Color Invariant SURF in Discriminative Object Tracking

Dung Manh 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.Smeulders}@uva.nl

Abstract. Tracking can be seen as an online learning problem, where the focus is on discriminating object from background. From this point of view, features play a key role as the tracking accuracy depends on how well the feature distinguish object and background. Current discriminative trackers use traditional features such as intensity, RGB and full body shape features. In this paper, we propose to use color invariant SURF features in the discriminative tracking. This set of invariant features has been shown to be of increased invariance and discriminative power. The resulting tracker inherits a good discrimination between object and background while keeping advantages of the discriminative tracking framework. Experiments on a dataset of 80 videos covering a wide range of tracking circumstances show that the tracker is robust to changes in object appearance, lighting conditions and able to track objects under cluttered scenes and partial occlusion.

Key words: tracking, surf, color, invariant

1 Introduction

In many visual object trackers [1–4], traditional features such as intensity, RGB and full body shape features are used. They reflect the state of the image directly and they are fast to compute. However, to cope with varying aspects of the object and the scene, features should be invariant to the undesired variations in the appearance of the object such as shadows, shadings and occlusions and discriminative enough to distinguish object from other objects and background. These above features are of limited invariance to such changes. The SIFT/SURF [5, 6] show increase in discriminative power [7, 8]. In particular Van de Sande et al. [9] show that the set of color and invariant SIFT obtains the best performance in the object recognition task. Moreover, the computations of SIFT and SURF are recently made fast enough for real-time application [10]. Inspired by these results, in this paper we aim to investigate invariant features in visual object tracking.

At large, trackers can be divided by three main mechanisms: background models [11, 12], foreground-based trackers [3, 4] and discriminative (foreground-background) trackers [2, 13, 14]. Many background-based trackers and foreground-based trackers resort to assumptions that an aspect of the background or the

foreground is constant (or at least predictable for the next image). They are designed to work well when disturbing scene-related circumstances develop slowly over time and place. Under that condition, the model of the background or the model of the foreground can be adapted. However, the assumption of slow development of the lighting and scene conditions is frequently violated in reality when there are abrupt changes in object appearance due to entering into shadow, abrupt albedo changes due to rotation, abrupt object motion changes, or abrupt silhouette changes due to occlusion. In many of such situations, discriminative trackers are in favor over the other two as they put in the center the distinction between foreground and background rather than modeling the foreground alone or background alone. Concentrating on discriminative trackers, invariant discriminative features are the natural ingredient to incorporate.

This paper proposes a novel tracking method using foreground/background discrimination. Unlike the above-mentioned methods, the proposed tracker uses color invariant SURF features for discrimination. The aim is to be robust to changes in object appearance and lighting conditions. And, the aim is to track objects under cluttered scenes and partial occlusions. An innovation of the research is the use of a broad dataset [15] developed to test the robustness of all sorts of tracking conditions as they occur in reality.

2 Related work

Our work is based on two components: discriminative tracking and color invariant features. We hence review these two topics in this section.

2.1 Discriminative Trackers

The discriminative trackers in [16, 17] are focused on classifier selection. A set of weak classifiers is trained on object features and background features. Grabner et al. [16] use online boosting to establish a strong classifier. Avidan [17] combines the weak classifiers into a decision by AdaBoost. Although online boosting and AdaBoost help to select best results from the weak classifiers, they disregard the spatial relation between object features. They suffer from a large number of free parameters to estimate, making the tracking computationally expensive and unstable under varying conditions.

The discriminative trackers in [18, 19] are focused on feature selection. Grabner et al. [18] propose a semi-supervised online learning method to select features. Mahadevan and Vasconcelos [19] define saliency measure for features, which ranks features how well they discriminate. Since the features are not invariant with respect to varying tracking conditions, feature selection methods will select best features on the fly. This method however leaves many degrees of freedom.

In [2], linear discriminant analysis is applied to discriminative tracking. An analytical incremental solution is found for updating the classifier online. It enables fast updating scheme with a small number of free parameters. The tracker also retains a spatial relation between object features. This allows the tracker to

overcome partial occlusions and compensate for global changes of illumination. Due to its computational simplicity and the small number of free parameters we follow this discrimination technique in our tracker.

2.2 Features in object trackers

Many trackers successfully replace grey features by color features (see an overview in [20]) and by SIFT/SURF features. He et al. [21] propose a SURF-based tracker where SURF-features are extracted from the object and its surrounding area using interest points. Object feature correspondence is estimated and then used to predict the object motion. Background features are only used to detect occlusions. The tracker imposes a smooth transition of the object appearance. Zhou et al. [22] apply original SIFT features into the mean shift tracking framework. Due to the discriminative power of SIFT, the resulting tracker outperforms the original version at the expense of considerably more computation. Tran and Davis [23] use SIFT in blob tracking, where objects are represented by a set of MSER regions. Object motion is estimated from the estimations of the blobs' motions. The tracker can track objects undergoing illumination changes due to the use of SIFT feature. These results show the potential of using SIFT/SURF in tracking.

The trackers in [24–26] successfully apply color features into the discriminative tracking framework. The tracker in Collins et al. [26] works on a pool of 49 linear combinations of R, G, B. For each feature, the log likelihood ratio between foreground and background feature histograms is computed, which is then used to rank the features. Similar mechanisms can also be found in [24] with multiple color spaces and color distribution models, or in [25] with 7 types of color histograms and gradient orientation histogram. These trackers demonstrate the usefulness of color features in discriminative tracking.

Our tracker is different from the above trackers. We use a different set of features in discriminative tracking. The features are the combinations of SURF with different color spaces and color invariants. These features are of enhanced discriminative and invariance power.

3 The Proposed Tracker

3.1 Discriminative Tracking Framework

Discriminative tracking treats tracking as a two-class instant classification problem between the object class and the background class. The object features are densely sampled in the object region and denoted by $\boldsymbol{f}_1^o, \ldots, \boldsymbol{f}_n^o$. The background features are also densely sampled in the neighbor background region and denoted by $\boldsymbol{f}_1^b, \ldots, \boldsymbol{f}_m^b$. As we aim to discriminate the object from background, with each object feature $\boldsymbol{f}_i^o$, a classifier g_i is trained to distinguish it from all the background features. The set of classifiers $\{g_1, \ldots, g_n\}$ constitutes the discrimination between the object and background. g_i should be fast to train in the

incremental mode and have few free parameters to arrive at a robust solution on few samples. To this end, we follow [2] with the use a linear classifier:

$$g_i(\boldsymbol{x}) = \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle + b_i, \tag{1}$$

where $\boldsymbol{a}_i \in \mathcal{R}^N$, $b_i \in \mathcal{R}$ and $\langle , \rangle$ denotes the inner product; N is the dimension of the used feature. The classifier g_i is trained such that

$$g_i(\boldsymbol{f}_i^o) > 0 \text{ and } g_i(\boldsymbol{f}_j^b) < 0 \text{ for all } \boldsymbol{f}_j^b. \tag{2}$$

When a new frame comes in, denote by θ the spatial transformation between the two frames and by $I(\boldsymbol{f}_i^o, \theta)$ the feature in the new frame that correspond to feature $\boldsymbol{f}_i^o$. The search for the object in the new frame is cast into the following maximization problem:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} g_i(I(\boldsymbol{f}_i^o, \theta)). \tag{3}$$

The maximization effectively pushes the object candidate as far away from the known background features as possible and pulls it close to the known object features. We notice that as $g_i(I(\boldsymbol{f}_i^o, \theta)) = \langle \boldsymbol{a_i}, I(\boldsymbol{f}_i^o, \theta) \rangle + b_i$ and b_i is independent from θ, we only need to compute $\boldsymbol{a}_i$.

Learning and updating the classifiers: given the object features $\boldsymbol{f}_1^o, \ldots, \boldsymbol{f}_n^o$ and the background features $\boldsymbol{f}_1^b, \ldots, \boldsymbol{f}_m^b$, we learn the classifiers g_i by solving the following optimization problem:

$$\min_{\boldsymbol{a}_i, b_i} \left[\left(\langle \boldsymbol{a}_i, \boldsymbol{f}_i^o \rangle + b_i - 1 \right)^2 + \sum_{j=1}^{m} \alpha_j \left(\langle \boldsymbol{a}_i, \boldsymbol{f}_j^b \rangle + b_i + 1 \right)^2 + \frac{\lambda}{2} ||\boldsymbol{a}_i||^2 \right], \tag{4}$$

where α_j are the weighting coefficients of the background features, $\sum_{j=1}^{m} \alpha_j = 1$. The closed-form solution of (4) is given by ([2]):

$$\boldsymbol{a}_i = c_i \left(\lambda \boldsymbol{I} + \boldsymbol{B} \right)^{-1} \left(\boldsymbol{f}_i^o - \bar{\boldsymbol{f}}^b \right), \tag{5}$$

where $\boldsymbol{B}$ and $\bar{\boldsymbol{f}}^b$ are the weighted covariance and mean of the background features; $\boldsymbol{I}$ is the identity matrix:

$$\bar{\boldsymbol{f}}^b = \sum_{j=1}^{m} \alpha_j \boldsymbol{f}_j^b, \tag{6}$$

$$\boldsymbol{B} = \sum_{j=1}^{m} \alpha_j \left(\boldsymbol{f}_j^b - \bar{\boldsymbol{f}}^b \right) \left(\boldsymbol{f}_j^b - \bar{\boldsymbol{f}}^b \right)^T, \tag{7}$$

$$c_i = \frac{1}{1 + 0.5 \left(\boldsymbol{f}_i^o - \bar{\boldsymbol{f}}^b \right)^T \left(\lambda \boldsymbol{I} + \boldsymbol{B} \right)^{-1} \left(\boldsymbol{f}_i^o - \bar{\boldsymbol{f}}^b \right)}. \tag{8}$$

Equations (6), (7) and (8) allow a fast learning step for the classifiers. We notice that the background features are compactly represented by the weighted mean

and the weighted covariance. It is hence not necessary to keep all the background features.

After each tracking step, we extract new object and background features. Suppose that $\hat{\theta}$ is the spatial transformation found by solving the optimization problem in Equation (3). Then $I(\boldsymbol{f}_1^o, \hat{\theta}), \ldots, I(\boldsymbol{f}_n^o, \hat{\theta})$ are the new object features. In order to allow the tracker to remember the past appearance of the object, we allow the old features to stay in the object representation with decreasing weights:

$$\boldsymbol{f}_i^{o(new)} = (1-\gamma)\boldsymbol{f}_i^o + \gamma I(\boldsymbol{f}_i^o, \hat{\theta}), \tag{9}$$

where γ is a predefined decay coefficient.

Suppose that $\boldsymbol{f}_{m+1}^b, \ldots, \boldsymbol{f}_{m+k}^b$ are the new background features. We put total weight for the new background to be γ, while the weight of each old background feature is downscaled $(1-\gamma)$. The updated background mean and covariance are given by:

$$\bar{\boldsymbol{f}}^{b(new)} = (1-\gamma)\bar{\boldsymbol{f}}^b + \gamma\frac{1}{k}\sum_{j=m+1}^{m+k}\boldsymbol{f}_j^b, \tag{10}$$

$$\boldsymbol{B}^{(new)} = (1-\gamma)\boldsymbol{B} + (1-\gamma)\bar{\boldsymbol{f}}^b\bar{\boldsymbol{f}}^{bT} - \bar{\boldsymbol{f}}^{b(new)}\bar{\boldsymbol{f}}^{b(new)T} + \frac{\gamma}{k}\sum_{j=m+1}^{m+k}\boldsymbol{f}_j^b\boldsymbol{f}_j^{bT}. \tag{11}$$

The set of Equations (5), (8), (9), (10) and (11) allows the tracker to update the classifiers in the incremental mode efficiently.

3.2 Features

The use of SURF in visual tracking is rather limited in few foreground-based trackers [21–23]. One of the reasons is due to the expensive procedure to compute SURF descriptors at interest points. We overcome this problem by extracting features $\{\boldsymbol{f}_1^o, \ldots, \boldsymbol{f}_n^o; \boldsymbol{f}_1^b, \ldots, \boldsymbol{f}_m^b\}$ densely and using the fast algorithm to compute SURF descriptors recently proposed in [10].

The original intensity-based SURF features have been extended to different color spaces and color invariant spaces. They have not yet been explored in visual tracking. Among the color spaces, we choose the opponent space as the high decorrelation between the 3 channels. Opponent color space contains one intensity channel and two chromaticity channels. As the three channels are highly decorrelated they are likely to improve the discriminative power when used together:

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix}. \tag{12}$$

With the color invariants, Geusebroek et al. [27] show an inclusion relationship: $H \subset C \subset W$, where H, C and W are three invariants derived from the

Kubelka-Munk photometric model under different assumptions. The inclusion implies that H has highest invariance and essentially H flattens out all patterns in an image. This is not a desired property for tracking since we want to keep a certain level of discriminative power to distinguish the object from the scene and from other objects. On the other hand, W lacks invariance. It does not wipe out accidental changes from illumination. We did experiments with the 3 invariants separately and observed consistently degraded performance of the H and W versions over the C version (the differences are approximately 58% and 8% respectively. Further data is not shown here). We hence will focus on C-SURF. We also use the intensity SURF (I-SURF) as baseline.

The C invariant [27] is an object reflectance property independent of the viewpoint, surface orientation, illumination direction and illumination intensity. The C color space consists of one intensity channel and 2 channels $\{C_\lambda, C_{\lambda\lambda}\}$ computed as follows:

$$\begin{array}{c} C_\lambda = \frac{E_\lambda}{E} \\ C_{\lambda\lambda} = \frac{E_{\lambda\lambda}}{E}, \end{array} \tag{13}$$

where $E(\lambda)$ is the energy distribution of the incident light over wavelength λ. $E, E_\lambda, E_{\lambda\lambda}$ are estimated from an RGB image as follows:

$$\begin{pmatrix} E \\ E_\lambda \\ E_{\lambda\lambda} \end{pmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.3 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \tag{14}$$

4 Dataset and Evaluation Metric

As we aim to design a tracker robust to the wide variety of tracking circumstances, we use the dataset in [15] covering 12 most important tracking conditions: lighting condition, object albedo, object specularity, object transparency, object shape, motion smoothness of object, motion coherence of object, clutter, confusion, occlusion, moving camera and zooming camera (the reference gives more detail on the selection and creation of the dataset). This dataset enables evaluation of a tracker with respect to different tracking circumstances. The dataset contains 80 videos covering both realistic videos and in-lab videos. The distribution of the videos over the categories are uniform. The videos are manually annotated in every $5th$ frame. Some example videos from the dataset are depicted in Figures 3, 4 and 5.

To measure the trackers' performance, [15] proposes to use a category-level average tracking accuracy measure (CATA), which indicates how much a tracker covers the object in each frame in average. CATA ranges from 0 to 1. The higher CATA is, the more accurate the tracker is. A CATA value of 0.6, for example, implies that in average in each frame where the object is present, the tracker covers at least 60% of the object and at least 60% of the tracked box is covered by the object.

5 Results

We demonstrate the performance of the proposed tracker in this section. For comparison purpose, three other state-of-the-art trackers are considered: the foreground background tracker (FBT) in [2]; the incremental visual tracker (IVT) in [3] and the Kalman predictive tracker (KAT) in [4]. We reimplemented the FBT and KAT, while the IVT is publicly available online from the author website.

5.1 Quantitative Comparison between Features

This section shows comparison of the proposed discriminative tracking framework with different types of features. In [2], the intensity Gabor feature is used. We extended it to include rudimental color information, resulting in the RGB Gabor feature. We compute CATA values of the discriminative tracking framework for 5 different types of features: intensity Gabor, RGB Gabor, I-SURF, C-SURF and Opponent-SURF. The data is visualized in Figure 1. As can be seen from the figure, the SURF-based versions outperform the Gabor-based versions in 11 out of 12 cases. This is attributed to the high discriminativeness of the SURF-based features, which especially is suited for our discriminative tracking framework. Large differences between the SURF-based versions and the Gabor-based versions can be seen in the following categories: albedo, transparency, clutter, confusion and occlusion.

Among the SURF-based versions, Opponent-SURF and C-SURF show better performance than I-SURF. This is attributed to the high decorrelation between three channels in the opponent color space, which contains one intensity channel decorrelated from the two chromaticity channels. The discriminative power of C-SURF regardless accidental shadows and shadings makes it well suited in combination with the online classifier which is at the core of this tracker. C-SURF improves the classification accuracy in our tracker especially in the confusion and occlusion cases where the object shares similar patterns with other neighbor objects or the object loses part of its appearance in occlusion.

Table 1. The average performance of the discriminative tracking framework with the 5 features in the whole dataset. This is computed by averaging all the CATA values of the 12 categories.

	Intensity Gabor	RGB Gabor	I-SURF	C-SURF	Opponent-SURF
Average	0.43	0.51	0.58	0.61	0.60

To conclude, SURF-based features outperform Gabor-based features. Further, color-based features outperform intensity-based features. As can be seen in Table 1, I-SURF gains improvement of 0.15 (35%), while Opponent-SURF and C-SURF gain even 0.17 (40%) and 0.18 (42%) respectively with respect to the original tracker in [2].

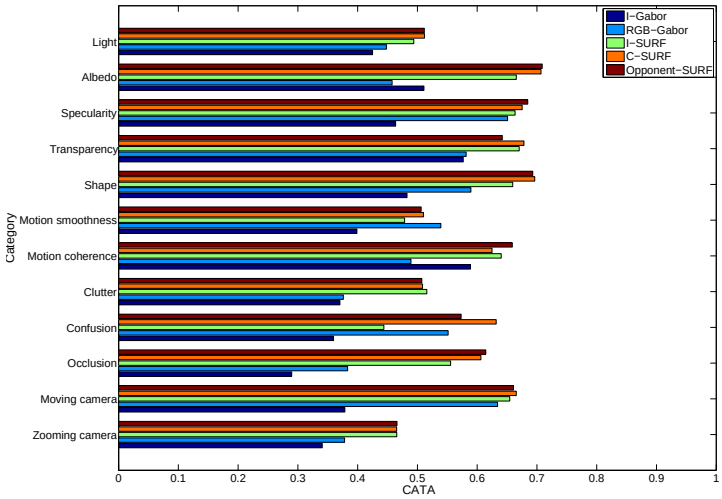

Fig. 1. Performance of the discriminative tracking framework with different types of features and different tracking circumstances. The x-axis indicates the CATA measure. The y-axis contains 12 different tracking categories in the dataset with total 80 videos.

5.2 Quantitative Comparison to Other Trackers

This section shows a quantitative comparison of the proposed tracker with the KAT and the IVT. We have integrated RGB, SURF, C-SURF and Opponent-SURF into the KAT and the IVT. However the SURF-based features do not improve the two trackers. The reason is that the numbers of free parameters in the IVT and the KAT are proportional to the feature's dimension. The use of the SURF-based features hence increases the number of free parameters to be estimated in the IVT and KAT with a limited number of samples. Hence the SURF-based features downgrade their performances. With IVT, we observe the best performance with the RGB feature, while the intensity feature is the best with KAT. The results of the proposed tracker with C-SURF, KAT with intensity and IVT with RGB are shown in Figure 2. As can be seen from the figure, the proposed tracker is more robust to changes in illumination conditions, object albedo and transparency. This is explained by the invariance of SURF to light intensity change and light intensity shift, which aids the tracker to overcome a certain level of illumination changes. The KAT gets affected most in the transparency case. The reason is that in such a case the object appearance reflects the color of the local background behind the object. Because of the inhomogeneous background, the object appearance changes abruptly, which violates the smooth assumption the KAT imposes on the object features.

Figure 2 also shows that the proposed tracker is more robust to confusion with the CATA value 63% while the scores for the IVT and the KAT are about 40%. The discriminative and invariance power of C-SURF enables the proposed tracker to distinguish the object from other nearby objects of similar appear-

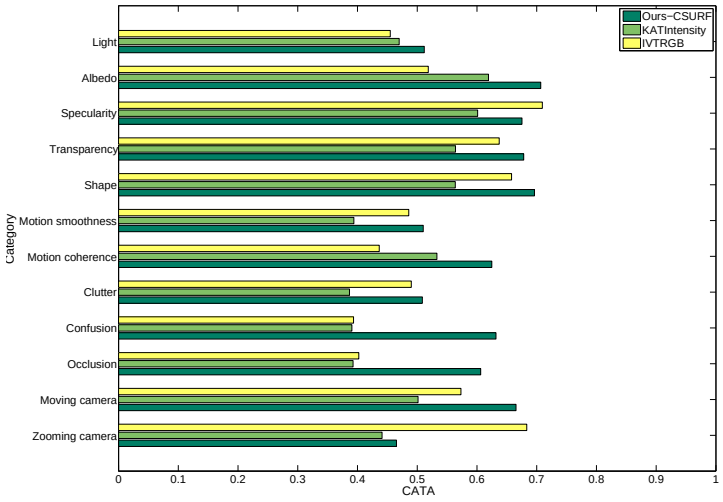

Fig. 2. Quantitative comparison between the proposed tracker with the IVT and the KAT. We select the best features: C-SURF for the proposed tracker, intensity for KAT and RGB for IVT. The x-axis indicates the CATA measure. The y-axis contains 12 different tracking categories in the dataset.

ance. We notice that confusion downgrades the IVT and the KAT as the two trackers have no mechanism to isolate the object even though they keep good representations of the object. The proposed tracker also outperforms the IVT and the KAT in the occlusion category. We notice that the IVT is the best in the zooming camera case. This is attributed to the scaling handling mechanism enabling the tracker to cope with objects with changing size due to camera's zooming. Overall, as can be seen from Table 2, the proposed tracker gains improvement of 0.12 (24%) and 0.07 (13%) over the KAT and the IVT respectively.

Table 2. The average performance of each tracker in the whole dataset. This value is computed by averaging the CATA values of the 12 categories.

	IVT-RGB	KAT-Intensity	Proposed-CSURF
Average	0.54	0.49	0.61

5.3 Robustness to Changes of Illumination and Object Appearance

In this section we analyze the performance of the proposed tracker with changes in illumination and object appearance. We select the best feature for each tracker: RGB with FBT, intensity with KAT, RGB with IVT and C-SURF with the proposed tracker.

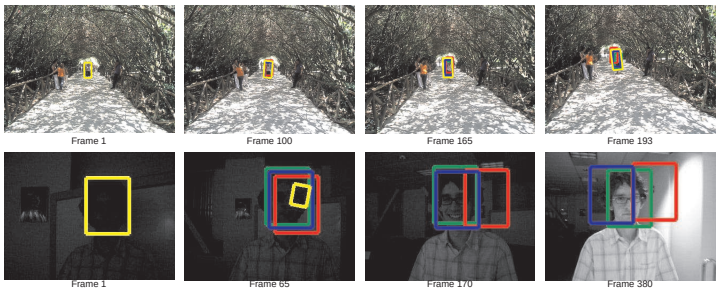

Fig. 3. The first row: a person undergoing foliage-like illumination. The second row: a person undergoing large changes in illumination intensity. Results of the 4 trackers are shown: yellow - IVT; red - FBT; blue - KAT; green - ours. Our tracker is able to track the targets despite abrupt changes of illumination over space and time.

In Figure 3, two targets undergoing different illumination conditions are being tracked. In the first row, the target is a person walking under dense foliage with abrupt changes in lighting over space and time. The FBT suffers small drift in each time step and eventually loses the object at frame 193 as the limited discriminative power of the feature and the presence of similar patterns in foreground and background. The KAT and IVT get small drifts at the end when the object turns left and the trackers are locked at an illuminated region in the background. The uneven illumination does not affect our tracker. Despite many false movements of the object, the results of tracker remain accurate.

In the second row of Figure 3, the target is a person moving from a dark area to a brighter area with the illumination intensity changing largely. The IVT gets difficulty at the beginning of the sequence since it confuses the face with the background. Due to lack of invariance of the features, the FBT and KAT drift away from the object at frame 170 and 380 respectively. Our tracker successfully tracks the object because of the invariance to light intensity changes of the SURF feature.

Figure 4 demonstrates the performance of our tracker with changes of object appearance. In the first row, a face undergoes translation and rotation movements. At frame 50, the other 3 trackers lose the object due to the rotation movement of the object. After that the KAT and the IVT accidentally recover the object. But only the KAT and the proposed tracker successfully follow the object till the end. This video shows the ability of our tracker in coping with new patterns when the object rotates in the vertical axis. The use of highly discriminative features enables the tracker to avoid the confusion of the black area of the head with the blackboard in the background. This is the reason why the FBT loses the object. In the second row of Figure 4, an experimental video is shown to demonstrate our tracker's robustness to changes in object appearance. The target is a rotating three-color ball. At frame 50 when the pink area occurs, the other 3 trackers drift away while our tracker still can follow the ball.

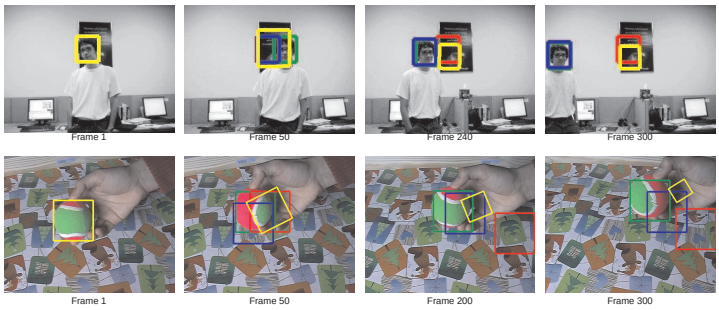

Fig. 4. The first row: A person with translation and rotation motion. The second row: a 3-color ball undergoing rotation. In both cases, the targets undergo large variations in appearance. Our tracker can adapt to new appearance patterns and successfully follow the targets.

5.4 Robustness to Confusion and Partial Occlusions

Figure 5 shows examples of tracking under clutter and confusion conditions. In the first row, a pupil in uniform runs in front of many other classmates. We notice that as all the pupils are in uniform, the object looks very similar to other nearby objects. This causes KAT and FBT to fail at the beginning and IVT to fail at frame 75. Our tracker succeeds in disregarding the confusion as the use of the discriminative feature, which allows it to focus on very distinct pattern of the object that discriminate it from the background patterns. A similar phenomenon can be seen in the second row and the third row of Figure 5, where our tracker successfully follows a person running in a marathon with similar objects in the vicinity and Waldo moving in front of a Where's Waldo picture.

The two videos in Figure 6 demonstrate the ability of the proposed tracker to cope with partial occlusion. Before the object enters the occlusion area, it enters a shadow area. As can be seen from the two videos, both IVT and KAT fail as the shadows change the object appearance abruptly. With the red ball video in the first row, the FBT overcomes the shadow area due to the distinct color of the object. It however fails to follow the toy car in the second row when it is occluded. Due the shadow invariance property of C-SURF, our tracker does not get affected by shadow and successfully follow the objects in both situations.

5.5 Failure Analysis

We search for failure cases of the proposed tracker. Figure 7 depicts 3 situations where the proposed tracker fails. In the first row, the target gets bigger as the camera is zooming in. The proposed tracker does not drift away from the target. However it cannot cope with the changing size of the object. The IVT however precisely follow the target. The reason is that the IVT considers scaling while searching for the object. The proposed tracker, on the other hand, uses a fixed template window. In the second row, the target is a flock of birds. We notice that

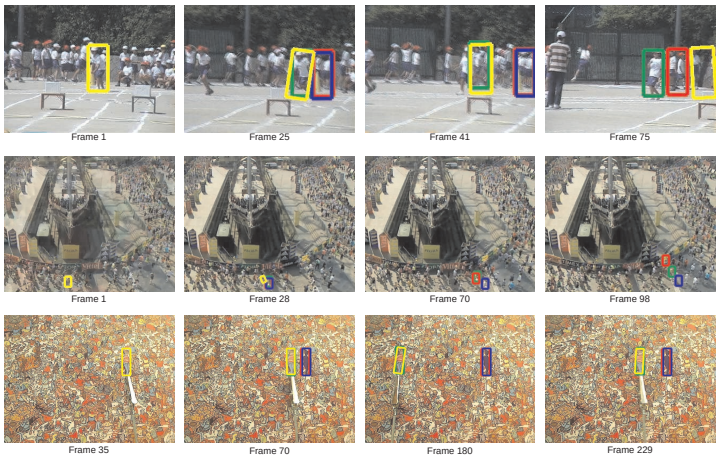

Fig. 5. Tracking under cluttered scene and confusion. The first row: a pupil running in front of other classmates in the same uniform. The second row: a person in a marathon. The third row: tracking Waldo. Our tracker successfully discriminates the targets from nearby objects with similar appearance and cluttered background due to the use of invariant feature.

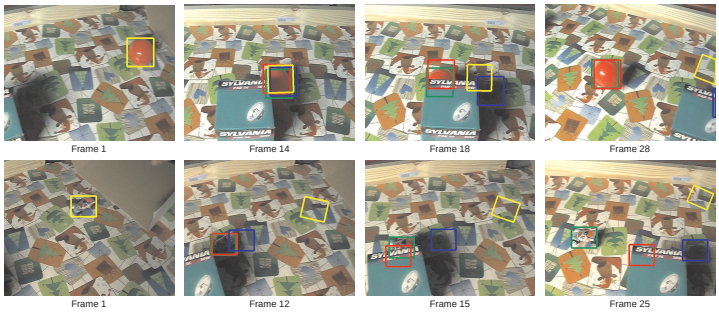

Fig. 6. Tracking under partial occlusion. The targets are the red ball and the toy car undergoing partial occlusion. Our tracker is able to follow the target accurately when they enter shadow and partial occlusion.

the dynamics of the flock shape makes it very difficult for the trackers to follow where many background patterns are present in the object region. In the third row, the changing light color is the challenge. At frame 13, the light becomes completely dark. We notice that the second and third rows represent two very extreme cases in visual tracking.

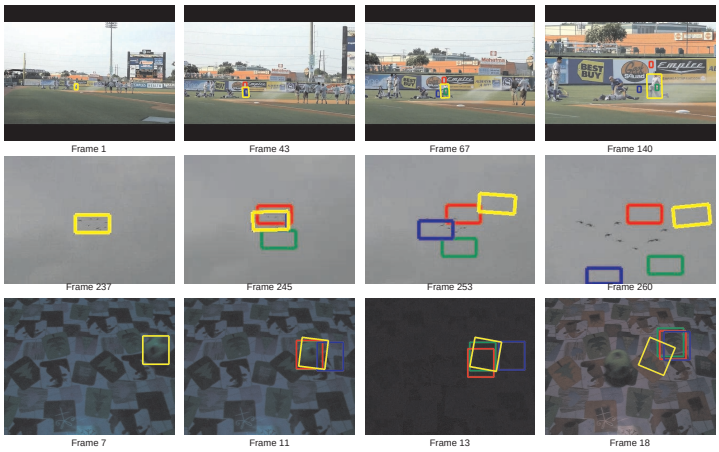

Fig. 7. Failure cases of the proposed tracker. The first row: tracking under zooming-in condition. The second row: tracking a flock of birds. The third row: tracking under changing light color.

6 Conclusion

We have presented a tracker that takes advantage of the discriminative tracking framework and highly discriminative power of SURF-based features. The resulting tracker is capable of tracking objects under changes in lighting conditions and object appearance and undergoing partial occlusion. The proposed tracker is also robust against confusion and cluttered scenes where there are similar objects in the vicinity of the tracked object.

The combination of SURF with the C invariant and the opponent color space are shown to be the best choice for the discriminative tracking framework. The conclusion goes along with the finding in Van de Sande et al. [9] in the object classification task. This makes an interesting link between object classification and discriminative tracking.

Acknowledgments

We thank Theo Gevers and Arjan Gijsenij for insightful comments and discussions.

References

1. Zivkovic, Z., Kröse, B.: An EM-like algorithm for color-histogram-based object tracking. CVPR (2004)
2. Nguyen, H., Smeulders, A.: Robust track using foreground-background texture discrimination. IJCV **68(3)** (2006) 277–294

3. Ross, D., Lim, J., R.S.Lin: Incremental learning for robust visual tracking. IJCV **77** (2008) 125–141
4. Nguyen, H., Smeulders, A.: Fast occluded object tracking by a robust appearance filter. PAMI **26** (2004) 1099–1104
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
6. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). CVIU **110** (2008) 346–359
7. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. PAMI **27** (2005) 1615–1630
8. Burghouts, G.J., Geusebroek, J.M.: Performance evaluation of local colour invariants. CVIU **113** (2009) 48–62
9. Van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. PAMI (2010)
10. Uijlings, J., Smeulders, A., Scha, R.: Real-time bag of words, approximately. In: CIVR. (2009)
11. Alper Yimaz, O.J., Shah, M.: Object tracking: a survey. ACM Computing Surveys **38** (2006)
12. Sheikh, Y., Javed, O., Kanade, T.: Background subtraction for freely moving cameras. In: ICCV. (2009)
13. Grabner, H., Grabner, M., Bischof, H.: Real-time tracking via on-line boosting. In: BMVC. (2006)
14. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: CVPR. (2009)
15. Authors: Twelve hard cases in visual tracking. In: Technical report; in preparation for Performance Evaluation of Tracking Systems Workshop. (2010)
16. Grabner, M., Grabner, H., Bischof, H.: Learning features for tracking. In: CVPR. (2007)
17. Avidan, S.: Ensemble tracking. PAMI **29** (2007) 261–271
18. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised on-line boosting for robust tracking. In: ECCV. (2008)
19. Mahadevan, V., Vasconcelos, N.: Saliency-based discriminant tracking. In: CVPR. (2009)
20. Tremeau, A., Tominaga, S., Plataniotis, K.N.: Color in image and video processing: most recent trends and future research directions. EURASIP Journal on Image and Video Processing **2008** (2008)
21. He, W., Yamashita, T., Lu, H., Lao, S.: Surf tracking. In: ICCV. (2009)
22. Zhou, H., Yuan, Y., Shi, C.: Object tracking using sift features and mean shift. CVIU **113** (2009) 345–352
23. Tran, S., Davis, L.: Robust object tracking with regional affine invariant features, ICCV (2007)
24. Stern, H., Efros, B.: Adaptive color space switching for tracking under varying illumination. IVC **23** (2005) 353–364
25. Wang, J., Yagi, Y.: Integrating color and shape-texture features for adaptive real-time object tracking. TIP **17** (2008) 235–240
26. Collins, R.T., Liu, Y., Leordeanu, M.: Online selection of discriminative tracking features. PAMI **27** (2005) 1631–1643
27. Geusebroek, J.M., van den Boomgaard, R., Smeulders, A.W.M., Geerts, H.: Color invariance. PAMI **23** (2001) 1338–1350