," in *Proc. Eurographics Symp. Virt. Environ. Short Paper*, 2007, pp. 59–64.

[10] V. F. da Camara Neto, D. B. de Mesquita, R. F. Garcia, and M. F. M. Campos, "On the design and evaluation of a precise scalable fiducial marker framework," in *Proc. 23rd SIBGRAPI Conf. Graph. Patterns Images*, 2010, pp. 216–223.

[11] S. Garrido-Jurado, R. Munoz-Salinas, F. J. Madrid-Cuevas, and R. Medina-Carnicer, "Generation of fiducial marker dictionaries using mixed integer linear programming," *Pattern Recognit.*, vol. 51, pp. 481–491, 2016.

[12] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognit.*, vol. 47, no. 6, pp. 2280–2292, 2014.

[13] M. G. Prasad, S. Chandran, and M. S. Brown, "A motion blur resilient fiducial for quadcopter imaging," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2015, pp. 254–261.

[14] L. Calvet, P. Gurdjos, C. Griwodz, and S. Gasparini, "Detection and accurate localization of circular fiducials under highly challenging conditions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 562–570.

[15] T. Birdal, I. Dobryden, and S. Ilic, "X-Tag: A fiducial tag for flexible and accurate bundle adjustment," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 556–564.

[16] H. Cruz-Hernández and L. G. de la Fraga, "A fiducial tag invariant to rotation, translation, and perspective transformations," *Pattern Recognit.*, vol. 81, pp. 213–223, 2018.

[17] J. Molineros and R. Sharma, "Real-time tracking of multiple objects using fiducials for augmented reality," *Real-Time Imag.*, vol. 7, no. 6, pp. 495–506, 2001.

[18] F. J. Romero-Ramirez, R. Muñoz-Salinas, and R. Medina-Carnicer, "Speeded up detection of squared fiducial markers," *Image Vis. Comput.*, vol. 76, pp. 38–47, 2018.

[19] L. B. Gatrell, W. A. Hoff, and C. W. Sklair, "Robust image features: Concentric contrasting circles and their image extraction," in *Proc. Cooperative Intell. Robot. Space II*, 1992, pp. 235–245.

[20] V. A. Knyaz, "The development of new coded targets for automated point identification and non-contact 3D surface measurements," *Int. Archives Photogrammetry Remote Sens.*, vol. 5, pp. 80–85, 1998.

[21] L. Naimark and E. Foxlin, "Circular data matrix fiducial system and robust image processing for a wearable vision-inertial self-tracker," in *Proc. 1st Int. Symp. Mixed Augmented Reality*, 2002, Art. no. 27.

[22] J. Sattar, E. Bourque, P. Giguere, and G. Dudek, "Fourier tags: Smoothly degradable fiducial markers for use in human-robot interaction," in *Proc. 4rth Can. Conf. Comput. Robot Vis.*, 2007, pp. 165–174.

[23] A. Xu and G. Dudek, "Fourier tag: A smoothly degradable fiducial marker system with configurable payload capacity," in *Proc. Can. Conf. Comput. Robot Vis.*, 2011, pp. 40–47.

[24] F. Bergamasco, A. Albarelli, L. Cosmo, E. Rodola, and A. Torsello, "An accurate and robust artificial marker based on cyclic codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2359–2373, Dec. 2016.

[25] L. Calvet, P. Gurdjos, and V. Charvillat, "Camera tracking using concentric circle markers: Paradigms and algorithms," in *Proc. 19th IEEE Int. Conf. Image Process.*, 2012, pp. 1361–1364.

[26] F. Bergamasco, A. Albarelli, and A. Torsello, "Pi-Tag: A fast image-space marker design based on projective invariants," *Mach. Vis. Appl.*, vol. 24, no. 6, pp. 1295–1310, 2013.

[27] J. Rekimoto, "Matrix: A realtime object identification and registration method for augmented reality," in *Proc. 3rd Asia Pacific Comput. Hum. Interact.*, 1998, pp. 63–68.

[28] H. Kato and M. Billinghurst, "Marker tracking and HMD calibration for a video-based augmented reality conferencing system," in *Proc. 2nd IEEE ACM Int. Workshop Augmented Reality*, 1999, pp. 85–94.

[29] J. Rekimoto and Y. Ayatsuka, "CyberCode: Designing augmented reality environments with visual tags," in *Proc. Designing Augmented Reality Environ.*, 2000, pp. 1–10.

[30] M. Rohs and B. Gfeller, "Using camera-equipped mobile phones for interacting with real-world objects," *Advances Pervasive Comput.*, vol. 176, pp. 265–271, 2004.

[31] D. Wagner, "ARToolKitPlus for pose tracking on mobile devices," in *Proc. 12th Comput. Vis. Winter Workshop*, 2007, pp. 139–146.

[32] K. Tateno, I. Kitahara, and Y. Ohta, "A nested marker for augmented reality," in *Proc. IEEE Virt. Reality Conf.*, 2007, pp. 259–262.

[33] F. Schweiger, B. Zeisl, P. Georgel, G. Schroth, E. Steinbach, and N. Navab, "Maximum detector response markers for sift and surf," in *Proc. Vis. Model. Vis. Workshop*, 2009, pp. 145–154.

[34] E. Costanza and J. Robinson, "A region adjacency tree approach to the detection and design of fiducials," in *Proc. Video Vis. Graph.*, 2003, pp. 63–69.

[35] E. Costanza, "D-touch: A consumer-grade tangible interface module and musical applications," in *Proc. Conf. Hum.-Comput. Interact.*, 2003, pp. 175–178.

[36] R. Bencina, M. Kaltenbrunner, and S. Jorda, "Improved topological fiducial tracking in the reactivision system," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2005, pp. 99–99.

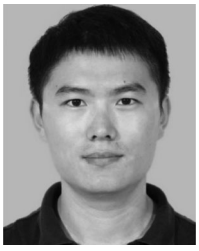
[37] R. Bencina and M. Kaltenbrunner, "The design and evolution of fiducials for the reactivision system," in *Proc. 3rd Int. Conf. Generative Syst. Electron. Arts*, 2005, pp. 97–106.

[38] M. Kaltenbrunner and R. Bencina, "reacTIVision: A computer-vision framework for table-based tangible interaction," in *Proc. 1st Int. Conf. Tangible Embedded Interact.*, 2007, pp. 69–74.

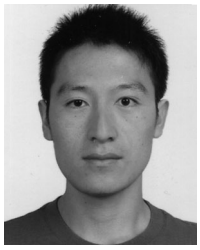
[39] C. N. Klokmose, J. B. Kristensen, R. Bagge, and K. Halskov, "BullsEye: High-precision fiducial tracking for table-based tangible interaction," in *Proc. 9th ACM Int. Conf. Interactive Tabletops Surfaces*, 2014, pp. 269–278.

[40] C. B. Owen, F. Xiao, and P. Middlin, "What is the best fiducial?" in *Proc. 1st IEEE Int. Workshop Agumented Reality Toolkit*, 2002, pp. 8–pp.

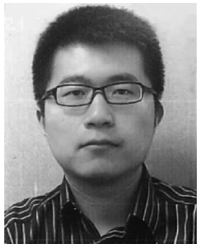
[41] D. Claus and A. W. Fitzgibbon, "Reliable fiducial detection in natural scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 469–480.
[42] D. Claus and A. W. Fitzgibbon, "Reliable automatic calibration of a marker-based position tracking system," in *Proc. 7th IEEE Workshops Appl. Comput. Vis.*, 2005, pp. 300–305.
[43] V. Lepetit and P. Fua, "Keypoint recognition using randomized trees," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1465–1479, Sep. 2006.
[44] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua, "Fast keypoint recognition using random ferns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 448–461, Mar. 2010.
[45] T. Collins and A. Bartoli, "Infinitesimal plane-based pose estimation," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 252–286, Sep. 2014.
[46] P. Chen and D. Suter, "Error analysis in homography estimation by first order approximation tools: A general technique," *J. Math. Imag. Vis.*, vol. 33, no. 3, pp. 281–295, Mar. 2009.
[47] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *J. Soc. Ind. Appl. Math.*, vol. 11, no. 2, pp. 431–441, 1963.
[48] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[49] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 157–173, 2008.
[50] K. Shabalina, A. Sagitov, M. Svinin, and E. Magid, "Comparing fiducial markers performance for a task of a humanoid robot self-calibration of manipulators: A pilot experimental study," in *Interactive Collaborative Robotics*, A. Ronzhin, G. Rigoll, and R. Meshcheryakov, Eds. Cham, Switzerland: Springer, 2018, pp. 249–258.
[51] M. Krogius, A. Haggenmiller, and E. Olson, "Flexible layouts for fiducial tags," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 1898–1903.



Guoxing Yu received the BEng degree in electronic information engineering from the Wuhan University of Science and Technology, Wuhan, China, in 2013, and the ME degree in information and communication engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2016. He is currently with Guangdong Virtual Reality Co., Ltd. (aka. Ximmerse) as an algorithm engineer. Prior to joining Ximmerse, he was an algorithm engineer with Wuhan Guide Infrared Company Ltd. Wuhan from July 2016 to August 2017. His research interests include computer vision, augmented reality, and virtual reality.



Yongtao Hu received the BEng degree in computer science from Shandong University, Jinan, China, in 2010, and the PhD degree in computer science from the University of Hong Kong, Hong Kong, in 2014. He is currently with Guangdong Virtual Reality Company Ltd. (aka. Ximmerse) as a research scientist. Prior to joining Ximmerse, he was a staff researcher with Image and Visual Computing Lab (IVCL), Lenovo Research, Hong Kong from January 2015 to October 2015, was a researcher assistant with IVCL from July 2014 to November 2014, and was a research intern with Internet Graphics Group in Microsoft Research Asia (MSRA) from March 2010 to June 2010. His research interests include computer vision, multimedia, machine learning, augmented reality, and virtual reality.



Jingwen Dai (Member, IEEE) received the BE degree in automation from Southeast University, Nanjing, China, in 2005, the ME degree in automation from Shanghai Jiao Tong University, Shanghai, China, in 2009, and the PhD degree in mechanical and automation engineering from the Chinese University of Hong Kong, Hong Kong, in 2012. He is currently with Guangdong Virtual Reality Company Ltd. (aka. Ximmerse) as cofounder and chief technology officer. Prior to joining Ximmerse, he was a manager and advisory researcher with Image and Visual Computing Lab (IVCL), Lenovo Research, Hong Kong from January 2014 to July 2015, and was a postdoctoral research associate with the Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina from October 2012 to December 2013. His current research interests include computer vision and its applications in human-computer interaction, augmented reality, and virtual reality.