PFT: A PROTOCOL FOR EVALUATING VIDEO TRACKERS

Tahir Nawaz and Andrea Cavallaro

Queen Mary University of London
Mile End Road, E1 4NS London, United Kingdom
Email: {tahir.nawaz,andrea.cavallaro}@eecs.qmul.ac.uk

ABSTRACT

The growing interest in developing video tracking algorithms has not been accompanied by the development of commonly used evaluation criteria to assess and to compare their performance. Researchers often present trackers' results on different datasets and evaluate them with different performance measures thus hindering both formative and summative quality assessment. In this paper, we present a protocol to evaluate the performance of tracking algorithms that tests video trackers using a set of trials and a pre-defined set of sequences and that enables objective and reproducible performance evaluation of trackers using ground truth information. Each trial highlights strengths and weaknesses of a tracker on simulated test scenarios on real sequences that represent real-world scenarios. Moreover a new evaluation measure is introduced that allows us to summarize the performance of a tracker based on the lost-track-ratio curve. The validation and the effectiveness of the proposed protocol is demonstrated experimentally on three trackers and its implementation is made available online to the research community.

***Index Terms*—** Video tracking, performance evaluation, protocol, trial, perturbation

1. INTRODUCTION

Video trackers are important temporal filters used in many applications ranging from human-computer interfaces to security and from behavior understanding to event detection. Despite their importance and their increasing diffusion, there is still a lack of a commonly accepted evaluation procedure that would allow effective evaluation and comparison of tracking algorithms.

Although several benchmark datasets and evaluation measures already exist, the analysis of the strengths and weaknesses of a specific tracker based only on the results reported in papers is still in most cases very difficult because of the lack of a commonly used evaluation protocol. Examples include the framework introduced by PETS (Performance Evaluation of Tracking and Surveillance), ETISEO (Evaluation du Traitement et de l'Interpretation de Sequences vidEO), CAVIAR (Context Aware Vision using Image-based Active Recognition), CLEAR (Classification of Events, Activities and Relationships). Other smaller-scale evaluation frameworks include comprehensive proposals such as the one in [1], and simple approaches such as the one based on "pseudo-synthetic video" sequences [2], on frame-based and object-based metrics [3], on the *Label and Size Based Evaluation Measure (LSBEM)* [4], or on measuring the tracking difficulty using a reflective model [5]. None of these frameworks has yet been widely taken up by the research community, thus limiting the opportunity to effectively assess and easily compare video tracking results.

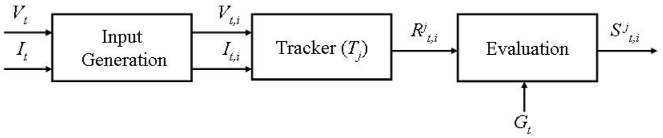


Fig. 1. Schematic diagram of the proposed evaluation Protocol for Tracking (PFT). V_t and I_t are the input video sequence and the corresponding target initialization data, respectively. $V_{t,i}$ and $I_{t,i}$ are the input to the tracker after appropriately modifying V_t and I_t for the tests. $R^j_{t,i}$ is the trajectory estimated by tracker T_j on trial i, $S^j_{t,i}$ is its final evaluation score computed with reference to the ground truth G_t.

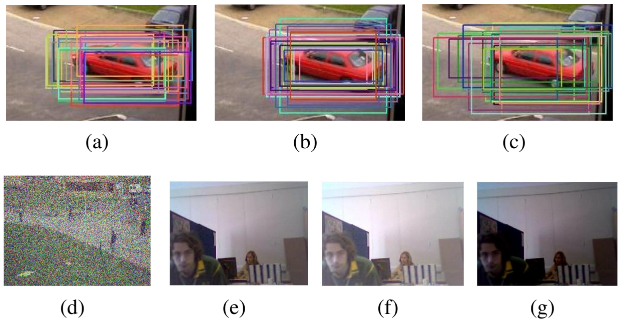
The contribution of this paper is threefold: (i) we propose a new comprehensive performance evaluation measure, (ii) we define an evaluation protocol and (iii) we make available its corresponding software implementation to facilitate its use by the research community. The protocol consists of a set of sequences and a set of evaluation procedures, or *trials*, which simulate real-world conditions such as changes in illumination, frame dropping, noise and various initialization errors. Each trial uses a predefined input-type generated by synthetically modifying either the test sequence or the initialization (bounding box) of the tracker. We also introduce an evaluation measure based on the area under the *lost-track-ratio* curve (AUC_λ), which effectively summarizes the performance of a tracker. The source code of the evaluation protocol is made available online at `http://www.eecs.qmul.ac.uk/~andrea/pft.html`.

This paper is organized as follows. Section 2 presents the proposed evaluation protocol and describes the trials, the evaluation criteria and the dataset. This is followed by the experimental validation in Section 3. Section 4 concludes the paper.

2. THE EVALUATION PROTOCOL

Seven trials have been defined to evaluate the performance of video trackers. Each trial represents a scenario that trackers are likely to face in real-world applications. These trials cover scenarios such as errors in the initialization of the tracker, variations in illumination, delayed generation of the tracking results and noisy input data. Figure 1 shows the schematic of the proposed evaluation Protocol for Tracking (PFT). Let a tracker T_j be evaluated on trial i of the protocol, with $j = 1, 2, ..., M$ and $i = 0, 1, ..., Z$, where M is the number of trackers under test and $Z + 1$ is the number of trials in the protocol ($Z = 6$), as discussed in the following subsection.

This work was supported in part by the EU, under the FP7 project APIDIS (ICT-216023).



(a) (b) (c)

(d) (e) (f) (g)

Fig. 2. (a-c) Example of perturbed initializations for Trials 1, 2 and 3, respectively; (d) Example of a test frame after adding the zero-mean Gaussian noise (Trial 4) with $\sigma^2 = 2(\sigma_r^2, \sigma_g^2, \sigma_b^2)$; (e-g) Example of frames in Trial 6 after changing illumination: (e) original frame, (f) the frame after increasing the illumination, (g) the frame after decreasing the illumination.

The first trial, *Trial* 0, is the reference trial that evaluates trackers using the original initialization, I_t, and test sequence, V_t,without any modifications. Then, the trial i initialization, $I_{t,i}$, and test sequence, $V_{t,i}$, are generated for the subsequent trials by adding a predefined perturbation to I_t, or to V_t, respectively.

Let a tracker T_j be tested on $V_{t,i}$ and $I_{t,i}$ to obtain the tracking result $R^j_{t,i}$, the estimated trajectory of the target on trial i. $R^j_{t,i}$ is evaluated with respect to the ground truth trajectory G_t of the target to compute performance scores $S^j_{t,i}$ for tracker j on trial i. The values of $S^j_{t,i}$ are finally used to compare the trackers under analysis.

2.1. Trials

Trials 1, 2 and 3 evaluate the robustness of trackers to *initialization errors* (e.g. possible errors of a detector). The tests are based on adding perturbations to the initialization data, namely the position and/or the size of the bounding box. *Trial* 1 adds perturbations to the position of the initializing bounding box only, while keeping its width and height unchanged (Fig. 2(a)); *Trial* 2 adds perturbation to the size (width and height) only (Fig. 2(b)); whereas *Trial* 3 adds perturbations both to the position and the size of the initializing bounding box (Fig. 2(c)). The amount of perturbation introduced in the data is a function of the size of the bounding box (width along x-direction and height along y-direction) with the constraint of a minimum 50% overlap between the modified bounding box and the original initialization. In each trial, the trackers are evaluated on a set of 20 perturbed initializations to test their robustness to initialization perturbations with a particular test sequence.

Trials 4, 5 and 6 evaluate the robustness of trackers to typical challenges such as noisy input generated by low-cost sensors, variations in illumination conditions and frame dropping. *Trial* 4 evaluates the robustness of trackers to the presence of *noise* in the video sequence. Zero-mean Gaussian noise is added to the three color channels, with $\sigma_r = 8.59$, $\sigma_g = 8.40$ and $\sigma_b = 11.96$, as estimated from a low-quality webcamera (Creative webcam VF0330). Three test sequences are generated with twice, four times and six times the estimated variances $(\sigma_r^2, \sigma_g^2, \sigma_b^2)$ to ensure trackers' evaluation on difficult noisy scenarios. Fig. 2(d) shows an image from PETS2010 datasets after adding the Gaussian noise with $2(\sigma_r^2, \sigma_g^2, \sigma_b^2)$. *Trial* 5 evaluates the robustness of trackers under *frame skipping* to simulate a potential latency of the tracker in processing the input and generating results. Frame skipping can possibly result in abrupt shifts in the position of the target [6]. In this trial, video sequences are generated by regularly dropping a certain number of frames $(m - 1)$ from the video sequence where $m = 2, 4, 6, 8$. Finally, *Trial* 6 evaluates the robustness of trackers to *illumination changes*. Two test sequences are generated by either *increasing* $(+\Delta L)$ and *decreasing* $(-\Delta L)$ synthetically the illumination over time. The illumination is changed by adding or subtracting (with saturation) $\Delta L = 0, 1, 2, ..., 200$ to or from the intensities of pixels of frames $(k = 1, 2, 3, ..., 201)$, respectively. When the number of frames in the video sequence $K > 201$, the amount of illumination change in the remaining frames is kept constant to $\Delta L = 200$. When $K < 201$, the amount of illumination change is $\Delta L = 0, 1, 2, ..., (K-1)$ in frames $k = 1, 2, 3, ..., K$ respectively. Fig. 2(e-g) shows the visualization of frames (sequence taken from `www.spevi.org`) after both an increase and a decrease of the illumination.

In the experiments, the trials are run several times on probabilistic trackers and the mean of their results is evaluated for an accurate performance evaluation.

2.2. Evaluation criteria

The evaluation criteria measure the performance of the tracking results with reference to a ground truth. We use as starting evaluation criteria the overlap measure, O_k, at each frame k and the lost-track ratio, λ. These criteria will allow us to define a new comprehensive measure of performance, the *area under the lost-track ratio curve* (AUC_λ), as described below.

The *overlap measure*, O_k, quantifies the amount of overlap between the estimated and the ground-truth bounding boxes. O_k is computed at every frame where the target exists [7]:

$$O_k = \frac{|TP_k|}{|TP_k| + |FP_k| + |FN_k|}, \tag{1}$$

where $|TP_k|$ is the number of pixels that are correctly detected at frame k as belonging to the target, $|FP_k|$ is the number of pixels incorrectly detected and $|FN_k|$ is the number of pixels missed by the tracker. The larger O_k, the better the tracking result.

The *lost-track ratio*, λ, is computed based on O_k over a test sequence [7]. A track in a frame k is considered to be *lost* when the amount of overlap between the estimated track and the ground truth is smaller than a certain value, i.e. $O_k \leq \tau$, where $\tau \in [0, 1]$. λ is the ratio between the number of frames with a lost track, N_l, and the total number of frames N of the estimated target trajectory:

$$\lambda = \frac{N_l}{N}. \tag{2}$$

Because the appropriate value of τ is different for different tracking applications, we consider the variation of λ for a full range of τ values, from $\tau = 0$ to $\tau = 1$ with an increment of $\Delta\tau = 0.01$. We refer to these parameterized values of the lost-track ratio as $\lambda(\tau)$. Based on $\lambda(\tau)$, we finally introduce a compact measure, AUC_λ, that quantifies the performance of trackers by computing the area under the *lost track ratio* curve as

$$AUC_\lambda = \Delta\tau \sum_{\tau=0}^{1} \lambda(\tau), \tag{3}$$

with $0 \leq AUC_\lambda \leq 1$. The lower the area under the *lost track ratio* curve (AUC_λ), the better the results (Fig. 3). The ideal performance of a tracker corresponds to $AUC_\lambda = 0$.

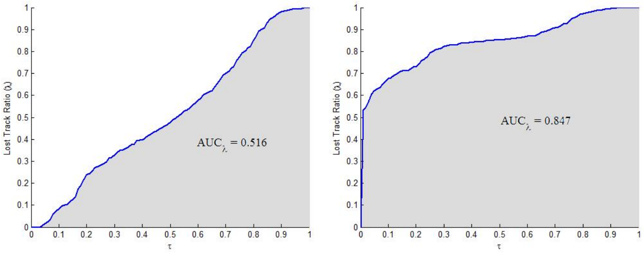


Fig. 3. Example of performance comparison of two tracking results in terms of AUC_λ. The tracking result on the left ($AUC_\lambda = 0.516$) is preferable to that on the right ($AUC_\lambda = 0.847$) because of a smaller AUC_λ.

2.3. Dataset

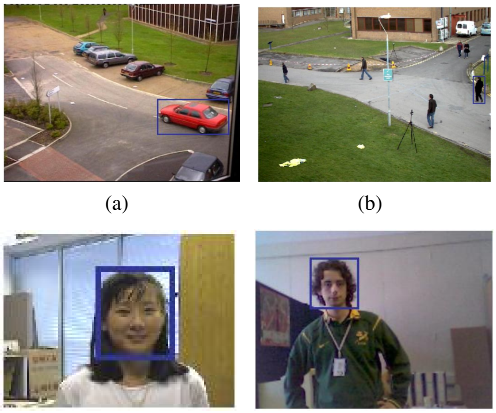
The selection of the dataset for PFT is made considering the diversity of targets (rigid and articulated) and the widespread availability of and access to the test sequences. The selected targets cover a range of tracking challenges such as partial and total occlusions, 360 degrees turnings, tilting, scale changes and random-path movements. The targets are shown in Figure 4: a *vehicle* ($H1$) from PETS2000[1], a *person walking* ($H2$) from PETS2010[2], a *head* ($H3$) from Clemson[3], and a *head* ($H4$) from the SPEVI[4] dataset. $H1$ is a rigid target (initial size: 227 × 108) whose minimum and maximum size in the sequence are 2067 and 24516 (pixels), respectively. The sequence containing $H1$ has 160 frames (frame size: 576×768). $H2$ is an articulated target having an initial size of 30 × 87 pixels. Its minimum and maximum size are 180 and 3444 (pixels), respectively. The sequence containing $H2$ is of 150 frames (frame size: 576×768). The initial size of $H3$ is 39 × 46 pixels and its minimum and maximum size are 192 and 2646 (pixels) respectively. The corresponding sequence has 501 frames (frame size: 96 × 128) and is characterized by pan, tilt and zoom movements of the camera. The initial size of $H4$ is 62 × 66 pixels and its minimum and maximum size are 370 and 40128 pixels, respectively. The corresponding sequence has 550 frames (frame size: 240 × 320).

Each tracker is evaluated on a total of 52 (13×4) sequences generated in the trials by modifying the dataset. In the following section, we shall highlight the strengths and weaknesses of three trackers by evaluating their performance using the proposed protocol.

3. EXPERIMENTAL VALIDATION

The proposed protocol is validated by evaluating and comparing the color-based adaptive Particle Filter tracker (PF) [8], the Mean Shift tracker (MS) [9] and the Hybrid tracker [10]. The results are summarized in Table 1 and Figure 5.

The performance evaluation results of the three trackers show that PF and the Hybrid tracker are more robust to cope with occlusions compared to MS, as shown for example in trial 0 with $H2$ and $H3$ (both involves occlusions) in Table 1. The evaluation with $H1$ and $H2$ shows that MS and Hybrid are more robust to perturbations in the initialization than PF, as reflected in the variations of their AUC_λ scores (Table 1) on trials 1-3 (note that in a few cases, the variations are comparable for all trackers). With $H3$, MS is found to be more sensitive to perturbations in initialization than PF and Hybrid. The evaluation results of PF with $H4$ show that the variations of its AUC_λ scores are smaller than those of MS and Hybrid; however its mean AUC_λ scores are considerably higher. AUC_λ scores for three trackers on trial $1 - 3$ are plotted in Fig. 5.



(a) (b)

(c) (d)

Fig. 4. Visualization of targets used in PFT: (a) Vehicle (PETS 2000); (b) Person (PETS 2010); (c) Head (Clemson); (d) Head (SPEVI).

The overall evaluation results show that PF and Hybrid are more robust to noise than MS. This is reflected in the variations of their AUC_λ scores computed for the three levels of Gaussian noise (twice, four times and six times $(\sigma_r^2, \sigma_g^2, \sigma_b^2))$) on trial 4 with all the targets (Table 1). The values of variations are lower for PF and Hybrid as they are more robust to noise than MS.

The Hybrid tracker is robust to fast movements of a target as shown in the variations of its AUC_λ scores on trial 5, which are generally smaller than those of MS and PF (Table 1). PF has a better performance on $H3$ and $H4$, whereas MS has a better performance on $H1$ and $H2$, as shown in the variations of the AUC_λ scores on trial 5 (Table 1). Interestingly, the variation of the AUC_λ scores of MS on $H2$ is even lower than that of the Hybrid tracker.

Finally, the evaluation results show that MS is sensitive to illumination changes compared to PF and Hybrid, as reflected in its performance on trial 6. Although the variations of its AUC_λ scores are mostly very small, its mean AUC_λ scores on trial 6 are generally very large (Table 1) showing that its performance declined significantly both for the case of increasing and decreasing illumination. PF and Hybrid have better performance. Figure 6 shows the mean AUC_λ scores of the trackers in each trial and the average AUC_λ scores of the trackers across all the trials: the Hybrid tracker has the best performance, followed by PF and then MS.

In summary, based on the outcomes of the protocol, the observations that were offered in [8, 9, 10] have been confirmed numerically, thus providing an objective ground for comparison among trackers and their characteristics. PF is found to be robust to occlusions. MS and Hybrid are found to be more robust to cope with initialization perturbations on $H1$ and $H2$; however, the performance of trackers changed on $H3$, which contains a larger target and camera motion. PF and Hybrid showed better performance than MS in the presence of noise in the sequence, as expected by methods using particle filtering verus a deterministic gradient descent approach. Hybrid is better at dealing with fast movements compared to the other two trackers. Finally, PF and Hybrid are better at coping with

[1] `ftp://ftp.cs.rdg.ac.uk/pub/PETS2000/`

[2] `http://www.cvg.rdg.ac.uk/PETS2010/a.html#s2l1`

[3] `http://www.ces.clemson.edu/~stb/`

[4] `http://www.eecs.qmul.ac.uk/~andrea/spevi.html`

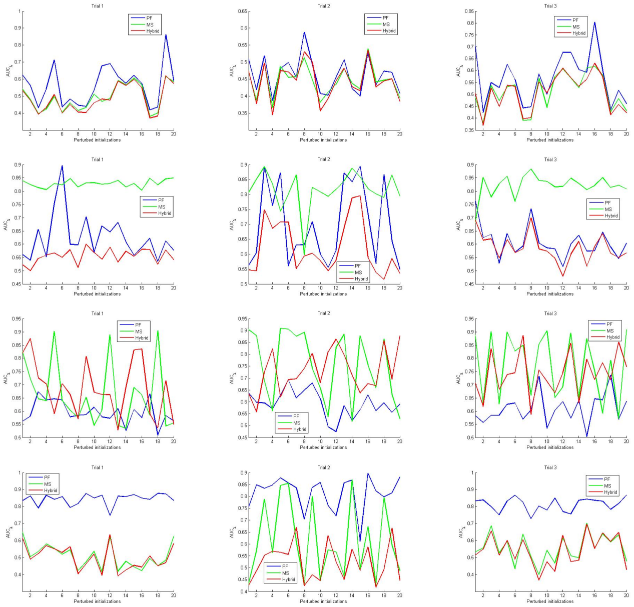


Fig. 5. AUC_λ scores of PF (blue), MS (green) and the Hybrid tracker (red) for 20 perturbed initializations on each trial. First column: Trial 1; Second column: Trial 2; Third column: Trial 3. First row: $H1$; Second row: $H2$; Third row: $H3$; Fourth row: $H4$.

Table 1. Mean (μ) and standard deviation (σ) of the AUC_λ scores for all the trials.

Trial	Target	PF	MS	Hybrid
		μ (σ)	μ (σ)	μ (σ)
0	$H1$	0.418	0.380	0.370
	$H2$	0.668	0.823	0.543
	$H3$	0.516	0.847	0.631
	$H4$	0.865	0.436	0.390
1	$H1$	0.563 (0.1170)	0.491 (0.0775)	0.483 (0.0794)
	$H2$	0.626 (0.0861)	0.829 (0.0137)	0.555 (0.0272)
	$H3$	0.595 (0.0430)	0.663 (0.1235)	0.686 (0.1061)
	$H4$	0.844 (0.0338)	0.509 (0.0718)	0.500 (0.0723)
2	$H1$	0.463 (0.0532)	0.446 (0.0449)	0.444 (0.0540)
	$H2$	0.698 (0.1311)	0.815 (0.0630)	0.621 (0.0898)
	$H3$	0.593 (0.0555)	0.755 (0.1447)	0.728 (0.0898)
	$H4$	0.814 (0.0716)	0.614 (0.1577)	0.522 (0.0800)
3	$H1$	0.567 (0.1015)	0.508 (0.0781)	0.506 (0.0759)
	$H2$	0.607 (0.0605)	0.820 (0.0430)	0.586 (0.0530)
	$H3$	0.609 (0.0576)	0.770 (0.1232)	0.741 (0.0816)
	$H4$	0.813 (0.0395)	0.555 (0.0818)	0.542 (0.0902)
4	$H1$	0.965 (0.0014)	0.946 (0.0088)	0.937 (0.0053)
	$H2$	0.848 (0.1210)	0.814 (0.1699)	0.768 (0.1082)
	$H3$	0.876 (0.0060)	0.877 (0.0102)	0.869 (0.0048)
	$H4$	0.899 (0.0058)	0.862 (0.0174)	0.881 (0.0086)
5	$H1$	0.530 (0.1936)	0.400 (0.0358)	0.362 (0.0080)
	$H2$	0.628 (0.2274)	0.896 (0.0692)	0.644 (0.1927)
	$H3$	0.594 (0.1508)	0.764 (0.1728)	0.663 (0.1562)
	$H4$	0.705 (0.1336)	0.698 (0.2491)	0.410 (0.0329)
6	$H1$	0.535 (0.1572)	0.501 (0.1787)	0.472 (0.1637)
	$H2$	0.699 (0.2613)	0.862 (0.0594)	0.688 (0.2467)
	$H3$	0.835 (0.1470)	0.931 (0.0145)	0.924 (0.0375)
	$H4$	0.864 (0.0044)	0.940 (0.0017)	0.938 (0.0006)

illumination changes.

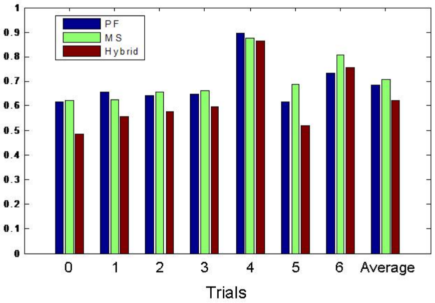


Fig. 6. Mean AUC_λ scores for each tracker in each trial (0 - 6) and overall mean AUC_λ scores (Average) for each tracker across all the trials.

4. CONCLUSIONS

We presented an evaluation protocol and a new compact performance measure to quantify the performance of video trackers and enable the comparison of their robustness in different real-world scenarios. The variability of these scenarios is controlled by incorporating perturbations to initializations and variations in illumination, noise and frame rate of the test sequences. The proposed protocol was validated on three well-known trackers and their evaluation results confirm the analysis of previous studies thus validating the effectiveness of the protocol. Our current work focuses on the performance evaluation and comparison of a larger set of trackers using the proposed evaluation protocol.

5. REFERENCES

[1] 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," *IEEE Trans. on PAMI*, vol. 31, no. 2, pp. 319–336, February 2009.

[2] J. Black, T. Ellis, and P. Rosin, "A novel method for video tracking performance evaluation," in *Proc. of IEEE Int. Workshop on VS-PETS*, 2003, pp. 125–132.

[3] F. Bashir and F. Porikli, "Performance evaluation of object detection and tracking systems," in *Proc. of IEEE Int. Workshop on PETS*, 2006, pp. 7–14.

[4] J. Popoola and A. Amer, "Performance evaluation for tracking algorithms using object labels," in *Proc. of IEEE ICASSP*, April 2008.

[5] P. Pan, F. Porikli, and D. Schonfeld, "A new method for tracking performance evaluation based on a reflective model and perturbation analysis," in *Proc. of IEEE ICASSP*, April 2009.

[6] G. Hua and Y. Wu, "Multi-scale visual tracking by sequential belief propagation," in *Proc. of IEEE Conf. on CVPR*, 2004.

[7] E. Maggio and A. Cavallaro, *Video tracking: theory and practice*, Wiley, 2011.

[8] K. Nummiaro, E. Koller-Meier, and L. Van Gool, "An adaptive color-based particle filter," *Image and Vision Computing*, vol. 21, no. 1, pp. 99–110, 2002.

[9] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. on PAMI*, vol. 25, no. 5, pp. 564–577, May 2003.

[10] E. Maggio and A. Cavallaro, "Hybrid particle filter and mean shift tracker with adaptive transition model," in *Proc. of IEEE ICASSP*, March 2005, pp. 221–224.