2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance

Thirteen Hard Cases in Visual Tracking

Dung M. Chu and Arnold W.M. Smeulders
Intelligent Systems Lab Amsterdam (ISLA), University of Amsterdam
Science Park 107, 1098 XG, Amsterdam, The Netherlands
{chu, A.W.M.Smeulder}@uva.nl

Abstract

Visual tracking is a fundamental task in computer vision. However there has been no systematic way of analyzing visual trackers so far. In this paper we propose a method that can help researchers determine strengths and weaknesses of any visual tracker. To this end, we consider visual tracking as an isolated problem and decompose it into fundamental and independent subproblems. Each subproblem is designed to associate with a different tracking circumstance. By evaluating a visual tracker onto a specific subproblem, we can determine how good it is with respect to that dimension. In total we come up with thirteen subproblems in our decomposition. We demonstrate the use of our proposed method by analyzing working conditions of two state-of-the-art trackers.

1. Introduction

While tracking has made considerable progress over the last few years, there is a lack of systematic evaluation benchmarks. In this paper we argue for the need of a systematic analysis of visual trackers. In this way progress towards robust and general trackers can be documented in terms of their conditions of functioning and failure.

Current trends in evaluating trackers are either using self-made videos or using benchmark datasets. Using self-made videos [9, 8, 14] is a good way to pinpoint the strengths of a tracker. However one can hardly assess what the working conditions precisely are as for such an assessment it is equally important to know the conditions of failure. Using benchmark datasets [15, 4, 5, 2] with evaluation metrics [11, 6] provides common frameworks to compare trackers. However state-of-the-art benchmark datasets still suffer from limitations. Firstly, datasets are task-specific with rather limited visual repertoires. Many of the datasets [6, 17, 18, 1] are surveillance data with static cameras, relatively stable background, rigid, opaque objects only. Trackers are left untested outside this domain. Secondly, the existing datasets do not yet sample the full breadth of tracking conditions. This paper aims to take a step into the direction of creating a systematic evaluation benchmark, which covers a broad range of tracking circumstances. The videos in the benchmark are ordered based on their difficulty levels.

In this paper we ask the question: *how to get a good sense of the full breadth of conditions of proper functioning and failure of a visual tracker?* Our contributions are: (1) we discuss thirteen different tracking conditions in a systematic way; (2) we collect data for each condition and order them from the easiest ones to the more difficult ones. The data are divided into two classes: laboratory videos and realistic videos; (3) we evaluate two state-of-the-art trackers on this dataset.

2. Related work on Benchmark Datasets

There are two trends in evaluating tracking algorithms: one is by considering tracking as a low-level task in a higher-level purpose. The performance is evaluated at the higher-level task. The other trend is to consider tracking itself as an end-level task and evaluate the performance directly. In this paper, we contribute to the latter.

Performance Evaluation of Tracking and Surveillance (PETS) [6] is a series of workshops devoted to boost performance evaluation of tracking and surveillance. Each year particular surveillance challenges are posed. Most of the PETS datasets are surveillance data and the challenges are focused on security issues on public places, *e.g.* detect lost luggages, detect loitering. Many tracking scenarios have been considered. The datasets are however not designed to measure trackers' ability to cope with different tracking conditions. We use some videos from PETS2007 and PETS2009 in our realistic dataset.

ETISEO [13] is a performance evaluation framework for visual tracking, object detection, localization, classification and event recognition in video surveillance systems. It aims at identifying suitable scene characteristics for a given video processing algorithm and to highlight algorithm weaknesses by underlining the dependencies between the algorithm and its conditions of use. ETISEO addresses two problems, namely shadows and weakly contrasted objects. We aim to

978-0-7695-4264-5/10 $26.00 © 2010 IEEE
DOI 10.1109/AVSS.2010.85

investigate the problem of determining strengths and weaknesses of trackers with a broader range of tracking conditions.

In ICCV 2009, the LabelMe Video [19] was introduced as an extension of the LabelMe image dataset. This ongoing project will allow internet users to upload their own videos and do annotation online. The purpose of the project is to get the prior knowledge of motion, location and appearance at the object and object interaction levels in real-world videos. The absence of systematic sampling is its weakness.

Video Verification of Indentity Databases [4] provides an opensource tracking testbed, which allows researchers to run and log tracking experiments, and compare different trackers. The dataset contains 10 video sequences. The dataset focuses on tracking ground vehicles from airborne cameras limiting the repertoire to multiple similar objects, moving cameras with occlusions.

Classification of Events, Actions and Relationships (CLEAR) [17, 18] and SPEVI [12] consider person tracking, face tracking and vehicle tracking. The datasets are recorded indoor from multiple fixed cameras. CLEAR also proposes a novel tracking metric, which calculates the basic types of errors made by multiple object trackers.

In summary, all existing datasets are developed in depth for specific high-level purposes. This has its own advantages. LableMe Video [19] has potential to cover the full breadth. However it lacks a systematic approach. We aim to cover a broad set of variations and also introduce a degree of difficulty for each dimension. Hence we are able to validate tracking algorithms in breadth.

3. Categorization of Thirteen Different Conditions in Visual Tracking

For a visual tracking system, four main dimensions will determine the complexity of the solution: light source, scene, objects and camera. Each of these dimension has a number of degrees of freedom influencing the performance of the tracking system, see Figure 1.

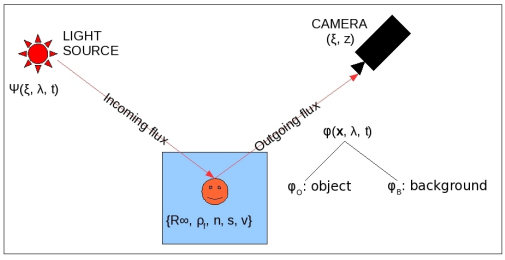

Figure 1. Photometric model and notations used in this paper.

1. We model the light source as incoming flux $\psi(\xi, \lambda, t)$, which is the spectral power distribution of light at spatial position ξ at time t. Variabilities may originate from changes of ψ over position ξ of the light source, wavelength λ or time t.

2. A point in the scene is modeled by a set of five most relevant parameters $\{R_\infty, \rho_f, n, s, v\}$. Fresnel surface reflectance ρ_f indicates how the incoming light is reflected at the surface of the scene. Body reflectance R_∞ describes how the incoming light is refracted to the object's body at the incidence point. Refractive index n indicates how transparent the scene at the incidence point is.

3. The light going from the scene to the camera is modeled by outgoing flux $\varphi(\mathbf{x}, \lambda, t)$, where $\mathbf{x}$ is 2-D position in the image. We denote by φ_O and φ_B the outgoing fluxes from the object and the background respectively.

4. The camera is modeled by its position in the scene and zooming level.

In total we will consider thirteen dimensions as the most relevant ones (see Table 1), where we will leave out in this paper less important degrees of freedom, such as backscatter, non-white balanced cameras and other more complicated models for the object, the scene and the camera. We acknowledge that there are interdependencies between the thirteen dimensions. We aim to start from the point where these interdependencies are low (so that they can be studied independently) and leave more complex interrelations for later. We now discuss the thirteen dimensions in detail.

(1) Light: the light source is modeled by $\psi(\xi, \lambda, t)$. We distinguish three situations:

Uneven light source: we consider changes of incoming light flux ψ over spatial position ξ resulting in uneven illumination of the object. Modelling the object appearance in the trackers is complicated due to the uneven object appearance. Trackers using point or line representations suffer this condition as it creates lots of artificial points and lines in the image. An uneven light source may also induce false movement of object albedo and hence cause drift. The fraction of unevenly illuminated areas is the ordering measure in this subcategory. A hard case of uneven light source is a scene under foliage, see Figure 2.

Light color: we consider changes of incoming light flux over wavelength λ. The object is illuminated by different light colors over time. We choose the speed of the color's change as the ordering parameter. A hard case of light color is a scene illuminated by city lights.

Changing light intensity over time: the object is illuminated with lights of intensities ψ varying over t. This results in different appearances of the object and of its shadows. The speed of the change determines how severe the sequence is. A hard case of changing light intensity over time is a scene under flash light.

Table 1. Description of the 13 categories. "Uneven" implies spatial changes. "Unstable" implies temporal changes.

No	Aspects of tracking	Effect on observed field	Ordering parameters	Examples of hard cases
1	Light	Uneven and unstable light	Fraction of unevenly illuminated area; speed of light color or intensity changes	Disco light
2	Multiple light sources	Multiplicity of shadow	The number of light sources	Mist
3	Albedo	Uneven and unstable albedo	Changes of R_∞	Person redressing
4	Specularity	Uneven and unstable specularity	Fraction of specular highlights	Mirror
5	Transparency	Uneven and unstable transparency	The amount of transparency n	Transparent ball
6	Shape	Uneven and unstable shape	Convexity complexity [20]	Octopus
7	Motion smoothness	Unstable speed	Object motion smoothness	Brownian motion
8	Motion coherence	Uneven motions of object parts	Variation of part motions	Flock of birds
9	Clutter	Clutter of object in background	Bravo and Garid [3]	Camouflage
10	Confusion	Similarity between objects	SSD similarity	Herd of cows
11	Occlusion	Object's presence or absence	Fraction of the occluded area	Object getting out of scene
12	Moving camera	Unstable camera position	Camera motion smoothness	Shaking camera
13	Zooming camera	Unstable zooming	Zooming speed	Abrupt close-up

Figure 2. The trellis sequence from Ross *et al*. [15] as an example of severe uneven light source. The trellis makes the illumination to the face uneven.

(2) Multiple light sources coming from many directions: the main issue with multiple light sources is that they create multiplicity of shadows or no shadow at all. For omnidirectional illuminant, it takes away shading, and hence there is no geometrical detail. Many trackers implicitly rely on shading to prevent drift. We can order sequences by the number of light sources in each sequence. A hard case of multiple light sources is mist with indefinitely many light sources.

(3) Albedo: when body reflectance R_∞ changes over time the object albedo changes as well, which induces changes in object appearance. The change of R_∞ determines the order of sequences. A hard case of body reflectance is a chameleon changing its skin to adapt with environment. Another example is a person redressing.

(4) Specularity: specularity is indicated by the Fresnel surface reflectance ρ_f. Specular-reflected light is the result of the light source, the geometry and the pose of the object. Specular highlights occlude the real object and bring abrupt changes. The fraction of specular highlights determines the video sequence order. Hard cases of specularity are mirrors.

(5) Transparency: we use the refractive index n to measure the transparency of an object with respect to the surrounding environment. With transparent objects there are 2 movements of 1 pixel. The refractive index n is the ordering parameter in this category. A hard case of transparency is a glass-like object.

(6) Shape: many trackers track objects by following the development of their shapes. The underlying assumption is that the objects' shapes do not change abruptly over time. The convexity measurement [20] of the object shape is the ordering parameter in this category. A hard case is a moving octopus.

Rotating of a non-convex object may also induce abrupt shape changes. In this situation, the projected shape is very different from the real 3D object shape and unstable inherently. Tracking a rotating mouse pad is an example.

(7) Motion smoothness: motion is a very useful cue in predicting future positions of the object. When the object's motion is smooth in both velocity and direction, prediction may be successful. The smoothness of the object motion determines its order. A bouncing ball is an example of hard cases. Another example is a Brownian-moving object.

(8) Motion coherence: with an articulated object, it may happen that the object parts move with different motions. Variation of the part motions determine the state of the object motion coherence. A hard case of motion coherence is a flock of birds or fireworks.

(9) Clutter: clutter occurs when the neighborhood background has many different patterns. We use the clutter measure in Bravo and Garid [3] as the ordering parameter. A

hard case of clutter is camouflage.

(10) Confusion: confusion happens when there are similarly-appearing objects close to the object of interest. The similarity causes trackers to confuse the object of interest others. We use the sum-of-squared-difference similarity measure between the objects to order the video sequences. A hard case of confusion is a herd of cows or zebras.

(11) Occlusion: we use the ratio between the area of the occluded portion and the whole object area to measure occlusion. A hard case of occlusion is a object getting out of the scene and returning in a different place, pose and illumination.

(12) Moving Camera: a moving camera induces changes in the object and also the background. We use the smoothness of the camera motion as the ordering parameter. An obvious hard case is a shaking camera. In this case, the object movement is very fast.

(13) Zooming Camera: when a camera zooms, it changes scale of the whole scene. The zooming speed determines how fast the change is. A hard case of zooming camera is abrupt close-up.

The thirteen categories are summarized in Table 1. The fourth column of the table are measures expressing the severeness of the dimension. The last column lists one example of each hard case, which indicate extreme cases. These are hard situations for visual tracking.

4. Datasets

4.1. The Laboratory Data

We setup a recording environment to make videos featuring the 13 dimensions. The recording setup is similar to the one used in [7]. In most cases we use one light source, which is mounted at a corner of the recording table. When special light conditions are needed we use strobe lights with different colors and mount them on the half-circle ring on top of the table. We use a ruler, a checkboard and a color checker to calibrate the camera. A single background, see the upper half of Figure 3, is used for all the cases (except for the clutter and confusion cases where the background needs to be changed).

The lower half of Figure 3 shows some of the objects that we used. For all the recordings, we roll the objects from the open end to the other end of the table. Table 2 gives an overview of the laboratory videos (since we have not been able to create a mist-like lighting condition in lab, we decided to leave out the multiple light category. However we do collect realistic data for this category). In making the laboratory videos, we vary the condition of interest and try to keep the other conditions constant as much as we can. In each category, the videos are divided into three subcategories, namely high, medium and low with decreasing level of difficulty corresponding to their ordering parameter values.

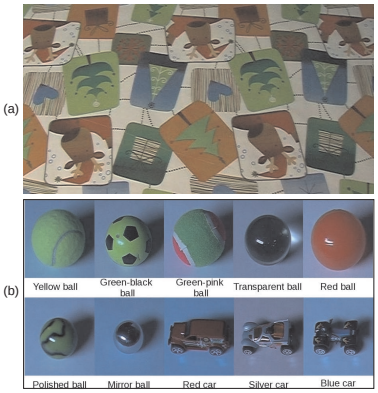

Figure 3. (a): the background used in making the laboratory data. When cluttered backgrounds are needed, we add more small objects with different shapes and color to the background; (b): some of the objects used in the recording.

4.2. The Realistic Data

In addition to avoid the peculiarities of only laboratory data we also collect a dataset following the 13 dimensions from realistic videos. The main sources of videos we collect are: existing tracking datasets [13, 12, 6, 16], YouTube and videos used in some tracking papers. In selecting realistic videos, we try to search for videos that feature the condition of interest and contain as few other conditions as possible. For example, with the occlusion category, a video from the SPEVI dataset [12] are chosen as it was made deliberately by the authors with the purpose of studying occlusion.

In total we have two datasets of about 6GB with approximately 100 videos. For both datasets annotation is done by a rectangular box every fifth image.

5. Performance Evaluation

5.1. Measures

We adapt the evaluation measures proposed by Kasturi *et al.* [11] and Kao *et al.* [10] to our framework. Let G^t denote the ground-truth object in frame t; D^t denotes the tracked object in frame t and N_{frames} denotes the number of frames where either the ground-truth object or the tracked object exists.

Average tracking accuracy (ATA) proposed in [11] is defined as the average ratio of the spatial intersection and union of the ground-truth object and the tracked object over all the frames.

$$ATA = \frac{1}{N_{frames}} \sum_{i=1}^{N_{frames}} \frac{|G^t \cap D^t|}{|G^t \cup D^t|}. \quad (1)$$

As in Kao *et al.* [10], we also want to define ROC-like curves for trackers' performance. To this end, we define

Table 2. Overview of the videos in the laboratory dataset.

Category	laboratory Data
Light	Yellow ball, red ball and green-black ball. Flashing light is made with a strobe. Color light is made with a color strobe. Uneven illuminated light is made with a board that allows light go through some holes on it.
Albedo	Yellow ball, green-pink ball and a ball with 4 different colors.
Specularity	Yellow ball, red ball, polished ball and mirror ball.
Transparency	Yellow ball, transparent ball and wine glass.
Shape	Pig toy, bear toy and beer opener.
Motion smoothness	Yellow ball and silver car. The ball moves either straight or bouncing against the wall. The car either moves straight or goes back and fourth.
Motion coherence	Yellow ball and small herd of similar balls moving together.
Clutter	Yellow ball, green-pink ball and a ball with 4 different colors.
Confusion	Yellow ball, a set of identical balls
Occlusion	Yellow ball, red ball, blue car and silver car. Occluders are chosen so that they occlude 0%, about 50% and 100% of the objects.
Moving camera	Red ball with 3 settings of the camera: stationary, smoothly moving and shaking.
Zooming camera	Yellow and green ball with 3 settings of the camera: stationary, slowly zooming and abrupt zooming.

the average tracking error (ATE) as follows

$$ATE = \frac{1}{N_{frames}} \sum_{i=1}^{N_{frames}} \frac{|D^t \setminus G^t|}{|D^t|}. \quad (2)$$

ATA and ATE can be interpreted respectively as the true positive rate and false positive rate. (ATE, ATA) together provides ROC-like curves to evaluate performance, where the top performance is the top-left corner point $(0, 1)$.

The two abovementioned measures evaluate performance of a tracker with respect to one video. With our categorization of tracking conditions, it is also interesting to consider tracking performance evaluation in category level. Category-level average tracking accuracy ($CATA$) is defined as:

$$CATA = \frac{\sum_{i=1}^{N_{videos}} ATA_i}{N_{videos}}, \quad (3)$$

and category-level average tracking error ($CATE$) is defined as:

$$CATE = \frac{\sum_{i=1}^{N_{videos}} ATE_i}{N_{videos}}, \quad (4)$$

where ATA_i, ATE_i are the average tracking accuracy and average tracking error of the tracker with the ith video respectively; N_{videos} is the number of videos of interest.

5.2. State-of-The-Art Trackers

We choose two state-of-the-art trackers to demonstrate the usefulness of our approach in gaining insight into their conditions of working and failure by doing performance evaluation from our categorization perspective of tracking conditions. The incremental visual tracker (IVT) [15] builds an object appearance model by making use of the new appearance information that comes available to incrementally improve a model of the target. Incremental PCA allows the tracker to perform an efficient subspace update.

The foreground-background tracker (FBT) [14] is a discriminant tracker. It follows an object by keeping an incremental classifier between features sampled from the object and its neighboring background. Effectively it tracks a hole in the background.

6. Results

We ran the two trackers on both datasets with the initial positions of the objects taken from the annotations. CATA and CATE values of each tracker are computed from the tracking results and the annotations.

6.1. The Orderings of the Videos

The first experiment is to see how our orderings of tracking conditions are reflected from the trackers' performance. Figures 4 and 5 show the performance of the IVT and FBT with the laboratory dataset respectively. With the IVT we observe that the ordering is reflected clearly in the light, albedo, specularity, confusion, occlusion and moving camera categories. Under mild transparency the IVT performance improves in this category, which can be explained by resistance of the object model in the tracker to uniform changes in the object appearance. In the case of the shape category, the tracker performs very well in three subcases. The lower score for simple-shape objects compared to the score for mildly-complicated shape objects is caused by the fact that our annotation are made using only vertically-

aligned rectangles. Nevertheless the difference is small. The performance in the zooming camera shows an opposite ordering, which is attributed to the use of scaling in the dynamics model of the IVT. Essentially the IVT is invariant to scaling of the object. In the "low" case of zooming camera, a moving ball is used with stationary camera while a static ball is used in the "medium" and "high" cases. The changes in appearance of the moving ball cause the low score of this sequence.

The performance of the FBT with the laboratory dataset (Figure 5) reflects our ordering of tracking conditions with exception in the cases of albedo, motion smoothness. Visual inspection of the FBT performance in the albedo videos shows that the FBT does not get affected when the albedo's change is small. However when the effect is bigger the tracker starts to perform worse. We note that with the motion smoothness case, the difference between the 3 levels is small. This is attributed to the searching strategy of the FBT, where the tracker puts equal weights in all the directions.

In conclusion, the performance of the IVT and FBT supports our assumption on the ordering of the tracking conditions with different difficulty levels.

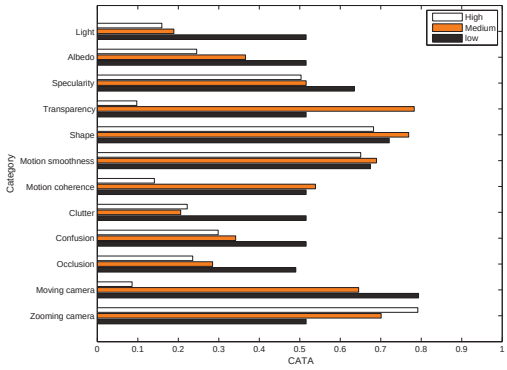

Figure 4. The performance of the IVT with respect to the categories in the laboratory dataset. The "low", "medium" and "high" bars indicate the videos corresponding to lowest, medium and highest values of the ordering measures with increasing level of difficulty.

6.2. The Correlation of the Laboratory Dataset and Realistic Dataset

We have considered the correlation between the laboratory dataset and the realistic dataset to have a first impression whether the laboratory dataset allows a similar performance as the realistic dataset. To this end, we compute the relative difference of the FBT and IVT on each dataset

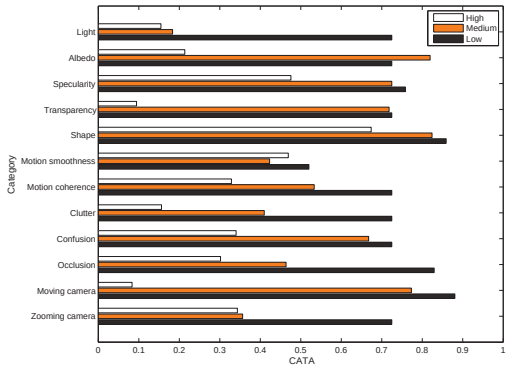

Figure 5. The performance of the FBT with respect to the categories in the laboratory dataset.

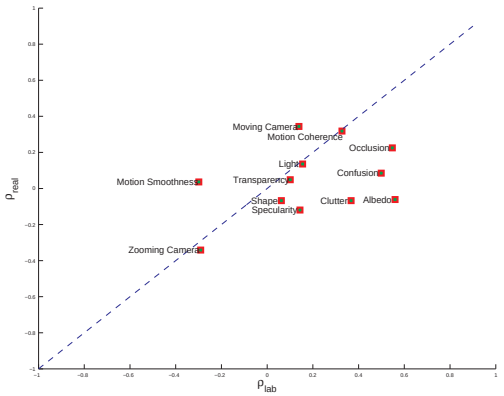

Figure 6. The correlation plot of the laboratory dataset and realistic dataset.

based on the following equations:

$$\rho_{lab}(i) = \frac{CATA^{FBT}_{lab}(i) - CATA^{IVT}_{lab}(i)}{CATA^{IVT}_{lab}(i)} \quad (5)$$

$$\rho_{real}(i) = \frac{CATA^{FBT}_{real}(i) - CATA^{IVT}_{lab}real(i)}{CATA^{IVT}_{real}(i)}, \quad (6)$$

where index i indicates one of the thirteen categories. The result is depicted in Figure 6. In general, the impression from Figure 6 is an acceptable correlation between the laboratory and realistic datasets. The number of videos ($N = 4$ in average per laboratory or realistic category) and especially the number of trackers ($N = 2$) is too low to draw general conclusion on the compatibility of the two datasets yet.

6.3. Assessing Strengths and Weaknesses of Visual Trackers

The dataset enables us to analyze the performance of a tracker with respect to the 13 tracking dimensions. We ran the two trackers with both datasets. Figures 7 and 8 depict the results. Figure 7 gives an overview of the IVT performance. As can be seen from the figure, the IVT performs

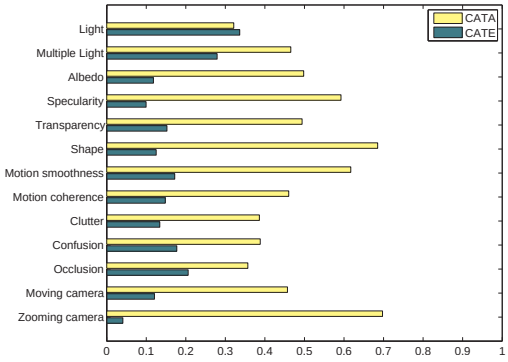

Figure 7. The performance of the IVT with the combined dataset.

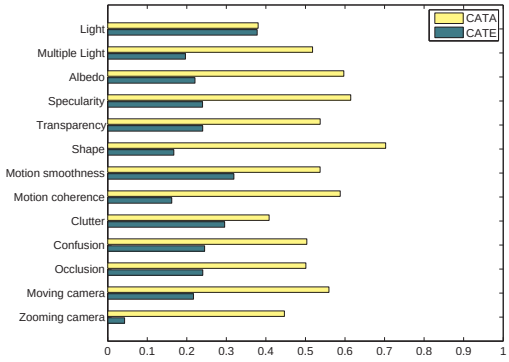

Figure 8. The performance of the FBT with the combined dataset.

best with the case of zooming camera where the tracking accuracy is high and the tracking error is very low. The specularity, shape and motion smoothness categories also get very high performance with the IVT. They are attributed to the fact that the IVT is focused on modeling object appearances and dealing with scale changes (as mentioned in the original paper [15]). The low score of the confusion (and also clutter) case suggests the inherent drawback of appearance-based tracking which does not take into account background information. The high CATE value with the light category shows that changes in lighting conditions are very challenging for the IVT.

From Figure 8 we observe that the FBT performs best with the shape case. This reflects the fact that the FBT effectively tracks a hole in the background. The FBT performance with the albedo, specularity, transparency is also high for the same reason. The low score of the FBT with the clutter case is caused by the large spread of the background pattern, which reduces the discrimination of the foreground and background. These properties go along with the claims in the FBT original paper [14]. Similarly to the IVT, changes in lighting conditions is also very challenging for the FBT with low CATA score and high CATE score. The low CATA score of the zooming camera case is attributed to the fact that we do not consider scaling in the implementation of the FBT.

In summary, the performance analysis derived from our dataset goes along with the claimed properties of the two trackers. We note that although the two trackers were tested with varying lighting conditions in the original papers, they do not perform well on the videos in our lighting category. This is because we created very challenging lighting conditions with changing color light, fast flashing light and foliage-like lighting in the laboratory dataset. These are in general difficult for any visual tracker.

6.4. Comparing Visual Trackers

Comparison between two trackers with respect to the 13 tracking conditions can also be done with the proposed

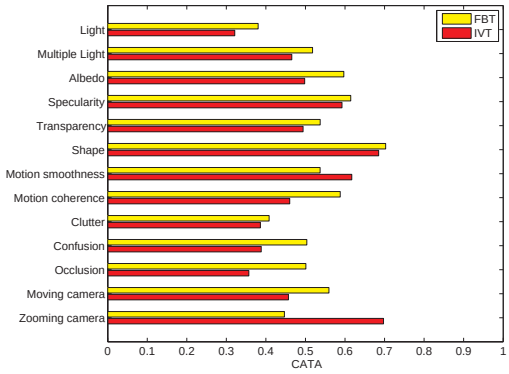

Figure 9. The true positive scores (CATA) of the IVT and FBT with the combined dataset. The FBT outperforms the IVT in most of the cases

dataset. For the IVT and FBT, the result is depicted in Figure 9 and Figure 10 for the CATA and CATE measures respectively (the data in this two figures are in fact already included in the Figure 7 and 8. We regroup them for the comparison purpose). As we can see from Figure 9, the FBT outperforms the IVT in the light, multiple light, albedo, transparency, motion coherence, confusion, occlusion and moving camera categories. The IVT is better in the zooming camera and motion smoothness categories. The Figure 10 shows that the false positive rate of the IVT is smaller than that of the FBT. This is attributed to the fact that the IVT can track objects with varying size better than the FBT.

7. Conclusions

We discuss a categorization of visual tracking consisting of thirteen representative categories in a systematic way. We propose parameters that order the space of videos in ordered sequences of complexity. For each of the tracking categories, we provide an ordered set of small laboratory videos recorded in our lab or selected from real life videos. The laboratory videos and realistic videos are shown to have acceptable correlation scores.

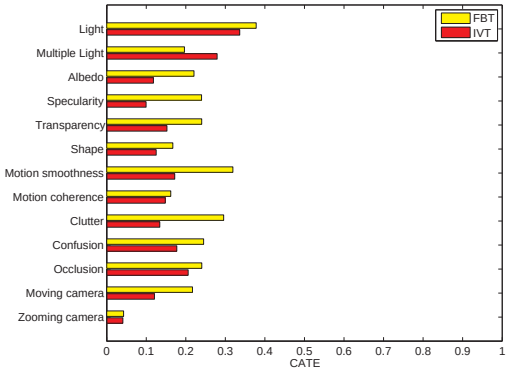

Figure 10. The false positive scores (CATE) of the IVT and FBT with the combined dataset. The IVT is in general better than the FBT in term of the false positive score.

We propose two kinds of tracking measures: category-level average tracking accuracy and category-level average tracking errors, which can be seen as true positive rate and false positive rate. The combined use of the proposed dataset and the tracking measures enables to assess a broad range of conditions of proper functioning and failure of a visual tracker.

We demonstrate usages of the datasets by analyzing and comparing two state-of-the-art trackers. The conclusions derived from our analysis go along with the claimed properties of the two trackers in their original papers. The comparison shows that while in term of true positive rate CATA the FBT is better, in term of false positive rate CATE the IVT is better.

The proposed datasets are still limited in the number of videos per category. The use of the two trackers does not yet allow us to give a general conclusion about the compatibility of the two datasets. We plan to address these issues in the future work.

The proposed datasets will be made publically available.

Acknowledgments

We thank Jan-Mark Geusebroek for insightful discussions and comments.

References

[1] Object tracking and classification beyond and in the visible spectrum. http://www.cse.ohio-state.edu/otcbvs-bench/. 1

[2] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *CVPR*, volume 0, pages 983–990, 2009. 1

[3] M. J. Bravo and H. Farid. A scale invariant measure of clutter. *Journal of Vision*, 8(1):1–9, 1 2008. 3

[4] R. Collins, X. Zhou, and S. K. Teh. An open source tracking testbed and evaluation web site. In *PETS*, 2005. 1, 2

[5] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, volume 2, pages 142–149, 2000. 1

[6] A. Ellis, A. Shahrokni, and J. Ferryman. Overall evaluation of the pets 2009 results. In *PETS*, 2009. 1, 4

[7] J.-M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The amsterdam library of object images. *IJCV*, 61(1):103–112, 2005. 4

[8] H. Grabner, C. Leistner, and H. Bischof. Semi-supervised on-line boosting for robust tracking. In *ECCV*, pages 234–247, 2008. 1

[9] W. He, T. Yamashita, H. Lu, and S. Lao. Surf tracking. In *ICCV*, 2009. 1

[10] E. K. Kao, M. P. Daggett, and M. B. Hurley. An information theoretic approach for tracker performance evaluation. In *CVPR*, pages 1523–1529, 2009. 4

[11] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *PAMI*, 31(2):319–336, 2009. 1, 4

[12] E. Maggio, E. Piccardo, C. Regazzoni, and A. Cavallaro. Particle phd filter for multi-target visual tracking. ICASSP, 2007. 2, 4

[13] A. T. Nghiem, F. Bremond, M. Thonnat, and V. Valentin. Etiseo, performance evaluation for video surveillance systems. In *AVSS*, pages 476–481, 2007. 1, 4

[14] H. Nguyen and A. Smeulders. Robust track using foreground-background texture discrimination. *IJCV*, 68(3):277–294, 2006. 1, 5, 7

[15] D. Ross, J. Lim, and R.S.Lin. Incremental learning for robust visual tracking. *IJCV*, 77:125–141, 2008. 1, 3, 5, 7

[16] S. Stalder, H. Grabner, and L. van Gool. Beyond semi-supervised tracking: tracking should be as simple as detection, but not simpler than recognition. In *IEEE Workshop on On-line Learning for Computer Vision*, 2009. 4

[17] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan. *The CLEAR 2006 Evaluation*, volume 4122 of *Lecture Notes in Computer Science*. Springer Berlin/Heidelberg, 2007. 1, 2

[18] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo. *The CLEAR 2007 Evaluation*, volume 4625 of *Lecture Notes in Computer Science*. Springer Berlin/Heidelberg, 2008. 1, 2

[19] J. Yuen, B. Russell, C. Liu, and A. Torralba. Labelme video: Building a video database with human annotations. In *ICCV*, 2009. 2

[20] J. Zunic and P. L. Rosin. A convexity measurement for polygons. In *BMVC*, volume 24, pages 173–182, 2002. 3