Tracking Manipulation Tasks (TMT) Benchmark and Evaluation

By:

Ankush Roy, Xi Zhang, Nina Wolleb, Camilo Perez

Instructor: Dr.Martin Jagersand

July 7, 2015

Contents

List of Figures

Tracking Benchmark and Evaluation for Manipulation Tasks

Ankush Roy, Xi Zhang, Nina Wolleb, Camilo Perez, Martin Jagersand
University of Alberta
Instructor: Dr. Martin Jagersand

July 7, 2015

Abstract

In this paper we present a public dataset to evaluate trackers used for human and robot manipulation tasks. For these tasks both high DOF motion and high accuracy is needed. We describe in detail, both the process of recording the sequences and how ground truth data was generated for the videos. The videos are tagged with challenges that a tracker would face while tracking the object. As an initial example, we evaluate the performance of six published trackers [5, 11, 12, 13, 15, 6] and analyse their result. We describe a new evaluation metric to test sensitivity of trackers to speed. A total of 100 annotated and tagged sequences are reported. All the videos, ground truth data, original implementation of trackers and evaluation scripts are made publicly available on the website so others can extend the results on their trackers and evaluation.

1 Introduction

Object tracking is a core component in visual servoing and manipulation. It is particularly challenging to precisely track the many degrees of freedom (DOF) needed to control the motion in robot arm and hand manipulation. In image-based servoing 2D video frames from a digital camera are processed by a tracking algorithm that returns the object state in each frame. Visual state is described in the image frame (e.g. homography parameters or 4 corner points of the patch) rather than world coordinate pose. This information guides a robotic manipulator to the desired position. This is much different from other works in tracking, e.g. surveillance where it is sufficient to track the 2 DOF object centre and possibly a loose bounding box.

Previous works either use synthetically generated data by applying random warps [8, 10] or test their algorithms on a small set of videos that are either recorded

Table 1: Comparision of Different Object Tracking Datasets

Datasets	**TR**	**RO**	**SP**	**IL**	**PR**	**SR**	**SC**	**OC**	**TX**	**DOF**
Metaio [2]	✓	✓	✓	✓	✓	✗	✓	✓	✓	8
Static [11, 12, 13]	✓	✓	✓	✗	✓	✗	✓	✗	✗	8
Zimmerman [9]	✓	✓	✓	✗	✓	✗	✓	✓	✗	6
CVPR [4]	✓	✓	✓	✓	✗	✓	✓	✓	✓	2
Gauglitz [3]	✓	✓	✓	✓	✓	✓	✓	✗	✓	8
ESM [11]	✓	✓	✗	✗	✓	✗	✓	✗	✗	8
BoBoT [22]	✓	✓	✓	✓	✗	✓	✓	✓	✓	2
VOT [14]	✓	✓	✓	✓	✗	✓	✓	✓	✓	2
TMT [24]	✓	✓	✓	✓	✓	✓	✓	✓	✓	8

TR = Translation; RO = Rotation; SP = Speed; IL = Illumination; PR = Perspective; SR = Specular; OC = Occlusion; TX = Texture

or pooled from the Internet. Zimmerman et al. [9] tested their algorithm on 3 grayscale videos (linTrack) that covered 6 DOF transformation of the object. They manually annotate the bounding box using either a special marker or visual texture based cues. Lin et al. [17] and Petit et al. [1] showed application of template tracking in augmented reality, but restricted themselves to a small set of videos. Some other popular sources that report videos for tracking are by Ross et al. in [5], Kalal et al. [6], Collins et al. in PETS2005 [7], Fisher et al. in CAVIAR [18]. However the challenge in tracking for these videos are targeted to evaluate trackers developed for surveillance applications. In a recent paper Yang et al. 2013 [4] categorized some of the above mentioned video sequences to publish a common benchmark. VOT Challenge [14] pools from a bigger set of videos. However, these video sequences are more suited for surveillance tracking.

Closest to our aim comes the Metaio dataset meant to evaluate planar homography tracking by Lieberknecht et al. [2]. They collect videos under various motions, illuminations and textures. They used a camera mounted on a Faro arm (a passive robot arm with very high precision joint encoders) that was calibrated and all transformations of pose were stored. The stored values were used to calculate ground truth data. However, instead of a 3D scene the Metaio benchmark like some other datasets [3] records a printed (2D planar) poster, making the benchmark somewhat artificial. Furthermore the setup with a moving camera on an arm means that the motions are not from natural tasks and restricted by the arm workspace and mass.

In this paper we provide a public video dataset to evaluate 2*D* marker-less

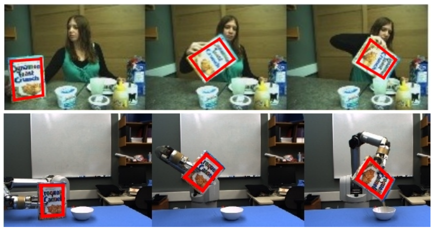

Figure 1: Sequence showing both the robot and human user performing identical tasks of pouring cereal in a bowl under normal light settings. The red rectangle shows the tracking result on the sequence using ESM [11]

single object tracking algorithms. The tasks are natural table top manipulations, frequently performed in our daily life. We provide videos of both a human and a robot robot perform the same task. Robot motion is normally planned with smooth velocity profiles and limited acceleration. Before grasping an object a robot normally comes to a smooth stop, then smooth start. Human motions are often less smooth, and a human may grasp an object on the fly, causing a tracked object to accelerate quickly. All videos are tagged with the challenges that a tracker would face while tracking. Using these tags susceptibility of a tracking algorithm can be properly narrowed down and subsequently improved. Each task is repeated with different speeds and under 2 lighting conditions.

We list four distinct contributions in this paper.

- **Dataset**: First a publicly available dataset which covers a wider range of challenges than most of the other publicly reported datasets (Table 1) for object tracking.
- **Performance Metric**: We introduce a metric that measures sensitivity of trackers to interframe motion.
- **Evaluation**:Several existing trackers [5, 6, 11, 12, 13, 15] are evaluated, their performance and suitability for arm/hand manipulation tasks discussed.
- **Code**: We make the tracker codes as well as the evaluation scripts publicly available. New trackers or different evaluation metrics [1] can be reported.

[1] Website: http://webdocs.cs.ualberta.ca/ vis/trackDB/

2 Dataset

The **Tracking Manipulation Tasks (TMT)** dataset consists of image sequences of manipulation task recorded by a human user and a robot arm.

2.1 Video Capture Set Up

All videos (Fig 4) were recorded using a GRAS-20S4C-C fire-wire camera equipped with a Kowa LM6NCM F1.2/6mm lens. For the human recorded videos the camera stood on a fixed position on a tabletop 90 cm from the edge of the table. Camera settings were implemented through coriander 2.0.1 [19]. Lighting was varied between normal and diffused light. The videos were recorded at 30 frames per second (f.p.s.) at a resolution of 600x800 in YUV colour space. Exact parameters of recording are listed in Table 2.

Table 2: Parameter Setting for Image Acquisition

Parameters	**Diffuse Light**	**Normal Light**
Exposure (in IL)	0.73	1.06
Gain (in dB)	6.02	0
Shutter (in s)	0.04	0.07
White Balance	Blue/U927 Red/V493	Blue/U757 Red/V490
Saturation (in %)	95.22	119.04

Fig. 2 shows a 7 DOF WAM arm with a Barret hand [20] performing the manipulation tasks. The camera was fixed on a tripod stand at a distance of 120 cm from the WAM arm. Other parameters were same as in Table 2.

2.2 Description of dataset

The tasks were selected in such a way so that one or more of the above challenges are covered (Table 3). The recorded videos were grouped under two broad categories *Single Motion Tasks* and *Composite Motion Tasks*.

2.2.1 Single Motion Tasks

refer to highly structured motion of the object

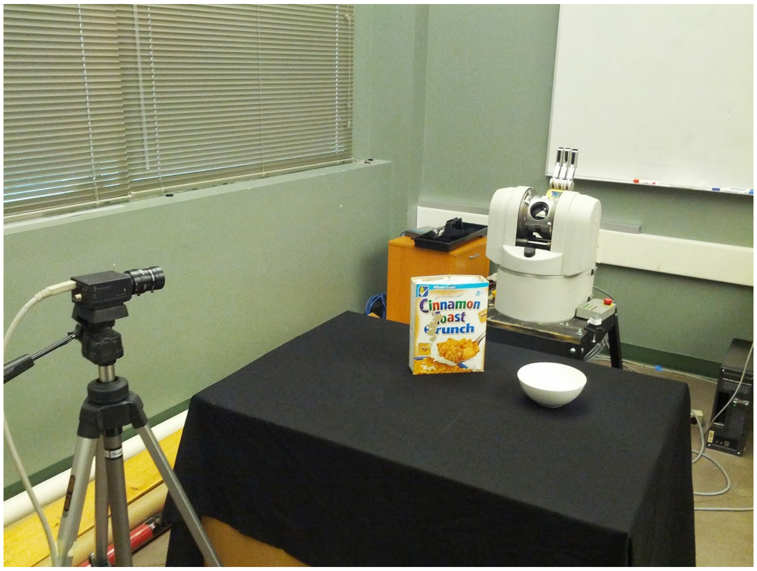

Figure 2: Set up for recording videos under normal light using a WAM arm and a Barret hand.

Juice Juice from a juice box is poured in a container. The goal is to track the juice box handling both translational (TR) and rotational (RO) motion of the object. The axis of rotation being parallel to the camera axis.

Cereal This task is similar to that of *Juice*, with an added challenge of large motion when the cereal is poured.

Book I The object (book) is tilted from a vertical upright position to finally lie flat on the table and vice versa. The challenge is to handle perspective (PR) deformation of the object.

Book II Here a book is brought near the camera parallel to the camera axis and away. The goal is to capture scale (SC) variation of the object.

Book III The challenge in this video is to track a book despite varying occlusion (OC) from the book holder.

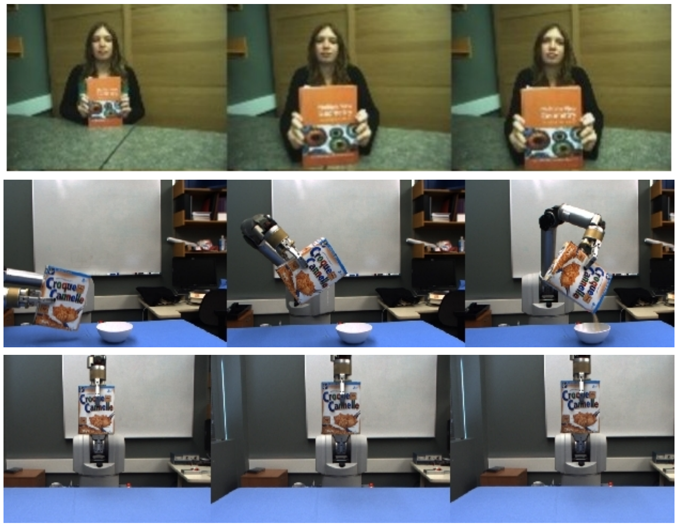

Figure 3: Image frames of the video sequences show sample tasks performed by both a human user and a robot hand

Mug I Pure translational (TR) motion of the object (coffee mug) is recorded. Specular reflection (SR) and low texture (TX) of the object present additional challenge in this case.

Mug II This sequence though similar to *Mug I*, but is more difficult because of perspective (PR) deformation when it is tilted to drink from it.

Mug III A low texture (TX) coffee mug is rotated (RO) from its initial position and back along the axis perpendicular to the principal camera axis.

2.2.2 Composite Motion Tasks

has videos that can be decomposed into simpler *Single Motion Tasks*.

Bus The planar surface in front of a toy bus that undergoes scale change (SC) and perspective deformation (PR) at varying speed (SP) is to be tracked.

Highlighting A portion of the newspaper is highlighted with a marker pen. The challenge is to track the object in the presence of changing texture (TX) caused by highlighting and occlusion (OC) by the pen.

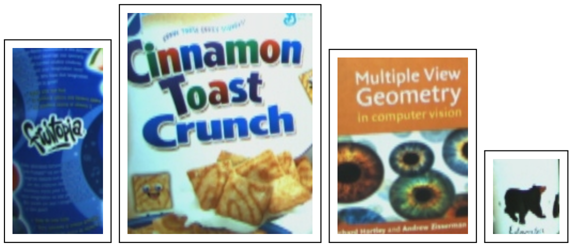

Figure 4: Planar projection of the objects used in the dataset are shown, both planar and curvilinear objects are chosen with varying texture, lambertian/specular to have a full spectrum of challenges

Letter The sequence records a letter being put inside an envelope and out again. The challenge in largely to tackle the perspective deformation (PR) of the object.

Newspaper A portion of the newspaper is to be tracked in the presence of perspective (PR) and scale (SC) changes under varying speed.

Single Motion Tasks are recorded at five different speeds, starting from very slow to very fast motion of the object. Table 3 summarises the challenges.

2.3 Ground Truth and Error Metric

2.3.1 Ground Truth

Ground truth (GT) refers to the four co-ordinate positions of the bounding box, denoting the object's location in the image plane. We report sub-pixel level co-ordinate positions based on tracking data. Three trackers [11, 12, 13] are initiated on the first frame. Ground truth is registered only when bounding box co-ordinates reported by all three of them lie within ±1 pixel. The reason for choosing the trackers are because of their high convergence.

The stringent convergence criteria (±1 pixel) sometimes result in the trackers failing to converge to a common bounding box. We reinitialize the three trackers using positional information from previous frames every time it fails to converge. Number of reinitialisations varied from 0 to 12, for more difficult sequences. As a final step we also verify all ground truth data manually.

2.4 Error Metric

To evaluate a tracker we need an error measure. Some of the commonly [4] used measures are centre distance of the object (X_{CT}) from ground truth (X_{CGTi}) ex-

Table 3: Description of Videos

Video	Object	Tracking Challenge
Juice	Juice Box	TR,RO,SP,IL
Cereal	Cereal Box	TR,RO,SP,IL
Book I	Book	PR,SP,IL,SR
Book II	Book	SC,SP,IL
Book III	Book	TR,OC,SP,IL
Mug I	Coffee Mug	TR,SP,IL,SR
Mug II	Coffee Mug	TR,PR,SP,IL,SR
Mug III	Coffee Mug	RO,SP,IL,SR
Bus	Toy Bus	TR,SC,PR,SP,IL
Highlighting	Newspaper	OC,SP,IL,TR
Letter	Envelope	PR,SP,IL
Newspaper	Newspaper	PR,SC,SP,IL

pressed as $E_C(E_C = \sqrt{(X_{CT} - X_{CGTi})^2})$ and area overlap E_A ($E_A = \frac{|A_T \cap A_{GT}|}{|A_T \cup A_{GT}|}$ where A_T is area of the target and A_{GT} area of ground truth). However, we define our error metric as that of [2]. The alignment error E_{AL} (eq. 1) is expressed as a root mean square distance of misalignment of the target image (X_{Ti}) with ground truth (X_{GTi}).

$$Error(E_{AL}) = \sqrt{\frac{\sum_{i=1}^{4}(X_{Ti} - X_{GTi})^2}{4}} \tag{1}$$

This score correctly captures the misalignment as shown in Fig. 5. The two errors E_C and E_A would be low but E_{AL} would be high since the target image (T) is 180° out of phase with the ground truth. Any further reference to error unless otherwise mentioned refers to E_{AL}.

3 Tracker Evaluation: Results and Analysis

3.1 Evaluation

We evaluate six trackers based on three measures *Overall Success (Robustness)*, *Average Drift (Convergence)* and a measure that we define, *Speed Sensitivity*. This last measure quantitatively captures a tracker's robustness to inter frame motion.

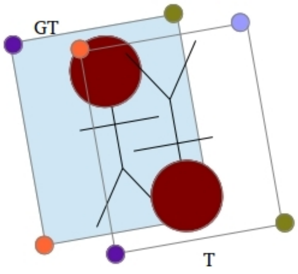

Figure 5: Displacement of each colour coded corner is expressed as a RMSE score to account for the misalignment of pose

3.1.1 Overall Success

of a tracker is defined as the fraction of frames that a tracker tracks within t_p pixel threshold. It can be expressed as $\frac{|S|}{|F|}$ (where $S = \{f^i \in F : e^i_{AL} < t_p\}$, F set of all frames, e^i_{AL} error of frame f^i).

3.1.2 Average Drift

computes the expected error of the tracker when it operates within a allowed drift of t_p pixels. This is similar to the one proposed by Stenger et al. in [16] $E[e|e < t_p]$ ($E[e|e < t_p] = \frac{\sum_{f_i \in F} e^i_{AL}}{|S|} : e^i_{AL} < t_p$). Yang et al. [4] used a pixel threshold t_p of 20 pixels. This is too high for manipulation tasks. We choose threshold t_p to be 5 pixels.

3.1.3 Speed Sensitivity

A common way to characterize convergence is to synthetically warp the standard ”Lena” image and let the tracker recover the warp. Warps are sampled on $\mathcal{N}(0,\sigma)$. However, σ as used in [11, 12, 13] overestimates convergence, since $\mathcal{N}(0,\sigma)$ with large σ will also contain small motions. From Fig 6, trackers perform poorly on large motion of 12 pixels (60%) compared to the corresponding sigma (90%) for NNIC.

It is more relevant to study the success rate against true inter frame motion m_i (Eq: 2).

$$m_i = rmse(gt^{i+1}, gt^i) = \sqrt{\frac{\sum_{k=1}^{4}(gt_k^{(i+1)} - gt_k^i)^2}{4}} \tag{2}$$

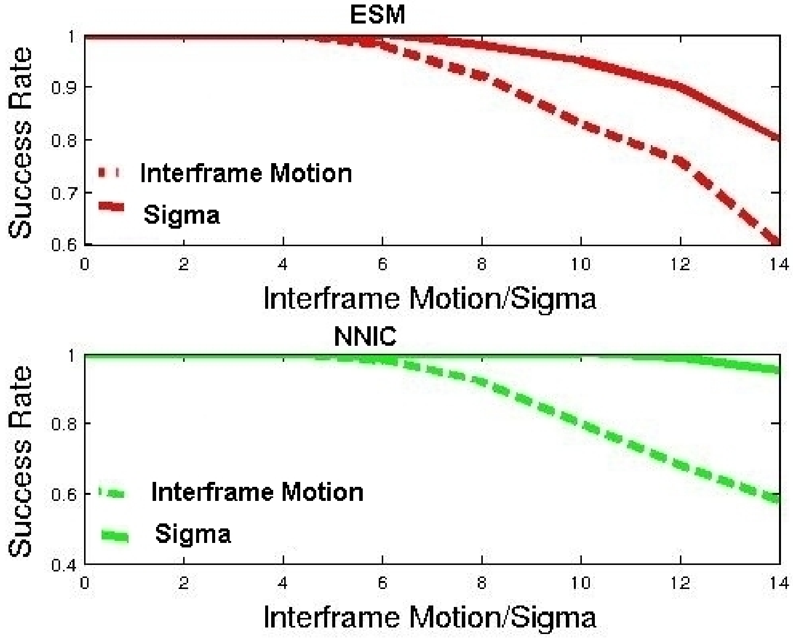

Figure 6: Success rate ($t_p = 2$) plotted against varying [11, 12, 13] and inter frame motion . It shows that even though the success rate for higher sigma is high it is not so for the corresponding inter frame motion.

3.2 Trackers Used

An equal sample of both registration and online learning based trackers were chosen. Each tracker was presented with the videos and the bounding box co - ordinates of the object in the first frame. Original implementation of NNIC [12], IVT [5], TLD [6] and L1-APG [13] were used. Parameters were set after cross validation to ensure best results.Each tracker was run through the video sequence and corresponding results were analysed.

3.2.1 Registration Based

We test ESM [11], NNIC[12], RKLT [23] and BMIC [13]. 30 iterations with a resolution of 100×100 were used in BMIC for convergence. NNIC uses a look up table of (0.06×0.04), (0.03×0.02) and (0.015×0.01) sigmas, each using 2000 samples and a resolution of 50×50. ESM uses 30 iterations with a resolution of

50×50. The parameters were chosen to give each tracker approximately the same execution time.

3.2.2 Online Learning Based

Three online learning based trackers were used, IVT [5] , L1-APG [13] and TLD [6]. Yang et al. in [4] analysed some state of the art trackers that used online learning. We choose three of the trackers that performed well in most of the videos. Both IVT[5] and L1-APG[13] used 600 particles for their particle filter search. Additionally L1-APG [13] tracker used a template subspace of 10 templates of the target image. Original parameters as specified in [6] were used for TLD.

3.3 Result and Analysis

Average Drift and Overall Success as described in Section 3.1 averaged over very slow, slow and medium speed are reported in Table 4 and Table 5.

3.3.1 Average Drift

Table 4: Average Drift Expressed as an Expectation Value

Sequence	**TLD**	**L1-APG**	**IVT**	**ESM**	**NNIC**	**BMIC**
Juice	N/A	2.39	3.27	**0.42**	0.51	**0.42**
Cereal	N/A	2.67	2.80	0.28	**0.27**	**0.27**
Book I	N/A	2.87	2.96	0.324	**0.29**	**0.29**
Book II	N/A	3.32	3.37	0.29	**0.27**	**0.27**
Book III	4.97	1.21	1.14	**0.22**	**0.22**	**0.22**
Mug I	3.48	3.3	1.34	**0.27**	**0.27**	**0.27**
Mug II	3.64	2.47	3.09	0.70	**0.49**	0.50
Mug III	3.20	2.51	2.28	**0.25**	1.22	1.29
Bus	3.25	0.68	2.18	0.63	**0.59**	0.63
Highlighting	N/A	3.71	1.61	1.23	1.21	**0.47**
Letter	4.53	1.79	1.68	**0.36**	0.505	**0.36**
Newspaper	N/A	2.49	3.18	0.42	0.53	**0.31**

N/A = There were no frames that had an error (E_{AL})less than 5 pixels

Registration based trackers in have low average drift compared to the online learned trackers for most of the sequences. The reason being, particle filter uses

Table 5: Overall Success Expressed as Fraction of Frames Tracked

Sequence	**TLD**	**L1-APG**	**IVT**	**ESM**	**NNIC**	**BMIC**
Cereal	0.00	0.24	0.99	**1**	**1**	**1**
Book I	0.00	0.10	0.48	**1**	**1**	**1**
Book II	0.00	0.79	0.30	**1**	**1**	**1**
Book III	0.00	0.42	**0.72**	0.34	0.32	0.32
Juice	0.00	0.16	0.98	**1**	0.41	**1**
Mug I	0.84	0.10	0.91	**1**	**1**	**1**
Mug II	0.41	0.30	0.72	**0.89**	**0.89**	**0.89**
Mug III	0.51	0.54	0.68	**1**	0.59	0.65
Bus	0.37	0.57	0.94	**1**	0.99	0.96
Highlighting	0.00	0.67	**0.95**	0.76	0.70	0.33
Letter	0.11	0.19	0.25	**1**	**1**	**1**
Newspaper	0.00	0.61	0.92	**1**	0.51	0.43

random samples to get an initial state estimate which is unlikely to hit the best alignment when iterations are limited, while the Gauss Newton method of the original registration trackers provide fast and higher convergence when they do converge.

We observe that L1-APG tracker fail to track in plane rotation (Fig 7). Although Mei et. al [15] use a 6 d.o.f motion model, it's an approximation. Their motion model use a velocity component $\vec{v} = (v_1, v_2)$ (horizontal and vertical velocities). This $\vec{v}$ is updated based on an average of the last few frames tracked. This accounts for the transitional motion of the object but fails to capture the full pose. This is not reported by existing work that we know of. TLD with it's 3 d.o.f. parametrization understandably fails to precisely estimate the object pose.

3.3.2 Overall Success

Online learned trackers build a model of the object (lower dimensional subspace of the initial template) to begin with. They update the model as tracking progresses. Newer templates are accounted for the appearance of the object. Though this makes the trackers more robust, this comes at the expense of convergence as newer templates (shifted, scaled from the original) are not exactly the same as the target image still they form a part of the appearance model.

Overall, the three registration based trackers show more usability in manipulation tasks (small threshold t_p) compared to some of the popular trackers(TLD [6],

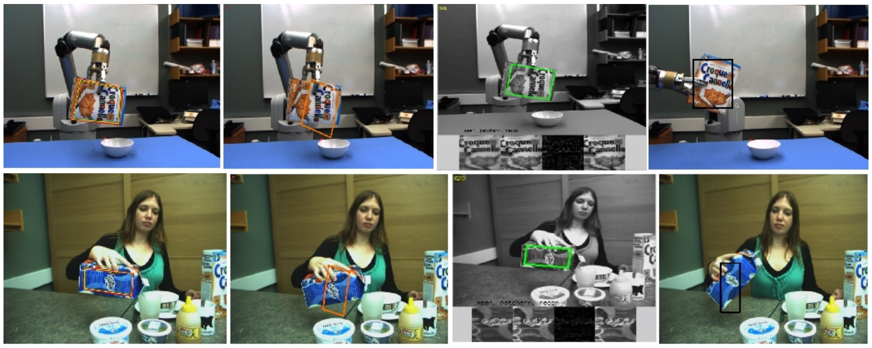

Figure 7: An instance of in plane rotation where L1-APG (orange) [15] (b) and TLD (black) [6] fails but the three homography based trackers i.e. ESM (red) [11], BMIC (yellow) [13] and NNIC (cyan) [12] (a) and IVT (green) [5] performs without considerable drift

L1-APG [15], IVT [5]).

The registration based trackers are further analysed to check how they perform when exposed to large motion and possible rank the difficulty of tracking a particular video sequence. For this we use "Speed Sensitivity".

3.3.3 Speed Sensitivity

Speed sensitivity (Fig: 6) if lower, trackers can track higher inter frame motion.

Planar Object "BookII" sequence is the easiest to track. The only challenge is scale (SC) variation. Compared to "BookII", "BookI" (Fig 8) has both perspective transformation (PR) and specularity (SR). When the book is tilted, (Fig 9) the appearance changes considerably which makes it hard for the tracker to track.

An interesting observation can be made (Fig: 10), though "Cereal" and "Juice" sequences have same sets of challenges, the "Juice" sequence is more sensitive to large motion.

The only difference between them being texture, we investigate further. Static image experiments [11, 12, 13] with optimal parameters are done.The results in Fig 11 show the effect of texture playing a crucial role, making the two sequences distinct in their own way. Cereal box having richer texture, is easier to track.

Curvilinear Object The mug sequences are on an average more difficult to track. With higher inter frame motion the success rate falls steeply. The specularity (SR)

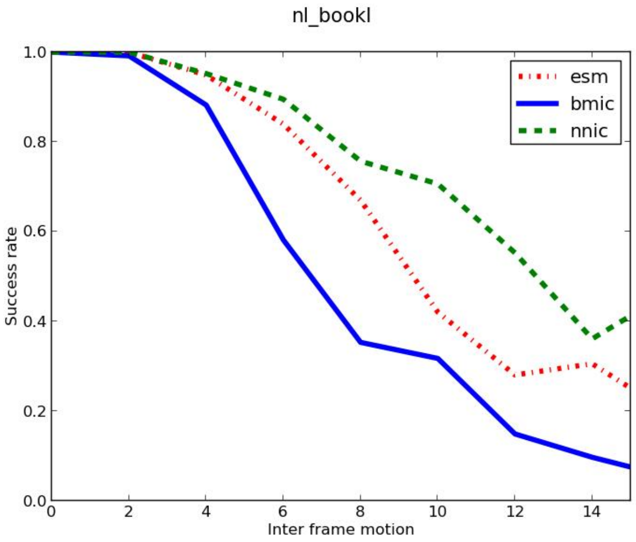

Figure 8: Speed Sensitivity result for “bookI” sequence

on the surface makes it hard to track. “MugII” is the most difficult sequence in the entire set, with the highest number of challenges (Table 3).

Speed sensitivity is reported taking into account all five speeds. We skip frames (track every second, third and fourth frame) to simulate fast motion. This gives us enough samples to have a statistically significant result. Videos reporting only one speed would suffer from the fact that enough natural motion samples could be sampled to report a significant result. We have made ground truth data publicly available, users can define their own evaluation criteria.

We also rank trackers 12 as done in [14]. This gives a global overview of how the trackers compare. However instead of using an area overlap error metric E_A we use E_{AL}. ESM and NNIC are the 2 top performing trackers. NNIC have higher success rates compared to ESM due of it’s ability to track larger object motion (Fig 6). TLD is the worst among the 6 trackers when used for precise tracking. It tracks 3 DOF state space and often re-initialises.

4 Conclusion and Discussion

We report a dataset to evaluate trackers designed for manipulation tasks. Complete ground truth data is made available for all the sequences. The source codes as well

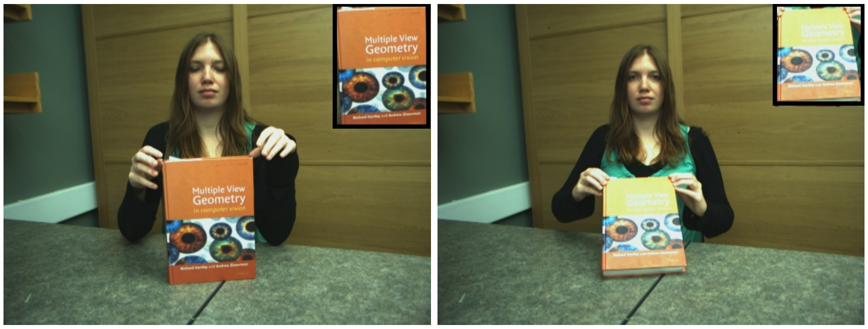

Figure 9: Frame 1 and Frame 190, in the “bookI” sequence, recorded at medium speed. The considerable change in appearance due to specular reflection is visually shown on top right corner

as the evaluation scripts are available for the trackers that were used for evaluation so that results could be replicated, or different evaluation metrics computed. Users are encouraged to use the dataset to evaluate their trackers.

We provide an initial analysis of some of the popular trackers in literature. Starting from an initial selection of trackers we systematically analyse the tracker’s performance. We introduce a new evaluation metric that tests the true robustness of a tracker to inter frame motion. This shows that the common way of evaluating the convergence on a static image [11, 12, 13] significantly overestimates convergence to dozens of pixels, while measured convergence on real videos is at best a few pixels inter frame motion.

Some interesting results are revealed during this analysis. We observe that even though L1-APG tracker has a 6 d.o.f motion model it cannot track in plane rotation. This is discussed further and explained. We also found that even though motions are identical of similar objects (planar and lambertian), the actual texture play a huge role in tracking.

In future we would like to augment the dataset by including videos from an eye in hand camera while manipulation is going on to have newer set of challenges.

References

[1] Petit, A., Caron, G., Uchiyama, H., Marchand, E. (2011). Evaluation of model based tracking with trakmark dataset. In 2nd Int. Workshop on AR/MR Registration, Tracking and Benchmarking.

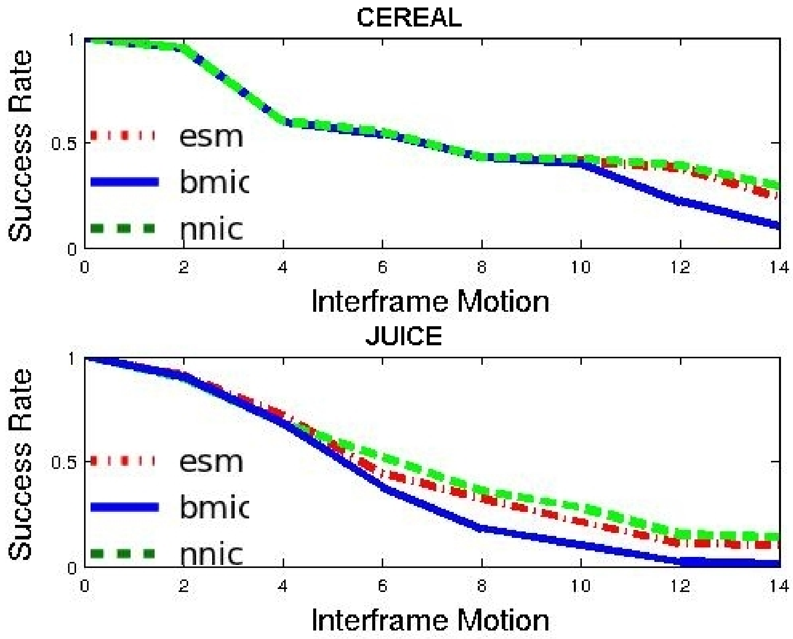

Figure 10: Speed sensitivity is plotted against inter frame motion for “Cereal” (top) and “Juice” (bottom) sequences

[2] Lieberknecht, S., Benhimane, S., Meier, P., Navab, N. (2009, October). A dataset and evaluation methodology for template-based tracking algorithms. In Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on (pp. 145-151)

[3] Steffen Gauglitz and Tobias Höllerer and Matthew Turk, Evaluation of Interest Point Detectors and Feature Descriptors for Visual Tracking, Evaluation of Interest Point Detectors and Feature Descriptors for Visual Tracking, 2011, vol-94, pp - 335-360

[4] Wu, Y., Lim, J., Yang, M. H. Online Object Tracking: A Benchmark, CVPR 2013

[5] Ross, D. A., Lim, J., Lin, R. S., Yang, M. H. (2008). Incremental learning for robust visual tracking. International Journal of Computer Vision, 77(1-3), 125-141.

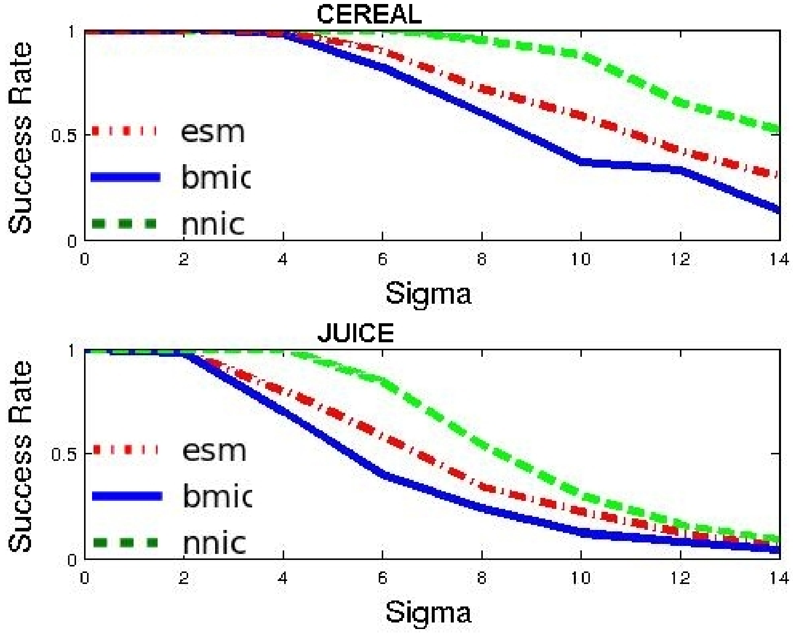

Figure 11: Static experiment [11, 12, 13] with 5000 trials for each sigma ($\sigma \in \{1, ..14\}$), is done on the cereal box (top) and juice box (bottom), cereal box having a rich texture is easier to track

[6] Kalal, Z., Mikolajczyk, K., Matas, J. (2012). Tracking-learning-detection. TPAMI, 34(7), 1409-1422.

[7] Collins, R., Zhou, X., Teh, S. K. An open source tracking test bed and evaluation web site. In IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (pp. 17-24).

[8] Baker, S., Matthews, I. (2004). Lucas-kanade 20 years on: A unifying framework. IJCV, 56(3), 221-255.

[9] Zimmermann, K., Matas, J.,Svoboda, T. (2009). Tracking by an optimal sequence of linear predictors. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31(4), 677-692

[10] Jurie, F. and Dhome, M. (2002). Hyperplane approximation for template matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 9961000

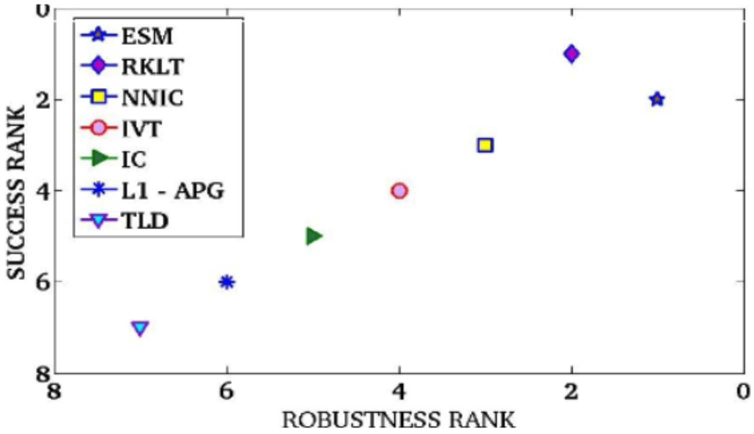

Figure 12: Trackers are ranked [14] based on robustness and overall success. The error measure used is E_{AL} with $t_p = 5$ pixels. RKLT and ESM are the two top performing trackers with TLD being the worst.

[11] Benhimane, S., Malis, E. (2004, September). Real-time image-based tracking of planes using efficient second-order minimization. In Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on (Vol. 1, pp. 943-948). IEEE.

[12] Dick, T., Perez, C., Shademan, A., Jagersand, M (2013, June). Realtime Registration-Based Tracking via Approximate Nearest Neighbour Search. In Proceedings of Robotics: Science and Systems, Germany

[13] Baker, S., Matthews, I. (2001). Equivalence and efficiency of image alignment algorithms. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 1, pp. I-1090). IEEE.

[14] Kristan, Matej, et al. ”The visual object tracking vot2013 challenge results.” Computer Vision Workshops (ICCVW), 2013 IEEE, 2013.

[15] Mei, X., Ling, H. (2009, September). Robust visual tracking using L_1 minimization. ICCV, (pp. 1436-1443).

[16] Stenger, B., Woodley, T., Cipolla, R. (2009, June). Learning to track with multiple observers. CVPR (pp. 2647-2654). IEEE.

[17] Lin, L., Wang, Y., Liu, Y., Xiong, C., Zeng, K. (2009). Marker-less registration based on template tracking for augmented reality. Multimedia Tools and Applications, 41(2), 235-252.

[18] Caviar : http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/

[19] Coriander : http://damien.douxchamps.net/ieee1394/coriander/

[20] WAM arm by Barett Technologies, Specification Available at: http://www.barrett.com/robot/products-arm.htm

[21] X. Mei and H. Ling, ”Robust Visual Tracking and Vehicle Classification via Sparse Representation”, IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 33(11):2259–2272, 2011

[22] D. A. Klein, D. Schulz, S. Frintrop, A. B. Cremers, Adaptive Real-Time Video-Tracking for Arbitrary Objects, IROS, 2010, pp : 772 - 777

[23] Xi Zhang, Abhineet Singh, Martin Jagersand, ”RKLT:8 DOF real-time robust video tracking combinin coarse RANSAC features and accurate fast template registration” , CRV, 2015

[24] Ankush Roy, Xi Zhang, Nina Wolleb, Camilo Perez, Martin Jagersand, ”Tracking Benchmark and Evaluation for Manipulation Tasks”, in IEEE Proc. of International Conference on Robotics and Automation (ICRA), 2015