Tracking Benchmark and Evaluation for Manipulation Tasks

Ankush Roy, Xi Zhang, Nina Wolleb, Camilo Perez Quintero, Martin Jagersand

***Abstract*— In this paper we present a public dataset to evaluate trackers used for visual manipulation tasks. 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 in the sequence and the task it simulates. As an initial example, we evaluate the performance of five published trackers [5], [12], [13], [14], [15] and analyse their result. A total of 100 annotated and tagged sequences are reported. All the videos and ground truth values are made publicly available on the website `http://webdocs.cs.ualberta.ca/~vis/trackDB/`.**

I. Introduction

Object tracking is a core component in visual servoing and similar methods of using real-time vision to guide robot motion. 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 passed on to a tracking algorithm that returns the object state in each image 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. The nature of the application calls for trackers that can accurately track high degrees of freedom (DOF) state transformation of the object. This is much different from others work in tracking, e.g. surveillance where it is sufficient to track the 2 DOF object centre and possibly a loose bounding box.

In this paper we provide a public video dataset to evaluate 2*D* marker-less single object tracking algorithms. We record two sets of videos (robot and human performing similar tasks) of natural motions, to cover a wide range of challenges. The tasks are natural table top manipulations frequently performed in our daily life. 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. To cover this subtle difference we record both sets of videos. All the videos are tagged with the task performed and the challenges that a tracker would face while tracking. Using these tags susceptibility of the tracking algorithm can be properly narrowed down and subsequently improved. Each task is repeated with different speed and under different light.

*This work is supported by NSERC and the Canadian Space Agency (CSA)

[1]Authors are with the Department of Computing Science, University of Alberta, Edmonton, AB, T6G2E8, Canada, `ankush2@ualberta.ca`

Fig. 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 [12]

Ground truth data is made available for all sequences that are reported. We define an error metric and analyse some of the existing trackers in the literature as an example. However, since ground truth data is provided users can also define their own error metrics and analyse trackers based on other evaluation schemes.

Previous works either use synthetically generated data by applying random warps [9], [11] or test their algorithms on a small set of videos that are either recorded or pooled from the Internet. Zimmerman et al. [10] 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 that they used to evaluate their tracking system. Some other popular sources that report videos for tracking are by Babenko et al. [4], Ross et al. in [5], Kalal et al. [6], Collins et al. in PETS2005 [7], Fisher et al. in CAVIAR [18] and Dalal et al. [8]. These datasets have sequences to test tracking algorithms mainly developed for surveillance applications.

In a recent paper Yang et al. 2013 [3] categorized some of the above mentioned video sequences to publish a common benchmark. Any new tracker can be evaluated and compared with other state of the art trackers in the literature using this benchmark. However, these video sequences are more suited to surveillance tracking. Furthermore, as we elaborate on in Section III, the accuracy measures used are too coarse and ill suited for manipulation tasks.

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 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 contribute a dataset of 100 videos for tracking objects with high pixel precision and high DOF state like Metaio [2]. However, unlike Metaio our videos have meaningful motions that are part of everyday manipulation tasks and are carefully chosen so that a wide range of challenges are covered. All the videos are of 3D objects moved in an environment with natural foreground and background. We provide ground truth data for all the videos. Users can use this data to evaluate their trackers using different error metrics and evaluation schemes or train if the algorithm involves a training phase. This makes the dataset more versatile compared to Metaio who withheld ground truth in order to make the evaluation more secure. Proper categorization is also ensured with each video tagged with the challenges that it present. This help in analysing a tracker's performance. Experimentally we analyse some of the most popular patch tracking algorithms, compare and uncover properties. The experimental comparison illustrates the use of the dataset and provides initial results, that others are encouraged to build upon by using the dataset.

II. Dataset

The dataset consists of image sequences recorded by a human and a robot arm.

A. Video Capture Set Up

All videos were recorded using a GRAS-20S4C-C firewire 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 with each sequence recorded at both the light settings. 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 I.

TABLE I
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

Figure 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 reported in Table I.

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

B. Motion types in video database

We focus on evaluating trackers to handle "translation (TR)", "rotation (RO)", "scale (SC)", "perspective (PR)", "occlusion (OC)", "specular reflection", "texture (TX)", "illumination (IL)" and "speed (SP)" variations. The tasks were selected in such a way so that one or more of the above challenges are covered (Table II) while performing each task.

The recorded videos were grouped under two broad categories *Single Motion Tasks* and *Composite Motion Tasks*. *Single Motion Tasks* refers to highly structured motion of the object. On the other hand *Composite Motion Tasks* have videos that can be decomposed into simpler *Single Motion Tasks*.

1) Single Motion Tasks:

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

b) Cereal: This task is similar to that of the juice box defined early, but with an additional challenge of tracking the cereal box even during large motion when cereal is poured.

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

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

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

f) Mug I: Pure translational (TR) motion of the object (coffee mug) is recorded. An additional challenge in this case is to track in the presence of specular reflection (SR) and low texture (TX).

g) Mug II: This sequence combines the challenges of Mug I sequence with perspective (PR) deformation of the object. The object (coffee mug) is not only lifted up but also tilted to drink the contents.

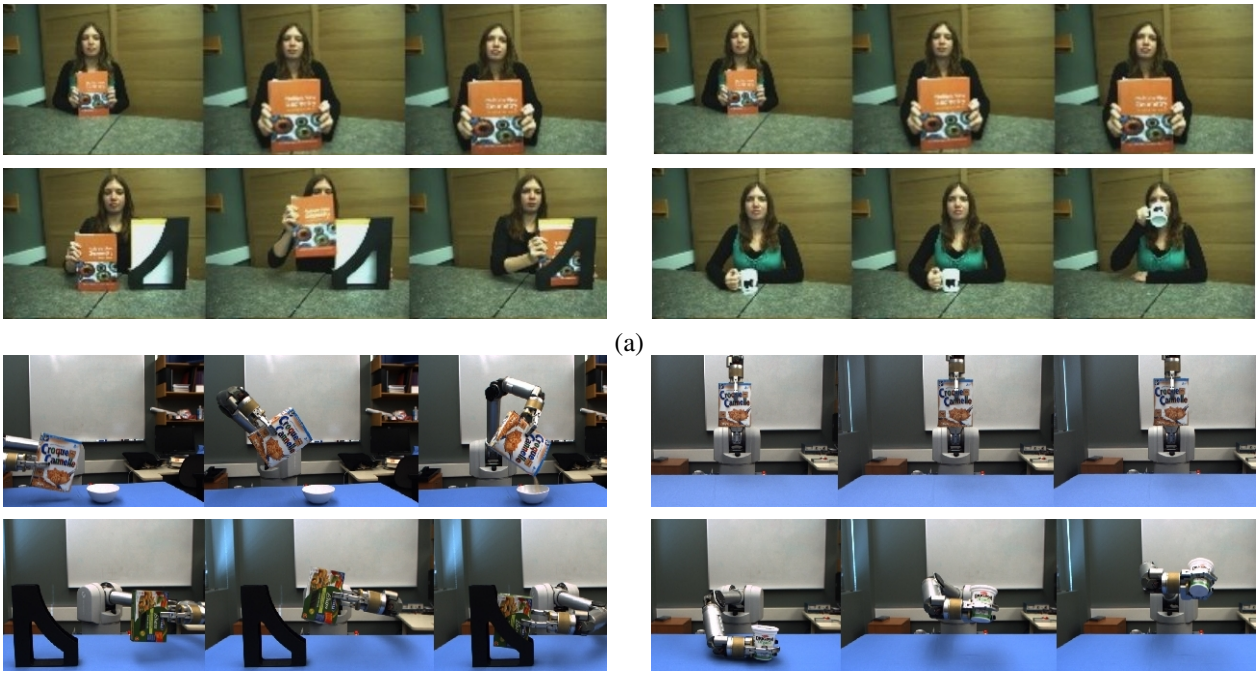


Fig. 3. Identical tasks are performed by a human and a robot arm. (a) shows the tasks performed by a human (Shaking, Moving, Placing, Drinking), corresponding tasks executed by the robot arm are shown in (b)

h) Mug III: A low texture (TX) coffee mug is rotated (RO) from its initial position and back again. The axis of rotation being perpendicular to the principal camera axis.

2) Complex Tasks:

a) Bus: A toy bus is moved around on a table. The goal is to track the planar surface in front of the bus that undergoes scale change (SC) and perspective deformation (PR) at varying speed (SP).

b) Highlighting: A portion of the newspaper is highlighted using a marker pen. The challenge is to track the object in the presence of varying texture (TX) caused by highlighting and occlusion (OC) caused by the portion of the pen within region of interest.

c) Letter: The object to be tracked here is a part of the envelope. The sequence records a letter being put inside the envelope and out respectively. The challenge in this case is largely to tackle the perspective deformation (PR) of the object.

d) Newspaper: A portion of the newspaper is supposed 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 motion to very fast motion of the object. Table II summarises the challenges that a tracker is likely to face. It also matches the task with the respective sequence. Some of the tasks are shown in Figure 3.

TABLE II
DESCRIPTION OF VIDEOS

Video	Object	Challenge	Task
Juice	Juice Box	TR,RO,SP,IL	Pouring
Cereal	Cereal Box	TR,RO,SP,IL	Shaking
Book I	Book	PR,SP,IL,SR	Lifting
Book II	Book	SC,SP,IL	Moving
Book III	Book	TR,OC,SP,IL	Placing
Mug I	Coffee Mug	TR,TX,SP,IL,SR	Raising
Mug II	Coffee Mug	TR,PR,TX,SP,IL,SR	Drinking
Mug III	Coffee Mug	RO,TX,SP,IL,SR	Rotating
Bus	Toy Bus	TR,SC,PR,SP,IL	Shifting
Highlighting	Newspaper	OC,SP,IL,TR	Marking
Letter	Envelope	PR,SP,IL	Putting
Newspaper	Newspaper	PR,SC,SP,IL	Reading

Acronym for challenges are as follow - translation (TR), rotation (RO), scale (SC), perspective (PR), occlusion (OC), illumination (IL), speed (SP), texture (TX) and specular reflection (SR)

III. Ground Truth and Error Metric

A. Ground Truth

Ground truth (GT) refers to the four co-ordinates positions of the bounding box, denoting the object's location in the image plane. Lieberknecht et al. [2] used a precisely calibrated set-up where stored pose information was used to generate ground truth data. Both Lieberknecht et al. [2] and Zimmerman et al. [10] also use visual markers to verify ground truth. In our dataset we report sub-pixel level co-ordinate positions solely based on tracking data. Three trackers [12], [13], [14] were initiated on the first frame. Ground truth is registered only when bounding box co-ordinates reported by all of them lie within 1 (± 1) pixel variation. The reason for choosing the following trackers were primarily because of their high convergence which is further illustrated in Table III.

One of the challenges that we faced while generating ground truth, was the highly sensitive tracker convergence requirement (± 1 pixel). So we reinitialize the trackers [12], [13], [14] 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.

B. Error Metric

Evaluation of a tracker first involves defining an error metric. Several error metrics are proposed in the literature. Some of the commonly [3] used ones are centre distance of the object (X_{CT}) from ground truth (X_{CGTi}) expressed as $E_C(E_C = \sqrt{(X_{CT} - X_{CGTi})^2})$ and area overlap E_A (E_A defined as $E_A = \frac{|A_T \cap A_{GT}|}{|A_T \cup A_{GT}|}$ where A_T is area of the target and A_{GT} area of ground truth). However, we define the error metric as the one used in [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}).

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

An advantage of using this score is that it correctly captures the misalignment information as shown in Figure 4. In this case the other two errors E_C and E_A would have very low error but E_{AL} would have a high error score. This is justified by the fact that the target image (T) is 180° out of phase with the ground truth. Using an alignment based error also helps in robotic tasks where an end effector needs to be aligned accurately before the task is performed like grasping or placing objects.

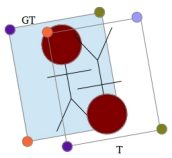


Fig. 4. Area overlap error measure. An error metric that looks only at displacement of the centre (E_C) point or area overlap E_A will have very small error in this case, but it clearly shows that the pose of the target image (T) image is 180° out of phase with the Ground Truth (GT). An error metric like E_{AL} will help to capture such misalignment

IV. Tracker Evaluation

A. Evaluation Measures

Several tracker evaluation strategies exist in the literature. Users are also encouraged to define their own evaluation criteria. Here as an example we evaluate five trackers based on two measures *Overall Success (Robustness)* and *Average Drift(Convergence)*. *Overall Success* of a tracker is defined as the fraction of frames that a tracker tracks within t_p pixel threshold. More specifically, 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 a frame f^i). *Average Drift* on the other hand 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. [3] used a pixel threshold $t_p = 20$ pixels. This is far too high for manipulation tasks. We set threshold t_p to be 5 pixels. Table III and TableIV summarises the performance of the trackers.

B. Trackers Used

Five trackers were evaluated on the dataset. Each tracker was presented with the sequence and the bounding box co - ordinates of the object in the first frame. Original implementation of NNBMIC [13], IVT [5] and L1 [14] was used. Parameters were set after cross validation to ensure best results. A single run through the sequences was done and results were further analysed.

1) Registration Based: These are 2D registration based patch trackers. We use ESM [12], NNBMIC[13] and BMIC [14], from the large number of template based trackers in literature. 10 iterations were used in BMIC for convergence. In NNBMIC 4000 $i.i.d.$ samples were used to build the initial warp index.

2) Online Learning Based: Two online learning based trackers were used, IVT by David Ross [5] and L1 tracker by Mei et al. [14]. Yang et al. in [3] analysed some of the state of the art trackers that used online learning to update the appearance model. We choose two of the trackers that performed well in most of the videos in [3]. Both these trackers used 600 particles for their particle filter search method. Additionally L1 tracker used a template subspace of 10 templates of the target image.

C. Result Analysis

The evaluation measures (Average Drift and Overall Success) as described earlier in Section IV-A are calculated for all the videos (averaged over very slow, slow and medium speed) and reported in Table III and Table IV.

TABLE III
Average Drift Expressed as an Expectation Value

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

Values in each field represent average drift, averaged over the frames where e_{AL} was less than t_p pixels. The values ($\frac{pixel\ drift}{frame}$) indicate convergence of the trackers. BMIC has the higher convergence among the registration based trackers for most of the sequences. Compared to L1 tracker IVT has higher convergence.

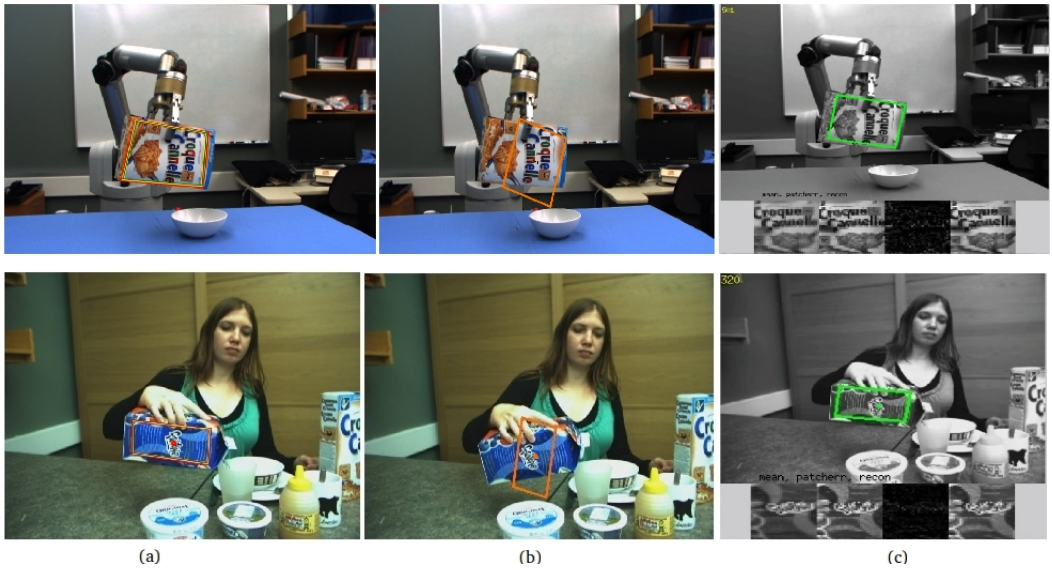


Fig. 5. An instance of in plane rotation where L1 (orange) [15] (b) fails but the three homography based trackers i.e. ESM (red) [12], BMIC (yellow) [14] and NNBMIC (cyan) [13] (a) and IVT (green) [5] (c) performs without considerable drift

Registration based trackers in general have a low average drift (high convergence) compared to the online learned trackers for all the sequences. The reason being, particle filter uses random sampling 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. Not only this the 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 (Figure 6), this comes at the expense of convergence. This is because newer templates (shifted, scaled from the original) which are not exactly the same as the target image now also form a part of the appearance model.

TABLE IV
OVERALL SUCCESS EXPRESSED AS FRACTION OF FRAMES TRACKED

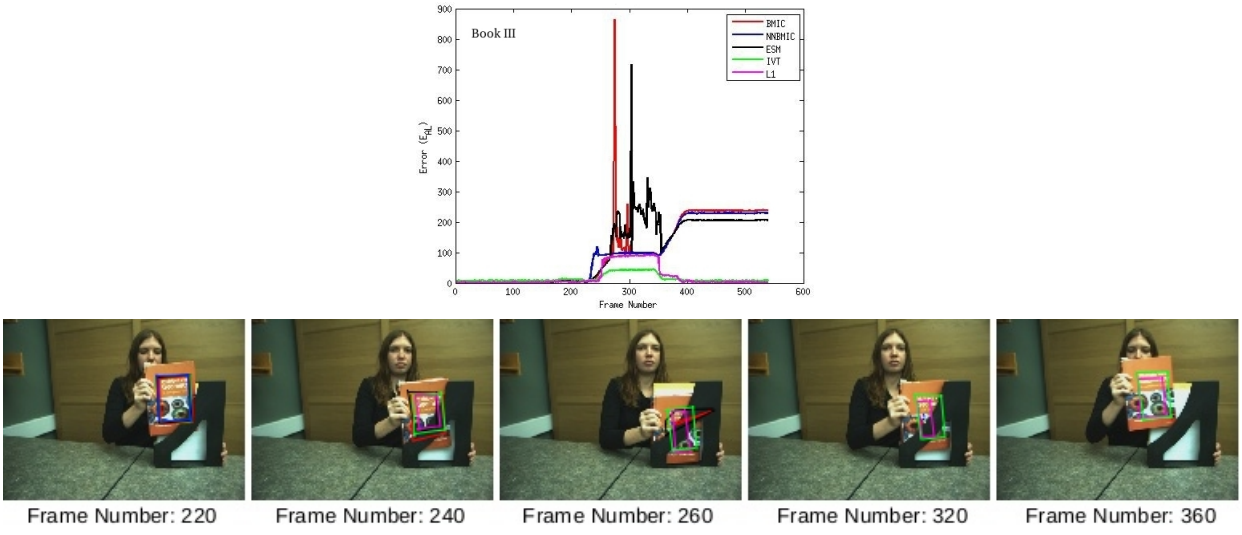
Sequence	L1	IVT	ESM	NNBMIC	BMIC
Cereal	0.24	0.99	**1**	**1**	**1**
Book I	0.10	0.477	**1**	**1**	**1**
Book II	0.79	0.3018	**1**	**1**	**1**
Book III	0.42	**0.72**	0.34	0.32	0.32
Juice	0.16	0.98	**1**	0.41	**1**
Mug I	0.10	0.91	**1**	**1**	**1**
Mug II	0.30	0.72	**0.89**	**0.89**	**0.89**
Mug III	0.54	0.68	**1**	0.59	0.65
Bus	0.57	0.94	**1**	0.99	0.96
Highlighting	0.67	**0.95**	0.76	0.70	0.33
Letter	0.19	0.25	**1**	**1**	**1**
Newspaper	0.61	0.92	**1**	0.51	0.43

Values in each field represent the fraction of successfully tracked frames. Within reasonable speed of object motion ESM converges most frequently followed by NNBMIC and then BMIC with threshold t_p set to 5 pixels. Among the online learned trackers IVT converges better than L1 but less than the registration based trackers.

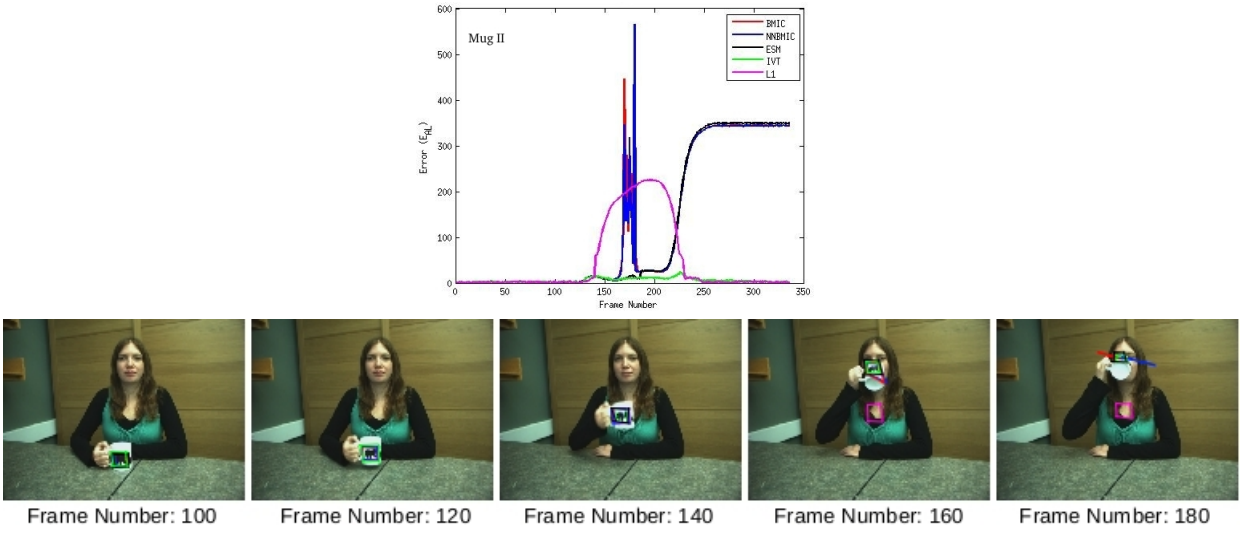
Overall success evaluates a tracker's robustness. For very small error margins ($t_p = 5$) and reasonable speed of object motion registration based trackers have a higher success rate for most of the sequences. Among the two online learned trackers IVT has a higher overall success rate compared to L1. A careful look at Table IV and Table II narrows down the reason for this. Challenge tags for the sequences where L1 has lower success rate is seen. A common set of challenges where it fails are rotation (RO) and specular reflection (SR).

L1 tracker builds a template subspace that models the appearance of the object. This template set is updated whenever it encounters occlusion. With a small template set of 10 templates, the algorithm fails to model the appearance change and hence drifts when appearance changes considerably. The inability of L1 tracker to handle rotation on the other can be attributed to the fact that with each iteration only the translation parameters are updated in the 6DOF state transition model described in [21]. The algorithm does not tackle full (6DOF) affine transformation of the object and hence the rotation parameters are not updated with each iteration.

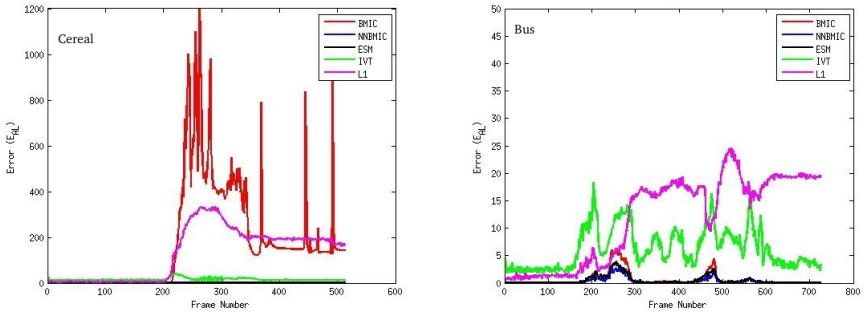
At moderate speeds though the registration based trackers have a higher overall success and low average drift, at high speed the nature changes. Overall accuracy for the registration trackers decreases as speed increases. In Figure 6 we see error plots (E_{AL} vs Frame number) of the trackers on four sequences having high speed object motion. Due to space limitations only a representative sample is chosen, so that all challenges as listed in Table II can be addressed. Figure 6a. shows that with adaptive learning of appearance change trackers can handle occlusion better than the registration based trackers. Although all the trackers drift substantially between frame number 200 and 300 i.e. the point where the book holder starts occluding the object



(a)



(b)



(c)

(d)

Fig. 6. Error (E_{AL}) profile (E_{AL} vs Frame Number) for each of the five trackers 1) BMIC (red), 2) NNBMIC (blue), 3) ESM (black), 4) IVT (green) and 5) L1 (magenta) is plotted for four sequences (BookIII (a), MugII(b), Cereal(c), Bus(d)). (a) points out that IVT and L1 handles occlusion (OC) better than the registration based trackers. Some of the frames are shown where the trackers fail due to substantial occlusion. (b) shows that L1 is incapable of handling large change in appearance due to specular reflection (SR). The frames below show that this happens particularly when the mug is tilted. (c) brings out the inability of L1 to track in plane rotation (RO). In (d) we see that the average drift of the registration based tracker (Gauss Newton search) is lower compared to the particle filter based trackers (IVT and L1) within the operating range of 5 pixels.

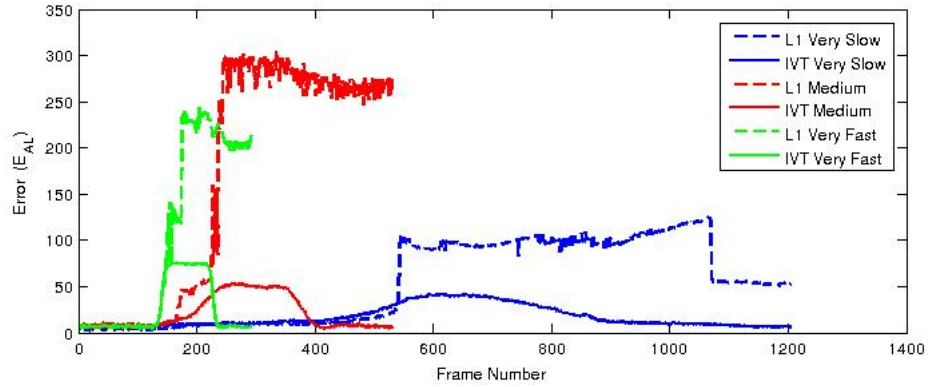


Fig. 7. Two different state of the art trackers i) IVT [5](-) and ii) L1 Tracker (- -) [15] are tested on the book sequence (bookI) under normal light. The video is recorded at five different speeds (SP), tracking results on all of the speeds are shown. Absolute error is plotted against the frame number. Here L1 doesn't converge back once it drifts whereas IVT does converges.

(Images shown below the graph), IVT and L1 recover and track for the rest of the sequence. From Table IV we can see that IVT handles occlusion in the best way possible among the trackers studied. Its overall success is higher even in slow object motion for this challenge (BookIII and Highlight). Figure 6b. validates the point that L1 fails to track when appearance of the object (mug) changes considerably. In this case, because of specular reflection (SR) on the mug (Images shown below the graph). Figure 6c. shows another instance of the L1 tracker failing to handle rotation. Finally Figure 6d. brings out the fact that the registration trackers with a Gauss Newton search step do operate with a low average drift when they converge.

Study of object speed (SP) variation is shown in Figure 7. Only the online learned trackers are used in this study. From the graphs of Figure 6. it is clearly established that these trackers are better in handling high speed object motion. So, we do a further study to show which among the two (IVT and L1) is better in doing so. We plot the error (E_{AL}) profiles (error vs frame number) for three different speeds(very slow, medium and very fast). The behaviour of the trackers are qualitatively same for other speeds which we don't plot. Both these trackers do fail initially but IVT recovers whereas L1 fails to do so which certifies that IVT is more robust compared to L1.

V. Conclusion and Discussion

We report a dataset of 100 video sequences to evaluate trackers meant for manipulation tasks. The dataset consists of videos recorded by both a robot and human to cover a wide range of challenges. Complete ground truth data is made available for all the sequences. The source codes 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 have provided an initial analysis of some of the popular trackers in the literature. Some interesting results are revealed during this analysis. It is observed that adaptive learning model of IVT helps it to recover from drift and also successfully handle occlusion at high speeds. However with a particle filter search method they have a higher average drift compared to the registration based tracker which uses a Gauss Newton search. The susceptibility of L1 to in-plane rotation and large appearance change is also narrowed down using the proper categorisation of the dataset.

In future we would like to increase the set of trackers we have analysed. Users of the dataset are also welcome to report meaningful analysis of their trackers (Ground Truth is made public) thus building a common benchmark that would help the community.

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.

[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). IEEE.

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

[4] Babenko, B., Yang, M. H., Belongie, S. (2009, June). Visual tracking with online multiple instance learning. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 983-990). IEEE.

[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.

[6] Kalal, Z., Mikolajczyk, K., Matas, J. (2012). Tracking-learning-detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 34(7), 1409-1422.

[7] Collins, R., Zhou, X., Teh, S. K. (2005, January). 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] Dalal, N. (2005). INRIA person dataset. 2011-04-10. http://pascal.inrialpes.fr/data/human

[9] Baker, S., Matthews, I. (2004). Lucas-kanade 20 years on: A unifying framework. International Journal of Computer Vision, 56(3), 221-255.

[10] 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 http://cmp.felk.cvut.cz/demos/Tracking/linTrack/data/index.html

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

[12] 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.

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

[14] 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.

[15] Mei, X., Ling, H. (2009, September). Robust visual tracking using L_1 minimization. In Computer Vision, 2009 IEEE 12th International Conference on (pp. 1436-1443). IEEE.

[16] Stenger, B., Woodley, T., Cipolla, R. (2009, June). Learning to track with multiple observers. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 2647-2654). IEEE.

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

[18] Caviar datasets available at `http://groups.inf.ed.ac.uk/vision/caviar/caviardata1/`

[19] Coriander, IEEE1394 Camera Gui, Available at: `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