The Enhanced Flock of Trackers

Tomáš Vojíř and Jiří Matas

Abstract The paper presents contributions to the design of the Flock of Trackers (FoT). The FoT estimates the pose of the tracked object by robustly combining displacement estimates from a subset of local trackers that cover the object and has been. The enhancements of the Flock of Trackers are: (i) new reliability predictors for the local trackers - the Neighbourhood consistency predictor and the Markov predictor, (ii) new rules for combining the predictions and (iii) introduction of a RANSAC-based estimator of object motion. The enhanced FoT was extensively tested on 62 sequences. Most of the sequences are standard and used in the literature. The improved FoT showed performance superior to the reference method. For all 62 sequences, the ground truth is made publicly available.

1 Introduction

The term ”visual tracking” covers a broad range of methods for estimation of the pose and state of some entity in a sequence of images assuming temporal dependence of the estimated quantities. The complexity of the tracked entity may range from a rectangular region to a deformable or articulated object like human or animal body. The pose refers to geometric parameters of the entity, in 2D tracking typically a position, often with scale and rotation. The state represents all other information about the object, e.g. its past appearance, dynamics or even a discriminative classifier for redection [8, 6] or pointers to objects in the image with correlated motion [5].

Tomáš Vojíř
The Center for Machine Perception, FEE CTU, Prague, Karlovo namesti 13, 121 35 Praha 2, Czech Republic, e-mail: `vojirtom@cmp.felk.cvut.cz`

Jiří Matas
The Center for Machine Perception, FEE CTU, Prague, Karlovo namesti 13, 121 35 Praha 2, Czech Republic, e-mail: `matas@cmp.felk.cvut.cz`

Short-term frame-to-frame tracking is the most widely used form of visual tracking. It formulates the problem as a sequential casual estimation of the pose of an object in the next frame given the pose in the current frame. Short term trackers do not consider the problems of object re-detection after occlusion or disappearance - some pose parameters are always output, regardless of the fact the tracked entity is no more (visible) in the field of view. Prominent examples of short term trackers are the Lucas-Kanade [11] and mean-shift [3] trackers. The popularity of short-term trackers stems from their simplicity and, consequently, high speed and applicability in a wide range of conditions.

The Flock of Trackers (FoT). Recently, Kolsch and Turk [10] and Kalal et al. [8, 6] have shown that a very robust short-term tracker is obtained if a collection (a ”flock”) of local short-term trackers covering the object is run in parallel and the object motion is estimated from the displacements or, more generally, from transformation estimates of the local trackers. Each local tracker is attached to a certain area specified in the object coordinate frame. Following [8, 6, 14], we adopted the Lucas-Kanade [11] algorithm for local tracking.

The block structure of the Flock of Trackers is illustrated in Fig. 1. In its simplest form, the FoT requires only two components: a local short-term tracker, multiple instance of which are run on different areas of the object and provide image-to-image correspondence, and a (global) object motion estimation module robustly combining the local estimates.

The FoT is a very attractive short-term tracker. In comparison to many recently published methods, it is relatively simple and transparent and yet its performance is close to the state of the art [14]. Its internal structure allows handling heavy partial occlusion and local non-rigid changes and it makes the pose estimation robust, since it does not depend on a single global property of the object but rather on a composition of many local (weak) features. The FoT is slower then a monolithic short-term tracker, but not by orders of magnitude since the local trackers operate on small patches are thus fast.

In this chapter we show that the performance of the FoT is significantly improved if the object motion module is provided with a confidence measure in the reliability of the local tracker motion estimates. We propose (i) new reliability predictors for the local trackers, (ii) new rules for combining the predictions and (iii) introduce a new, RANSAC-based estimator of the object motion.

The local tracker reliability predictors presented in the chapter fall into two groups. The first group contains methods that are applicable to any short-term tracker and includes estimators based on the apparent magnitude of the intra-frame appearance change like the sum of squared intensity differences (SSD), the normalized cross-correlation (NCC) and the forward-backward procedure (FB). The forward-backward procedure runs the Lucas-Kanade tracker [11] twice, once in the forward direction time, as in a standard implementation, and then a second (extra) run is made in the reverse direction. The probability of being an oulier, i.e. of tracker failure, is a function of the distance of the initial position and the position reached by the FB procedure.

The second group of local tracker reliability predictors includes two estimators applicable only to trackers running multiple local trackers, such as the FoT. One, a new predictor based on the consistency of motion estimates in a local neighbourhood (P_N), exploits the fact that it is unlikely for a local motion estimate to be correct if it differs significantly from other motion estimates in its neighbourhood. The second new predictor reflects past performance of the local tracker. If a local tracker motion estimate has (often) been an outlier in the (recent) past, i.e. it was inconsistent with the global motion estimate, it is not likely to be correct in the current frame. This occurs for instance when the area covered by the local tracker is occluded or because the area is not suitable for tracking (e.g. it has near constant intensity). This local predictor of tracker reliability is called the Markov predictor (P_M), since it models the sequence of predicted states (either inlier or outlier) as a Markov chain.

The Markov predictor uses the global object motion estimates as ground truth in judging the correctness of local tracker motion. Naturally, the global motion esti-

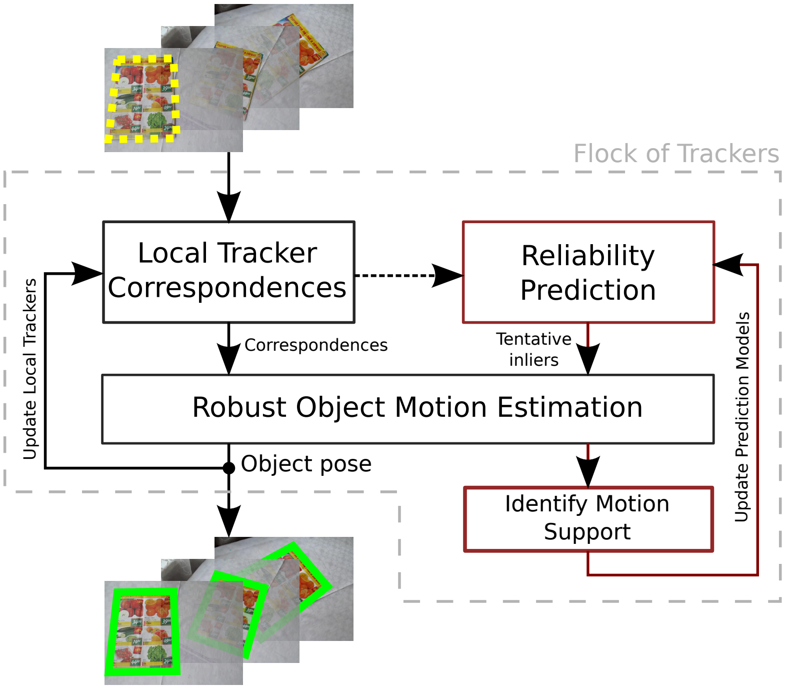


Fig. 1: Block structure of the Flock of Trackers (FoT). Correspondences (motion estimates) between two images, given the previous object pose and two consecutive images, are produced by local trackers. Simultaneously, reliability is estimated for each motion estimate. The object pose in the next frame is robustly estimated from a subset of most reliable motion estimates called tentative inliers.

mate may be correct or incorrect, but the latter case need not be considered since the FoT has failed anyway.

Combination of predictors. With the exception of the forward-backward procedure, the evaluation of the reliability prediction is fast in comparison with the time it takes to calculate the local motion estimate. It is therefore reasonable to combine all fast predictors to achieve high accuracy and avoid, if possible, the FB procedure.

We show that the Markov and Neighbourhood predictors, both on their own and when combined with the normalized cross-correlation predictor P_{ρ}, are more reliable than the normalized cross-correlation predictor combined with the FB procedure used in the reference method [7]. The new predictors are computed efficiently at a cost of about 10% of the complete FoT procedure whereas the forward-backward procedure slows down tracking approximately by a factor of two, since the most time consuming part of the process, the Lucas-Kanade local optimization, is run twice. With the proposed combination of reliability predictors, a FoT with much higher robustness to local tracker problems is obtained with negligible extra computational cost.

We introduce and compare two *predictor combination* schemes: a predictor combination method approximating a likelihood-based decision (denoted as $\mathcal{P}_{\Theta}$) and a simple ad-hoc predictor combination (denoted as $\mathcal{P}_{\wedge}$ combination). The ad-hoc combination sets a reliability threshold for each predictor (i.e. P_{ρ}, P_{M}, P_{N}) and the local tracker has to satisfy all of the condition to be used for pose estimation. The likelihood-based method orders the local trackers based on their likelihood of being correct. It allows choosing either the *n* best local trackers or a variable size subset that on average maintains a certain level of the inlier ratio for robust object pose estimation. In experiments, we set the operating point of the $\mathcal{P}_{\Theta}$ combination so that the number of the local trackers in the predicted inlier set (i.e. points, from which the object pose is estimated) is the same in each frame for the $\mathcal{P}_{\wedge}$ and the $\mathcal{P}_{\Theta}$ methods. The methods are evaluated by inlier prediction precision and by how many true inliers were in a predicted set.

Finally, we turn our attention to *robust object motion estimation* that takes as input the local motion estimates equipped with their reliability predictions.

The reference method is the Median-Flow (MF) [7] tracker which was shown to be comparable to the state-of-the-art where object motion, which is assumed to be well modelled by translation and scaling, is estimated by the median of a subset of local tracker responses.

Theoretically, the median with the breakdown point 0.5 is robust up to 50% of corrupted data. Since a single displacement vector give an estimate of the translation, the median as a translation estimator is robust up to 50% of incorrect local trackers. For scale estimation a ratio of pairwise distances of local trackers is used as an estimate of scale change, therefore a median is robust up to $100 \times (1 - \sqrt{0.5})\% \doteq 29\%$ of incorrect local trackers for scale estimation step.

In practice, the outlier tolerance is often lower since the outliers ”conspire”. The outlier motion estimates originate from occluded or background areas. All local motion estimates in such areas are typically consistent with a motion of the occluding object or the background, i.e. they are higher or lower than the tracked object motion and bias the median based estimate. In challenging tracking scenarios presented in Section 6, the inlier percentage was often not sufficient for the median-based estimation of global motion and it failed when used without local tracker reliability prediction.

We show that RANSAC [2, 4] followed by least square fitting of inliers (LS) as model estimator is a preferable alternative to the median estimator. There are three main advantages of using the RANSAC+LS estimator: the model is estimated consistently (i.e. translation estimation is not separated from scale estimation), the motion model is not constrained to translation, scale and rotation; affine transformation or a homography requires only to change the sample size and it handles higher outlier percentages.

The rest of the paper is structured as follows. Section 3 proposes two new predictors of local tracker failure and discusses the predictor parameters selection. Section 4 discusses predictor combinations. Section 5 introduce RANSAC as a model estimator. Finally, Section 6 evaluates the proposed improvements. Conclusions are given in Section 7. This paper is an extension of a workshop paper [14].

2 Related Work

The work presented in the chapter builds on Kalal et al. [7] who mainly used the FoT as a tracking component of the powerful Tracking-Learning-Detection system, or TLD in short, long-term tracker [8]. Interestingly, with the improvements in presented in the chapter, the FoT with the combined new reliability prediction of local trackers approaches performance of the TLD framework on sequences where redetection is not needed, and yet is significantly faster.

The baseline FoT [7] places local trackers on a regular grid, i.e. the local trackers cover the object uniformly. Object motion, which is assumed to be well modelled by translation and scaling, is estimated by the median of a subset of local tracker displacement estimates (translation) and the median of the relative change of distance between positions of local tracker pairs (scale).

For reliability prediction of local trackers, Kalal et al. [7] use several standard local tracker filtering methods, namely the normalised cross-correlation (or sum of squared differences) of the corresponding patches, and the consistency of the forward-backward procedure.

The original idea of exploiting a collection of trackers goes back at least to Kölsch et al. [10] who proposed the Flock of Features for fast hand tracking using local trackers (Lucas-Kanade [11]) with color histograms for replenishing of failed local trackers. They also enforce ”flock behaviour” [12] to detect failing local trackers. The output of their tracker is the median position of the local trackers, which manifests the flock behaviour.

Adam et al. [1] introduced FragTrack, which represents object by multiple patches (histograms of local areas). During tracking, each patch votes for an object pose by comparing its histogram to neighbourhood patch histograms. Robust statistics is then used to combine votes from multiple patches. Nejhum et al. [13] combine global description (histogram over the whole object) and a small number of rectangular blocks (weighted histograms) to determinate the most probable object location. An approximate boundary contour is then extracted using graph-cut segmentation. Block positions and weights are then updated. Kwon et al. [9] use local patch-based appearance model and an efficient scheme for online evolution of the local patch topology. For each frame, the Maximum a Posteriori (MAP) estimate is computed from the observation and transition models of local patches in a Bayesian manner.

3 Tracker reliability prediction methods

In this section, two novel methods for the local tracker reliability prediction are presented: section 3.3 describes the Neighbourhood consistency reliability predictor and section 3.4 presents the Markov predictor based on the long-term behaviour of the local tracker. Before that, two predictors used in the literature are described: the reliability predictor P$_\rho$ based on normalised cross-correlation of the corresponding patches in consecutive frames (section 3.1) and the forward-backward predictor (section 3.2

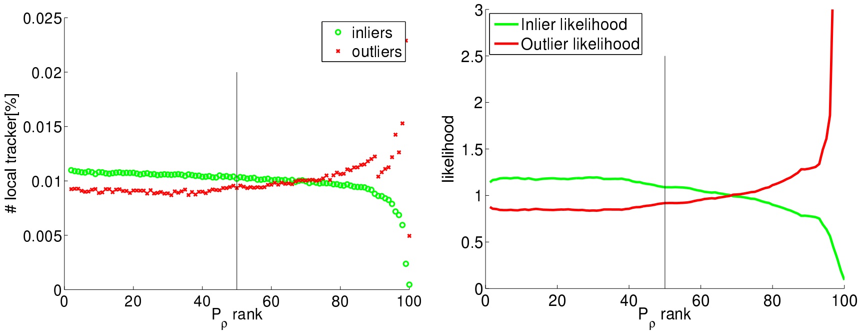
3.1 The NCC reliability predictor P$_\rho$

The first step of the predictor is to calculate for each local tracker the normalized cross-correlation NCC, eq. 1 between the patches T_1 and T_2 at corresponding positions and size (w, h) given by the motion estimate:

$$\begin{aligned} T_1'(x,y) &= T_1(x,y) - 1/(w \cdot h) \cdot \textstyle\sum_{x',y'} T_1(x',y') \\ T_2'(x,y) &= T_2(x,y) - 1/(w \cdot h) \cdot \textstyle\sum_{x',y'} T_2(x',y') \\ \text{NCC} &= \frac{\sum_{x,y}(T_1'(x,y) \cdot T_2'(x,y))}{\sqrt{\sum_{x,y} T_1'(x,y)^2 \cdot \sum_{x,y} T_2'(x,y)^2}} \end{aligned} \tag{1}$$

The P$_\rho$ predictor, introduced in [7] works as a ranking filter. It is difficult to find a general function linking the NCC to tracker reliability, since NCC values for all local trackers may change dramatically from frame to frame due to an illumination change, shadows, small drifts, etc. The local trackers are thus only sorted by NCC and their rank is used as a predictor.

The top 50% of the local trackers are predicted to be inlier (correct motion estimate), the rest as outliers (incorrect motion estimate). The threshold was selected empirically. Figure 2a shows the histogram of ranks for both inliers and outliers and supports the choice to filter 40%-50% of the worst local trackers, as the probability of being an inlier in the bottom half of the ranks is smaller than the probability of



(a)

(b)

Fig. 2: Properties of the P_ρ predictor averaged over a subset of the test sequences and all frames. (a) The histogram of NCC ranks ρ for local trackers with correct motion estimates (green) and incorrect motion estimates (red). (b) The correct/incorrect motion estimate ratio as a function of NCC rank ρ (green), the reciprocal value in red.

being an outlier. This is illustrated in figure 2b in terms of the likelihood ratio of being an inlier/outlier. Another interesting fact is that probability of being an outlier slightly rises around the 1-5 rank. This is caused by local trackers that are placed on the background (due to the bounding box representation of object or tracker drift) where a zero motion is estimated. The NCC values are very high on the static background.

Experimentally we observed that the P_ρ predictor is sensitivity to local tracking precision of the model and candidate patch - small misalignment may induce arbitrary large similarity difference. This often happens for articulated or non-rigid objects.

3.2 The forward-Backward reliability predictor P_{FB}

This underlying idea behind the forward-backward predictor is that the process of motion estimation between two images is independent of the order of the images. In an error-free situation, tracking an image region using Lucas and Kanade [11] gradient optimization from frame $1 \rightarrow 2$ and then the resulting image region from $2 \rightarrow 1$ will end up in the original position in the frame 1.

When the deviation from the original position in frame 1 is large, then at least one of the two motion estimates is inaccurate. It is not unreasonable to assume that reliability of the motion estimate is a monotonic function of the distance of the original position and the position reached by the forward-backward procedure.

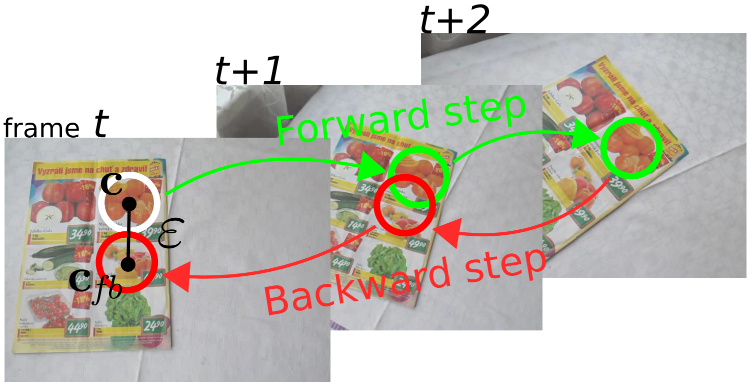


Fig. 3: A reference point of a regions of interest is tracked forward in time (from frame $t \rightarrow t+1 \rightarrow t+2$) and then backward. The positional forward-backward error $\varepsilon = \| \mathbf{c} - \mathbf{c}_{fb} \|^2$ is then used as a measure of tracker reliability.

The process may be generalised and the forward and backward direction tracking computed over larger number of frames. This is illustrated in Fig. 3.

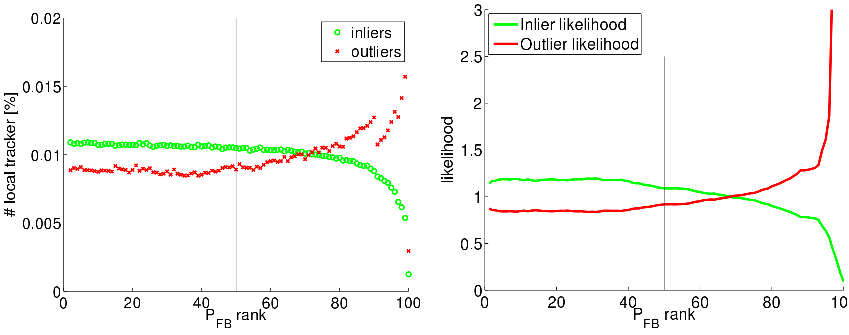
Figure 4a shows the histogram of FB distance ranks for correct and incorrect motion estimates and supports the choice to filter $30\% - 50\%$ of the worst local trackers, as the probability of being an inlier in the bottom half of the ranks is smaller than the probability of being an outlier. Figure 4b depicts the ratio of being an inlier or outlier respectively as function of the rank. Similarly to P_ρ predictor, the probability of being an outlier rises around the 1-5 rank. This is also caused by local trackers that are on the background and thus are consistent with FB procedure.

3.3 The neighbourhood consistency predictor P_N

The assumption behind the neighbourhood consistency predictor is that the motion of neighbouring local trackers is often very similar, whereas a failing local tracker returns a random displacement.

The P_N predictor is implemented as follows. For each local tracker i, a set of neighbouring local trackers N_i is defined. In all experiments, N_i included the four nearest neighbours of i. The neighbourhood consistency score S_i^N, the number of the neighbourhood local trackers that have a similar displacement. The process is visualised in Fig. 5.

We tested two definitions of the scoring functions given in eq. 2 and eq. 3. The latter has superior performance and was adopted.



(a) (b)

Fig. 4: Properties of the $\mathrm{P_{FB}}$ predictor averaged over a subset of the test sequences and all frames. (a) The histogram of FB ranks for local trackers with correct motion estimates (green) and incorrect motion estimates (red). (b) The correct/incorrect motion estimate ratio as a function of the FB rank (green), the reciprocal value in red.

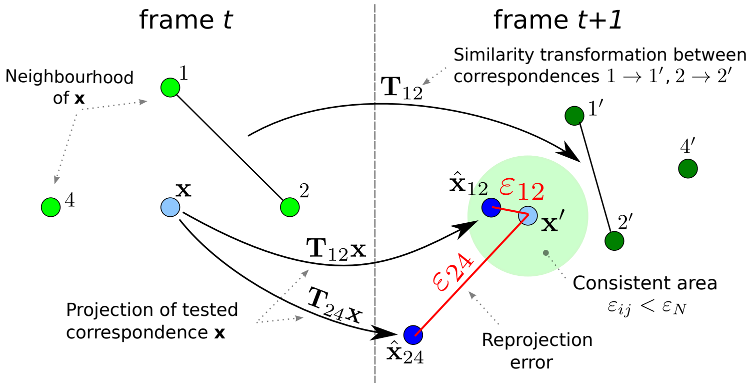


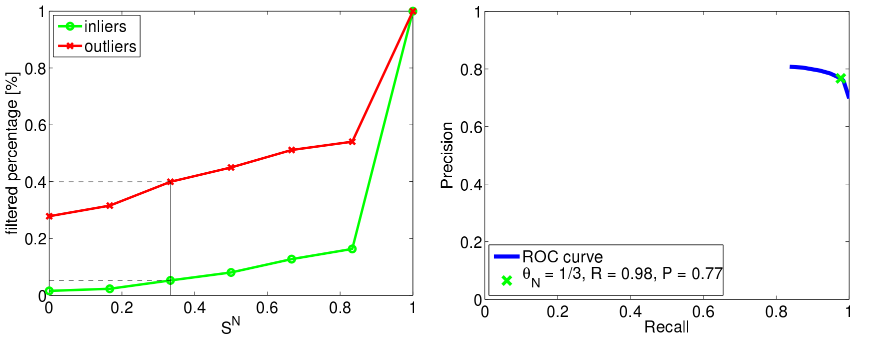
Fig. 5: Neighbourhood score computation for two pairs of correspondences. Each unique pair of correspondences (green) $i, j \in 1, 2, 3, 4$ generate a similarity transformation $\mathbf{T}_{ij}$. The tested (blue) correspondence $\mathbf{x}$ is transform by the estimated similarities and the reprojection error $\varepsilon_{ij} = \parallel \hat{\mathbf{x}}_{ij} - \mathbf{x}' \parallel^2$ is computed. The final score is the number of $\varepsilon_{ij} < varepsilon_N$ (number of $\hat{\mathbf{x}}_{ij}$ points inside green circle around $\mathbf{x}'$).

$$
\begin{aligned}
&\mathrm{S'}^{N}_{i} = \frac{1}{\mathrm{Z}} \sum_{j \in N_i} \left[|\angle_{ij}| < \varepsilon_{\angle} \quad \& \quad \frac{\|\Delta_j\|}{\|\Delta_i\|} \in (\varepsilon_l, \varepsilon_h) \right] \\
&\text{where} \quad [expression] = \begin{cases} 1 & \text{if } \textit{expression} \text{ is true} \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
\tag{2}
$$

and where $\varepsilon_\angle$ is the maximum angle threshold, $(\varepsilon_l, \varepsilon_h)$ bounding range for the ratio of displacement magnitudes, Δ_i is the displacement of local tracker *i* and $\mathrm{Z} = \frac{4}{N_i}$ is normalization to 4-neighbourhood (to account for corners and sides of bounding box). A local tracker is defined to be consistent if $\mathrm{S}_i^N \geq \theta$, where θ is a threshold for this predictor.

$$\mathrm{S}_i^N = \frac{1}{\mathrm{Z}} \sum_{\substack{j,k \in N_i \\ j \neq k}} [\| T_{jk}\mathbf{x}_i - \mathbf{x}'_i \|^2 < \varepsilon_N] \\ \text{where} \quad [expression] = \begin{cases} 1 & \text{if } \textit{expression} \text{ is true} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Scoring function S_i^N counts the number of triplets of consistent local tracker. The transformation T_{jk} calculated from motion estimates of trackers j and k is applied on the reference point $\mathbf{x}$ of tracker i. If the transformed position $T_{jk}\mathbf{x}_i$ is within ε_N of its corresponding point $\mathbf{x}'_i$, one is added to the score. In experiments, ε_N was set to 2.



(a)

(b)

Fig. 6: Properties of the P_N predictor averaged over a subset of test sequences and all frames, (a) The normalized cumulative histogram of the local tracker state for S^N, (b) The Precision-Recall curve for P_N predictor

When used as a decision function which is required in one of the predictor combination methods described in the next section, there are finite number of possible thresholds depending on the number of neighbourhood local trackers.

Figure 6a shows a normalized cumulative histogram of the local tracker state for values of S^N normalized to range $< 0, 1 >$. Threshold $\theta_N = 1/6$ is chosen (i.e. S^N greater or equal to $1/3$ to predict an inlier state) as a good trade off between the ratio of filtered outliers and the false negative rate. Figure 6b shows the operating point of this threshold on the Precision-Recall curve.

3.4 The Markov reliability predictor P_M

The Markov reliability predictor (P_M) is based on the model of the past performance of a local tracker bound to a region specified by object coordinate frame. The model is in the form of a Markov chain with two states, $s_t \in \{0,1\}$, depicted in Fig. 7.

The predicted state (i.e. being correct - inlier or incorrect - outlier) of the local tracker depends on its state in the previous time instance and on the transition probabilities. The behaviour of each local tracker *i* at time *t* is modeled by transition matrix $\mathbf{T}_t^i$ described in Eq. 4, where s_t is the current state of the local tracker and whose columns sum to 1.

$$\mathbf{T}_t^i = \begin{bmatrix} p^i(s_{t+1} = 1 \mid s_t = 1) & p^i(s_{t+1} = 1 \mid s_t = 0) \\ p^i(s_{t+1} = 0 \mid s_t = 1) & p^i(s_{t+1} = 0 \mid s_t = 0) \end{bmatrix} \tag{4}$$

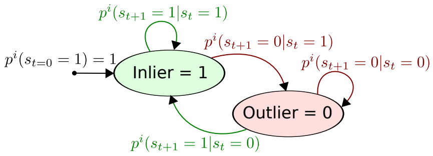

$p^i(s_{t+1} = 1|s_t = 1)$
$p^i(s_{t+1} = 0|s_t = 1)$
$p^i(s_{t=0} = 1) = 1$
Inlier = 1
$p^i(s_{t+1} = 0|s_t = 0)$
Outlier = 0
$p^i(s_{t+1} = 1|s_t = 0)$

Fig. 7: The state diagram of the Markov chain for the local tracker in the form of a two-state probabilistic automaton with transition probabilities p^i, where *i* identifies the local tracker and initial state $s_{t=0} = 1$.

The prediction that certain local tracker would be an tentative inlier (or an outlier) is done according to equation 5.

$$\begin{bmatrix} p^i(s_{t+1} = 1) \\ p^i(s_{t+1} = 0) \end{bmatrix} = \mathbf{T}_t^i \cdot \begin{bmatrix} p^i(s_t = 1) \\ p^i(s_t = 0) \end{bmatrix} \tag{5}$$

where $p^i(s_t = 1) \in \{0,1\}$ is binary and depends on the previous state (inlier/outlier) of the i^{th} local tracker. The left side of equation 5 are then probabilities that next state would be inlier (outlier).

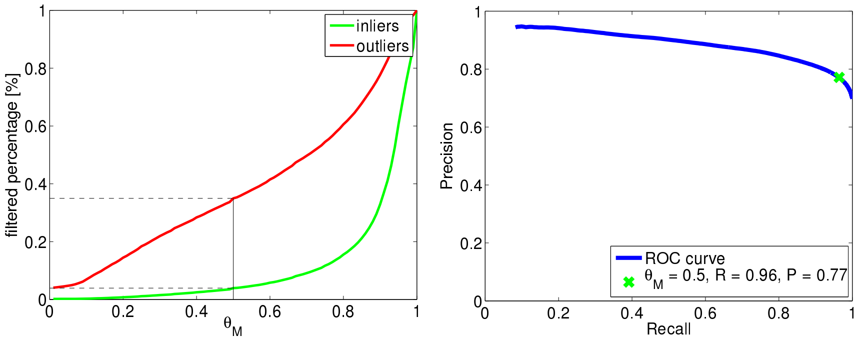
In the update stage, transition probabilities are re-estimated as follows :

$$\begin{aligned} p^i(s_{t+1} = 1 \mid s_t = 1) &= \frac{n_{11}^i}{n_1^i} \\ p^i(s_{t+1} = 1 \mid s_t = 0) &= \frac{n_{01}^i}{n_0^i} \end{aligned} \tag{6}$$

where n_1 and n_0 are numbers for the local tracker *i* being inlier (outlier respectively), and n_{11} and n_{01} are relative frequency for event that the local tracker *i* was inlier (outlier respectively) in the time t and inlier in the time $t+1$, for $t \in (0, t\rangle$.

The current state of the local tracker being inlier (outlier) is obtained by identifying local trackers that support the estimated global motion model.

When used as a decision function which is required in one of the predictor combination methods described in the next section, the observed characteristics support the natural choice of tresholding the inlier probability at 0.5. Figure 8a depicts the normalized cumulative histograms of a local tracker state for the Markov predictor values quantized to 100 bins. It shows how many inliers/outliers would be filtered out for different values of the θ_M threshold. The selected threshold 0.5 filtered out 4% of inliers and more than 35% of outliers. Figure 8b shows the operating point for threshold 0.5 on the Precision-Recall curve.



(a) (b)

Fig. 8: Properties of the P_M predictor averaged over a subset of test sequences and all frames, (a) The normalized cumulative histograms of a the local tracker state for $p(s_{t+1} = 1)$ values quantized to 100 bins, (b) The Precision-Recall curve for the P_M predictor

4 Methods for combining tracker reliability predictions.

This section describes two predictor combination methods – $\mathcal{P}_\Theta$ and $\mathcal{P}_\wedge$ and discusses their advantages and disadvantages. The explanation of the combination methods is elaborate for the combination of three predictors P_ρ, P_N, P_M.

4.1 The $\mathcal{P}_\Theta$ combination method

The $\mathcal{P}_\Theta$ combination estimates the likelihood of a local tracker being an inlier. The local tracker inlier likelihood is a function of three variables (i) P_ρ rank $\in \{1, 2, \ldots, 100\}$ quantized equally to 25 bins, $\rho = \lceil \frac{\text{rank}}{25} \rceil$ (ii) The P_N score $\in \{0, 1, 2, 3, 4\}$ in case of four-neighbourhood (iii) P_M probability $\in (0, 1)$ quantized equally to 25 bins. In the training phase a inlier/outlier likelihood ratio is estimated for all the combinations of variables using a Bayesian approach. resulting in a table with dimensions $25 \times 5 \times 25$. The combination can work in two modes (1) choose the fix threshold for local trackers inlier/outlier likelihood (2) take the n best local trackers, to form a local trackers subset for object pose estimation.

The advantage of this combination is a possibility to take an quasi-optimal decision (assuming independence of the individual predictors). The problem is formulated as a hypothesis test whether a local tracker is an inlier (outlier) given the likelihood ratio using a standard criterion such as Neyman–Pearson or min-max. The disadvantage is the need of the learning phase to the estimate local tracker inlier likelihood, which may overfit to the training data. In practice, the likelihood estimate generalized well enough to work in various scenarios.

4.2 The $\mathcal{P}_\wedge$ combination method

The $\mathcal{P}_\wedge$ predictor combination method computes responses of its constituent predictors and makes a binary decision for each of them (reliability below a threshold is interpreted as an outlier and visa versa). The final decision about the local tracker failure is a logical ”and” function:

$$\begin{aligned} f(\mathrm{P}_\rho, \mathrm{P}_N, \mathrm{P}_M) = {} & \rho > \text{median}(\rho) \\ & \wedge\ \mathrm{S}^N > \theta_N \\ & \wedge\ p(s_{t+1} = 1) > \theta_M \end{aligned} \tag{7}$$

The $\mathcal{P}_\wedge$ combination method assumes that since local tracker predictors exploit complementary information (i.e. P_ρ predictor – local appearance, P_M – temporal behaviour, P_N predictor – spatial consistency), parameters and threshold values of the inlier/outlier decision may be set independently.

5 RANSAC

The median estimator is robust and has a breakdown point 0.5. However, as shown in the experimental section, the percentage of correct local motion estimates is lower in many situations. Moreover, the median is biased if the noise is biased, which causes drifting of the tracker. This drifting happens in cases, where the background is static or locally static around the object of interest, e.g. when the bounding box is not a precise representation of the object shape and some local trackers are placed on the background.

We propose to use RANSAC for transformation estimation and show experimentally its superiority. This method has two main advantages over the median: (1) Is more robust to outliers (2) using unbiased least-square fitting to estimate transformation (up to homography).

6 Performance evaluation

6.1 The test data

The performance of the FoT with combined reliability prediction of local trackers and RANSAC-based object motion estimation was tested on challenging video sequences collected from a number of recently published papers. The sequences include object occlusion (or disappearance), illumination changes, fast motion, different object sizes and object appearance variance. The videos vary in length, contain highly articulated object and background clutter; some have poor visual quality. Targets in the sequences exhibit out-of-plane and in-plane rotation and some have homogeneous surfaces almost without texture. The sequences are described in Tab. 1. For details about the sequences visit `http://cmp.felk.cvut.cz/~vojirtom/dataset`. The lists of authors who kindly provided the sequences is available on the web site.

6.2 The experimental set-up

In all experiments, a frame was considered correctly tracked if the overlap with the ground truth is greater than 0.5, with the exception of experiment 6.6 where the influence of the initialization of the tracker was assessed. Since in this case the bounding boxes are randomly generated and may not fully overlap the object, the threshold was lower to 0.3, see Fig. 12. The overlap was measured as $o = \frac{area(T \cap G)}{area(T \cup G)}$, where T is object bounding box reported by the tracker and G is ground truth bounding box.

In the experiments, the predictor of neighbourhood consistency (P_N) and the Markov predictor (P_M) were run as explained in Section 3. The normalized cross-correlation (P_ρ) and the forward-backward procedure rank local trackers and treat the top 50% as inliers. Combinations of two or more predictors use the $\mathcal{P}_\wedge$ approach. Predictors are denoted by the names of their error measure, except for the combination $\mathrm{P}_M + \mathrm{P}_\rho + \mathrm{P}_N$ which is abbreviated to Σ.

6.3 Comparison of $\mathcal{P}_\wedge$ combination vs. $\mathcal{P}_\Theta$ combination

The $\mathcal{P}_\wedge$ predictor combination is compared with the $\mathcal{P}_\Theta$ combination in terms of inlier prediction precision. To make results comparable the measurement was done at the operating point of $\mathcal{P}_\wedge$ combination, since this method does not guarantee a number of predicted inliers and does not have any means for choosing n-best in contrast to $\mathcal{P}_\Theta$ combination.

The $\mathcal{P}_\Theta$ combination needs to learn likelihoods for the combined likelihood table of three criterion variables. A leave one out cross-validation was used to split the dataset to the training and validation sets. That means that for evaluation on sequence i the table is learned on all sequences except the sequence i. True inliers were extracted by comparing frame-to-frame tracking results with corresponding ground truth positions and criteria variables were recorded. The recorded values (P_N Score, P_M probability, P_ρ rank) were quantized (to 5, 25, 25 bins) and used to compute the inlier - outlier likelihood. Entries of the combined likelihood table are addressed by the quantized criteria values.

Results in table 2 show that the two combination methods perform similarly. The $\mathcal{P}_\wedge$ predictor combination has an advantage that it does not require learning in advance. We choose to use the $\mathcal{P}_\wedge$ predictor combination to keep the tracker as independent as possible of the training data and other external variables (e.g. the precision of the ground truth used for extracting true inliers, the size of the dataset, diversity of dataset, etc.).

6.4 Comparison of the reliability prediction methods

We compared performance of individual predictors and combinations $\mathrm{P}_{FB \circ \rho}$ (reference [7]), $\mathrm{P}_{N \circ M}$ and P_Σ. All parameters for predictors were fixed for all sequences, as described in Section 4.2.

The performance was measured by the recall and the number of reinitialization needed to track the whole sequences (reinitialization after object disappearance are not counted). The recall is defined as the ratio of the number of frame where the estimated object rectangle had an overlap with the ground truth rectangle higher then 0.5 and the number of frames where the object is visible. Approximately speaking, recall is the percentage of the frames with the tracked object visible where the object was correctly tracked.

The results are summarized in tables 3 and 4. Both tables have the same structure. Each line starting with a number presents results on one of the 62 sequence. The last two lines summarize performance. The #best line compares the median flow object motion estimator (m, left) and the RANSAC-based estimator (r, right) by counting the number of sequences when median flow outperformed RANSAC (the number before the ":"), where RANSAC dominated (the number after the ":"), the number of "draws" is given in parentheses.

According to both the recall (table 3) and reinitialization (table 4) criteria, RANSAC performs better for all reliability predictors and their combinations. Results for different predictors and combinations are presented in different columns. The final line of the table compares the mean recall and reinitialization. RANSAC performs better in terms of the mean too.

The "mean" row allows comparison of the the reliability predictors, both individually and in combination. The combinations P$_{N \circ M}$ and P$_{\Sigma}$ perform the best, clearly better than any individual tracker and slightly better than the forward-backward procedure combined with the NCC. Note that the P$_{\Sigma}$ and even more P$_{N \circ M}$ are significantly faster than the FB procedure.

Fig. 9 visualizes the performance for selected combinations of predictors in a manner facilitating comparison. Two combinations of predictors P$_{\Sigma}$ and P$_{N \circ M}$ are clear the most reliable methods.

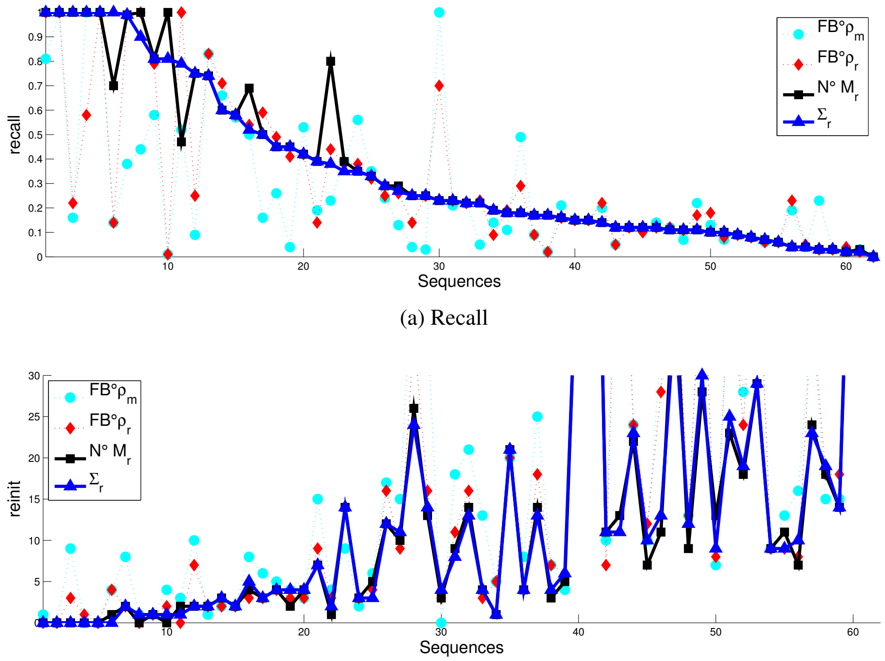
Visualization of predictor performance on selected frames from two challanging sequences are shown in Figs. 10 (*motor-bike*) and 11 *woman*. Predictor score is encoded in a "heat map" (red - high score, blue - low score). Green/Red boxes below predictor score encodes false positive (red dot with red background), false negative (green dot with red background), true positive (green dot with green background) and true negative (red dot with green background). On the right side of the image, a cut out shows the outlier (red) and inlier (green) motion estimates. The green-on-black images shows the area covered by inlier local trackers.

For the *motor-bike* sequence, it is somewhat surprising that the motion estimates on the biker are small. The biker is tracked by the cameraman and the position of the bike in the image stays roughly the same, the background exhibits fast apparent motion in the oposite direction. The FoT handle are rather large change of apperance of the biker between frames #31 and #77.

The *woman* sequence is more challenging, due to occlusion and changes of appearance due to walking, the number of local trackers providing correct motion estimates is small, as low as 19 out of 90 in frame # 18.

6.5 Comparison the speed of the reliability prediction methods

The FoT tracker is intended for real-time performance and thus the speed of local tracker predictor is important. The experiment was performed on all sequences listed in Tab. 1 and then the results were averaged. Speed was measured as the average time needed for frame-to-frame tracking. For results see Tab. 5. Processing time for I/O operations, including image loading, and other tasks not relevant to tracking were excluded. The P$_{\Sigma}$ predictor performs 41% faster than P$_{\mathrm{FB} \circ \rho}$. Most of the additional computation of P$_{\Sigma}$ over the P$_{\emptyset}$ lies in computation of normalized



(a) Recall

(b) Reinit count

Fig. 9: Comparison of the best performing predictor combinations and estimators in terms (a) Recall and (b) the number of reinitialization. Sequences (x-axis) are sorted by the recall measure of the P_{Σ} with RANSAC estimator.

cross-correlation. Therefore, the $\mathrm{P}_{N\circ M}$ overhead is negligible compared to reference predictor $\mathrm{P}_{\emptyset}$ (i.e. tracker without any predictor) and is more than two times faster then $\mathrm{P}_{\mathrm{FB}\circ\rho}$.

6.6 Robustness to bounding box initialization

For a tracking algorithm, it is highly desirable not to be sensitive to the initial pose specified by the object bounding box as it is often selected manually, with unknown precision.

If a part of the bounding box does not cover the object, the P_M predictor soon discover that the local trackers are consistently in the outlier set. This property can be used to define the object more precisely, e.g. as the set of object parts that are likely to be inliers according to P_M (see Figs. 10 and 11). Thus, with P_M, the global tracker may be insensitive to initialization.

This experiment tested the assumption on the challenging sequence Pedestrian 1, where an articulated object is tracked in a sequence containing background clutter and fast motions, which emphasize the need for correct initialization. We randomly generated 100 initial bounding boxes overlapping the object of interest (Fig. 12) and counted the correctly tracked frames (Tab. 6).

In the experiment, a frame was declared as correctly tracked if the overlap with the ground truth was greater than 0.3. The tracker with the P_{Σ} predictor performed about 90% better than the tracker with the $\mathrm{P}_{\mathrm{FB}\circ\rho}$ predictor and it was able to track the object correctly up to frame 84 on average.

Figs. 13a and 13b show the histograms of the number of correctly tracked frames for 100 runs with different initialization and Fig. 13c shows the 2D histogram of the number of correctly tracked frames by $\mathrm{P}_{\mathrm{FB}\circ\rho}$ and P_{Σ} initialized with the same random bounding box (to compare performance for individual random initialization).

7 Conclusions

We have presented a set of enhancements of the Flock of Trackers. First, new reliability prediction methods were introduced - the Neighbourhood consistency predictor and the Markov predictor.

Next, two methods for combining predictors, the ad-hoc $\mathcal{P}_{\wedge}$ and the likelihood thresholding $\mathcal{P}_{\Theta}$, were proposed and compared and similar performance was achieved. We decided to use $\mathcal{P}_{\wedge}$, because it is a straightforward approach without the need of learning the relevant statistics in advance.

Combined with the normalized cross-correlation predictor, the new Markov and Neighbourhood consistency predictors form a reliable compound predictor P_{Σ}. The P_{Σ} predictor was compared with the published $\mathrm{P}_{\mathrm{FB}\circ\rho}$ predictor and outperformed it in all criteria, i.e. in speed, recall, the number of reinitialization and the robustness to bounding box initialization. The simpler $\mathrm{P}_{N\circ M}$ combination performed almost identically and is faster. Finally, we have shown that the RANSAC-based global object motion estimator outperforms the published median flow algorithm.

The enhanced FoT was extensively tested on 62 sequences. Most of the sequences are standard and used in the literature. The improved FoT showed performance superior to the reference method, which competes well with the state-of-the-art [14].

For all 62 sequences, the ground truth is available at `http://cmp.felk.cvut.cz/~vojirtom/dataset`. For some of the sequences the ground truth has not been in the public domain till now.

Acknowledgements

The authors were supported by Czech Science Foundation Project GACR P103/12/G084 (JM) and by the Grant Agency of the CTU Prague project SGS13/142/OHK3/2T/13 (TV).

References

1. A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. In *CVPR*, 2006.
2. O. Chum, J. Matas, and J. Kittler. Locally optimized ransac. *Pattern Recognition*, 2003.
3. D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, 2000.
4. M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 1981.
5. H. Grabner, J. Matas, L. Van Gool, and P. Cattin. Tracking the invisible: Learning where the object might be. In *CVPR*, June.
6. Z. Kalal, J. Matas, and K. Mikolajczyk. P-N Learning: Bootstrapping Binary Classifiers by Structural Constraints. *CVPR*, 2010.
7. Z. Kalal, K. Mikolajczyk, and J. Matas. Forward-Backward Error: Automatic Detection of Tracking Failures. *ICPR*, 2010.
8. Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *PAMI*, 2012.
9. J. Kwon and K. Lee. Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling. In *CVPR*, 2009.
10. M. Kölsch and M. Turk. Fast 2d hand tracking with flocks of features and multi-cue integration. In *Workshop at CVPR*, 2004.
11. B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision (ijcai). In *IJCAI*, 1981.
12. C. W. Reynolds. Flocks, herds and schools: A distributed behavioral model. In *SIGGRAPH*, 1987.
13. S. Shahed Nejhum, J. Ho, and M.-H. Yang. Visual tracking with histograms and articulating blocks. In *CVPR*, 2008.
14. T. Vojíř and J. Matas. Robustifying the flock of trackers. In *CVWW*, 2011.

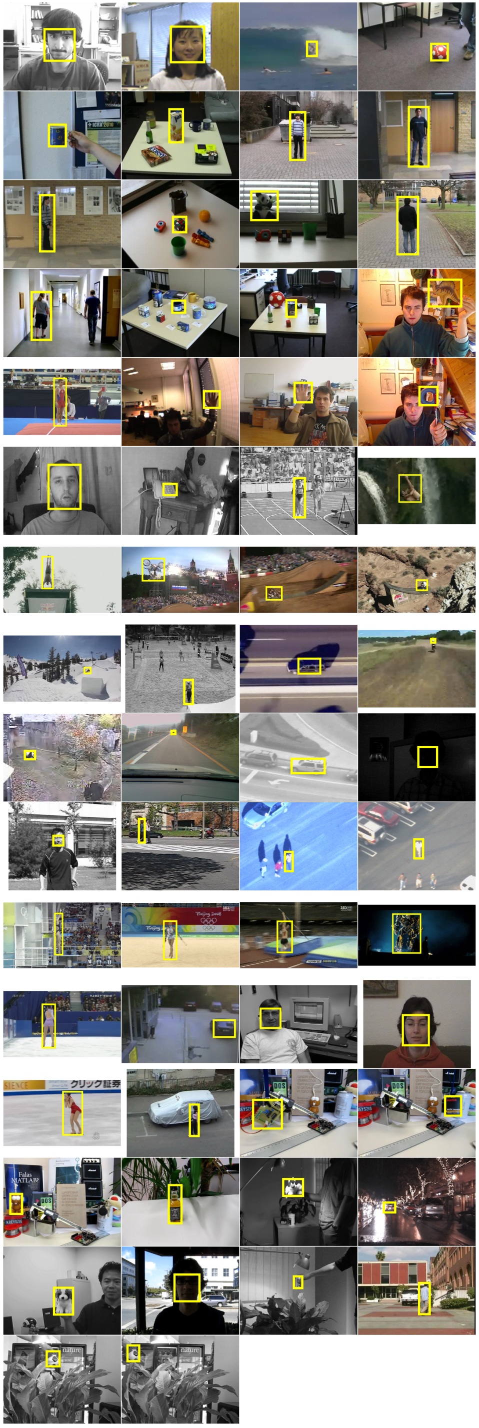
Seq. ID	name	#frames	#target visible	preview
1	girl	501	475	
2	OccludedFace2	815	815	
3	surfer	842	762	
4	Vid_A	602	602	
5	Vid_B	629	629	
6	Vid_C	404	404	
7	Vid_D	947	947	
8	Vid_E	305	305	
9	Vid_F	453	416	
10	Vid_G	716	716	
11	Vid_H	412	412	
12	Vid_I	1017	994	
13	Vid_J	388	383	
14	Vid_K	1020	1020	
15	Vid_L	1308	1308	
16	dinosaur	326	326	
17	gymnastics	567	567	
18	hand	244	244	
19	hand2	267	267	
20	torus	264	264	
21	head_motion	2351	2351	
22	shaking_camera	990	990	
23	track_running	503	397	
24	cliff-dive1	76	76	
25	cliff-dive2	69	61	
26	motocross1	164	164	
27	motocross2	23	23	
28	mountain-bike	228	228	
29	skiing	81	81	
30	volleyball	500	500	
31	car	945	860	
32	CarChase	9928	8660	
33	david	761	761	
34	jumping	313	313	
35	Motocross	2665	1412	
36	Panda	3000	2730	
37	pedestrian3	140	140	
38	pedestrian4	338	266	
39	pedestrian5	184	156	
40	Volkswagen	8576	5141	
41	diving	231	218	
42	gym	767	767	
43	jump	122	111	
44	trans	124	124	
45	Asada	661	661	
46	drunk2	1821	911	
47	dudek-face	1145	1145	
48	faceocc1	899	899	
49	figure_skating	624	624	
50	woman	597	597	
51	board	698	698	
52	box	1161	1129	
53	lemming	1336	1305	
54	liquor	1741	1704	
55	car11	393	393	
56	dog1	1353	1350	
57	Sylvestr	1344	1344	
58	trellis	569	569	
59	coke	292	270	
60	person	331	326	
61	tiger1	354	354	
62	tiger2	365	365	

Table 1: Overview of the test sequences. Basic information (left) and sample images with the selected object of interest (right) are shown. Full information about the sequences (authors, papers reporting results on the data, etc.) and the data are available at `http://cmp.felk.cvut.cz/~vojirtom/dataset`.

Seq.	$\varTheta$	$\wedge$
17	0.713±0.132	0.738±0.134
20	0.875±0.022	0.919±0.021
31	0.894±0.040	0.922±0.043
32	0.857±0.060	0.895±0.058
33	0.952±0.029	0.773±0.166
34	0.943±0.007	0.965±0.005
35	0.958±0.008	0.977±0.008
36	0.945±0.006	0.966±0.004
37	0.680±0.073	0.730±0.068
38	0.623±0.053	0.684±0.060
39	0.925±0.013	0.945±0.026
40	0.967±0.002	0.986±0.001
55	0.980±0.006	0.986±0.006
59	0.924±0.008	0.967±0.006
Mean	0.874±0.033	0.890±0.043

Table 2: The comparison of the $\mathcal{P}_{\wedge}$ predictor combination and the $\mathcal{P}_{\varTheta}$ combination in terms of inlier prediction precision ± variation. Averaged performance over a subset of sequences is reported in the last row. The subset of sequences was selected such that it includes mainly rigid objects; in some sequences also articulated objects (pedestrians) are tracked.

P	$\emptyset$	ρ	N	FB	M	FB ∘ ρ	$N \circ M$	Σ
Seq.	m ◇ r	m ◇ r	m ◇ r	m ◇ r	m ◇ r	m ◇ r	m ◇ r	m ◇ r
1	0.13 ◀ 0.18	0.12 ◀ 0.18	0.12 ◀ 0.18	0.11 ◀ 0.19	0.12 ◀ 0.18	0.11 ◀ 0.19	0.13 ◀ 0.18	0.13 ◀ 0.18
2	1.00 ▶ 0.40	1.00 ▶ 0.47	0.47 = 0.47	1.00 ▶ 0.70	0.47 ▶ 0.23	1.00 ▶ 0.70	0.22 ◀ 0.23	0.22 ◀ 0.23
3	0.02 ◀ 0.06	0.02 ◀ 0.06	0.07 = 0.07	0.06 = 0.06	0.07 ▶ 0.06	0.06 = 0.06	0.07 ▶ 0.06	0.07 ▶ 0.06
4	0.11 = 0.11	0.11 = 0.11	0.11 = 0.11	0.12 = 0.12	0.11 = 0.11	0.12 = 0.12	0.13 ▶ 0.11	0.12 ▶ 0.11
5	0.22 ◀ 0.38	0.24 ◀ 0.35	0.35 ◀ 0.38	0.23 ◀ 0.44	0.47 ◀ 1.00	0.23 ◀ 0.44	0.38 ◀ 0.80	0.38 = 0.38
6	0.50 ◀ 1.00	0.51 ◀ 1.00	0.47 ◀ 1.00	0.44 ◀ 1.00	0.48 ◀ 1.00	0.44 ◀ 1.00	0.47 ◀ 1.00	0.46 ◀ 0.90
7	0.57 ▶ 0.39	0.57 ▶ 0.39	0.57 ▶ 0.35	0.39 ▶ 0.38	0.58 ▶ 0.39	0.39 ▶ 0.38	0.58 ▶ 0.39	0.54 ▶ 0.35
8	0.57 ◀ 0.58	0.57 ◀ 0.58	0.57 ◀ 0.58	0.57 ◀ 0.58	0.57 = 0.57	0.57 ◀ 0.58	0.57 ◀ 0.58	0.57 ◀ 0.58
9	0.23 ◀ 0.32	0.23 ◀ 0.29	0.28 ◀ 0.29	0.24 ◀ 0.25	0.36 ▶ 0.28	0.24 ◀ 0.25	0.28 ◀ 0.29	0.36 ▶ 0.29
10	0.83 ◀ 1.00	0.83 ◀ 1.00	0.84 ◀ 1.00	0.81 ◀ 1.00	0.84 ◀ 1.00	0.81 ◀ 1.00	0.82 ◀ 1.00	0.83 ◀ 1.00
11	1.00 = 1.00	1.00 = 1.00	1.00 = 1.00	1.00 = 1.00	1.00 = 1.00	1.00 = 1.00	1.00 = 1.00	1.00 = 1.00
12	0.09 ◀ 0.12	0.10 ◀ 0.11	0.09 ◀ 0.11	0.07 ◀ 0.11	0.08 ◀ 0.11	0.07 ◀ 0.11	0.08 ◀ 0.11	0.08 ◀ 0.11
13	0.20 ▶ 0.17	0.20 ▶ 0.17	0.20 ▶ 0.17	0.21 ▶ 0.16	0.27 ▶ 0.16	0.21 ▶ 0.16	0.31 ▶ 0.16	0.31 ▶ 0.16
14	0.64 ▶ 0.42	0.64 ▶ 0.54	0.52 ◀ 0.97	0.52 ◀ 1.00	0.64 ▶ 0.43	0.52 ◀ 1.00	0.52 ▶ 0.47	0.52 ◀ 0.79
15	0.16 ◀ 0.78	0.16 ◀ 0.74	0.16 ◀ 0.74	0.16 ◀ 0.59	0.16 ◀ 0.56	0.16 ◀ 0.59	0.16 ◀ 0.50	0.16 ◀ 0.50
16	0.25 ◀ 0.39	0.25 ◀ 0.39	0.25 ◀ 0.38	0.19 ▶ 0.14	0.27 ◀ 0.39	0.19 ▶ 0.14	0.39 = 0.39	0.39 = 0.39
17	0.15 = 0.15	0.15 = 0.15	0.15 = 0.15	0.15 = 0.15	0.14 ◀ 0.15	0.15 = 0.15	0.15 = 0.15	0.15 = 0.15
18	0.09 ◀ 0.16	0.09 ◀ 0.15	0.11 ▶ 0.09	0.09 = 0.09	0.13 ▶ 0.09	0.09 = 0.09	0.16 ◀ 0.17	0.16 ◀ 0.17
19	0.09 ◀ 0.22	0.09 ◀ 0.14	0.07 ◀ 0.26	0.04 ◀ 0.14	0.05 ◀ 0.14	0.04 ◀ 0.14	0.05 ◀ 0.25	0.05 ◀ 0.25
20	0.20 ◀ 0.52	0.20 ◀ 0.56	0.21 ◀ 0.60	0.16 ◀ 0.22	0.46 ◀ 0.58	0.16 ◀ 0.22	0.54 ◀ 1.00	0.54 ◀ 1.00
21	0.77 ◀ 0.80	0.76 ▶ 0.52	0.77 ◀ 0.80	0.58 ◀ 0.79	0.77 ◀ 0.81	0.58 ◀ 0.79	0.77 ◀ 0.81	0.77 ◀ 0.81
22	0.15 = 0.15	0.15 = 0.15	0.15 = 0.15	0.15 = 0.15	0.15 = 0.15	0.15 = 0.15	0.15 = 0.15	0.15 = 0.15
23	0.09 ◀ 0.21	0.10 ◀ 0.21	0.09 ◀ 0.21	0.20 ◀ 0.22	0.13 ◀ 0.82	0.20 ◀ 0.22	0.13 ◀ 0.14	0.13 ◀ 0.14
24	0.34 ◀ 0.42	0.34 ◀ 0.41	0.34 ◀ 0.41	0.53 ▶ 0.42	0.42 ▶ 0.41	0.53 ▶ 0.42	0.43 ▶ 0.42	0.43 ▶ 0.42
25	0.15 ▶ 0.13	0.16 ▶ 0.11	0.15 ▶ 0.11	0.13 ◀ 0.18	0.11 ◀ 0.13	0.13 ◀ 0.18	0.15 ▶ 0.10	0.15 ▶ 0.10
26	0.18 ▶ 0.04	0.18 ▶ 0.03	0.45 ▶ 0.04	0.23 ▶ 0.03	0.16 ▶ 0.03	0.23 ▶ 0.03	0.05 ▶ 0.03	0.05 ▶ 0.03
27	0.83 ▶ 0.70	0.83 ▶ 0.70	0.83 ▶ 0.70	0.83 = 0.83	0.57 ◀ 0.91	0.83 = 0.83	0.57 ◀ 0.74	0.57 ◀ 0.74
28	0.40 ◀ 0.99	0.40 ◀ 0.99	0.43 ◀ 0.99	0.38 ◀ 0.99	0.82 ◀ 0.99	0.38 ◀ 0.99	0.82 ◀ 0.99	0.82 ◀ 0.99
29	0.07 ◀ 0.10	0.07 ◀ 0.10	0.07 ◀ 0.10	0.09 = 0.09	0.06 ◀ 0.07	0.09 = 0.09	0.06 ◀ 0.09	0.06 ◀ 0.09
30	0.23 ▶ 0.22	0.23 ▶ 0.22	0.22 = 0.22	0.22 = 0.22	0.22 = 0.22	0.22 = 0.22	0.22 = 0.22	0.22 = 0.22
31	0.50 ◀ 1.00	0.48 ◀ 0.58	1.00 ▶ 0.57	1.00 ▶ 0.58	0.75 ▶ 0.50	1.00 ▶ 0.58	0.61 ◀ 1.00	0.61 ◀ 1.00
32	0.01 ◀ 0.02	0.01 ◀ 0.02	0.02 = 0.02	0.03 ◀ 0.04	0.01 ◀ 0.02	0.03 ◀ 0.04	0.02 = 0.02	0.02 = 0.02
33	0.45 ◀ 0.60	0.59 ▶ 0.32	0.59 ▶ 0.50	0.01 = 0.01	0.39 ◀ 1.00	0.01 = 0.01	0.59 ◀ 1.00	0.59 ◀ 0.81
34	0.13 ◀ 0.24	0.14 ▶ 0.11	0.11 ◀ 0.24	0.05 = 0.05	0.14 ▶ 0.12	0.05 = 0.05	0.18 ▶ 0.12	0.18 ▶ 0.12
35	0.00 = 0.00	0.00 = 0.00	0.00 = 0.00	0.00 = 0.00	0.00 = 0.00	0.00 = 0.00	0.00 = 0.00	0.00 = 0.00
36	0.02 = 0.02	0.02 = 0.02	0.02 = 0.02	0.03 ▶ 0.02	0.02 ◀ 0.03	0.03 ▶ 0.02	0.02 ◀ 0.03	0.02 = 0.02
37	0.06 ◀ 0.19	0.06 ◀ 0.09	0.07 ◀ 0.09	0.14 ▶ 0.09	0.11 ◀ 0.32	0.14 ▶ 0.09	0.04 ◀ 0.19	0.04 ◀ 0.19
38	0.58 ▶ 0.54	0.58 ▶ 0.53	0.50 ◀ 0.70	0.66 ◀ 0.71	1.00 ▶ 0.56	0.66 ◀ 0.71	1.00 ▶ 0.60	1.00 ▶ 0.60
39	1.00 = 1.00	1.00 = 1.00	1.00 = 1.00	1.00 = 1.00	0.92 ◀ 1.00	1.00 = 1.00	0.89 ◀ 1.00	0.89 ◀ 1.00
40	0.05 ▶ 0.04	0.05 = 0.05	0.18 ◀ 0.24	0.19 ◀ 0.23	0.05 = 0.05	0.19 ◀ 0.23	0.18 ▶ 0.04	0.19 ▶ 0.04
41	0.13 ▶ 0.12	0.13 ▶ 0.12	0.13 ▶ 0.12	0.12 = 0.12	0.17 ▶ 0.12	0.12 = 0.12	0.16 ▶ 0.12	0.16 ▶ 0.12
42	0.04 ▶ 0.03	0.07 ▶ 0.03	0.03 = 0.03	0.03 = 0.03	0.03 = 0.03	0.03 = 0.03	0.03 = 0.03	0.03 = 0.03
43	0.06 ◀ 0.09	0.06 ◀ 0.10	0.10 = 0.10	0.11 ▶ 0.10	0.15 ▶ 0.14	0.11 ▶ 0.10	0.14 ▶ 0.12	0.14 ▶ 0.12
44	0.51 ▶ 0.38	0.44 ▶ 0.39	0.41 ◀ 0.50	0.56 ▶ 0.38	0.35 ◀ 0.40	0.56 ▶ 0.38	0.35 = 0.35	0.35 = 0.35
45	0.08 = 0.08	0.08 ◀ 0.15	0.15 ▶ 0.09	0.08 = 0.08	0.07 ◀ 0.09	0.08 = 0.08	0.09 ▶ 0.08	0.09 ▶ 0.08
46	0.04 ◀ 0.20	0.04 ◀ 0.17	0.03 ◀ 0.19	0.02 = 0.02	0.01 ◀ 0.61	0.02 = 0.02	0.01 ◀ 0.17	0.01 ◀ 0.17
47	0.18 = 0.18	0.18 = 0.18	0.18 ◀ 0.29	0.49 ▶ 0.29	0.18 = 0.18	0.49 ▶ 0.29	0.18 = 0.18	0.18 = 0.18
48	0.10 ◀ 0.58	0.10 ◀ 0.69	0.10 ◀ 0.58	0.09 ◀ 0.25	0.07 ◀ 0.36	0.09 ◀ 0.25	0.07 ◀ 0.75	0.07 ◀ 0.75
49	0.05 = 0.05	0.05 ▶ 0.04	0.05 ▶ 0.03	0.04 ◀ 0.05	0.04 ◀ 0.05	0.04 ◀ 0.05	0.08 ▶ 0.04	0.08 ▶ 0.04
50	0.06 ◀ 0.12	0.07 ◀ 0.11	0.06 ◀ 0.12	0.14 ▶ 0.12	0.42 ▶ 0.12	0.14 ▶ 0.12	0.05 ◀ 0.12	0.05 ◀ 0.12
51	0.06 ◀ 0.21	0.06 ◀ 0.63	0.08 ◀ 0.56	0.05 ◀ 0.23	0.22 = 0.22	0.05 ◀ 0.23	0.48 ▶ 0.22	0.48 ▶ 0.22
52	0.05 ◀ 0.26	0.08 ◀ 0.24	0.09 ◀ 0.27	0.13 ◀ 0.26	0.05 ◀ 0.26	0.13 ◀ 0.26	0.10 ◀ 0.29	0.10 ◀ 0.27
53	0.02 ◀ 0.25	0.02 ◀ 0.25	0.03 ◀ 0.25	0.03 ◀ 0.25	0.09 ◀ 0.25	0.03 ◀ 0.25	0.09 ◀ 0.25	0.09 ◀ 0.25
54	0.21 ◀ 0.23	0.21 ◀ 0.23	0.21 ◀ 0.23	0.21 ◀ 0.23	0.23 = 0.23	0.21 ◀ 0.23	0.23 = 0.23	0.23 = 0.23
55	0.26 ◀ 0.43	0.26 ◀ 0.43	0.26 ◀ 0.40	0.26 ◀ 0.49	0.26 ◀ 0.40	0.26 ◀ 0.49	0.26 ◀ 0.45	0.26 ◀ 0.45
56	0.58 ▶ 0.52	0.58 ▶ 0.53	0.65 ▶ 0.54	0.50 ◀ 0.54	0.48 ◀ 0.73	0.50 ◀ 0.54	0.53 ◀ 0.69	0.53 ▶ 0.52
57	0.31 ◀ 0.33	0.32 ◀ 0.34	0.33 ▶ 0.32	0.35 ▶ 0.32	0.34 ▶ 0.32	0.35 ▶ 0.32	0.35 ▶ 0.33	0.35 ▶ 0.33
58	0.04 ◀ 0.67	0.04 ◀ 0.45	0.04 ◀ 0.45	0.04 ◀ 0.41	0.04 ◀ 0.44	0.04 ◀ 0.41	0.04 ◀ 0.45	0.04 ◀ 0.45
59	0.14 = 0.14	0.14 = 0.14	0.14 = 0.14	0.14 = 0.14	1.00 ▶ 0.14	0.14 = 0.14	1.00 ▶ 0.70	1.00 = 1.00
60	0.05 ◀ 0.06	0.05 ◀ 0.06	0.05 ◀ 0.06	0.06 = 0.06	0.11 ▶ 0.08	0.06 = 0.06	0.10 ▶ 0.07	0.10 ▶ 0.07
61	0.07 ◀ 0.08	0.07 ◀ 0.08	0.08 = 0.08	0.07 ◀ 0.08	0.11 ▶ 0.08	0.07 ◀ 0.08	0.11 ▶ 0.10	0.11 ▶ 0.10
62	0.11 ▶ 0.09	0.11 ▶ 0.09	0.16 ▶ 0.11	0.22 ▶ 0.17	0.11 = 0.11	0.22 ▶ 0.17	0.23 ▶ 0.11	0.23 ▶ 0.11
#best	15:36 (11)	18:34 (10)	14:33 (15)	15:28 (19)	19:31 (12)	15:28 (19)	21:30 (11)	21:27 (14)
mean	0.26:0.34	0.26:0.32	0.27:0.35	0.27:0.32	0.30:0.35	0.27:0.32	0.30:**0.36**	0.30:**0.36**

Table 3: The recall of the FoT on 62 sequences. For details, see text.

P	$\emptyset$	ρ	N	FB	M	FB ∘ ρ	$N \circ M$	Σ
Seq.	m ◇ r	m ◇ r	m ◇ r	m ◇ r	m ◇ r	m ◇ r	m ◇ r	m ◇ r
1	26 ◀ 21	24 ◀ 22	23 = 23	20 = 20	25 ◀ 18	20 = 20	24 ◀ 21	24 ◀ 21
2	0 ▶ 4	0 ▶ 3	2 ▶ 3	0 ▶ 3	3 ▶ 4	0 ▶ 3	2 ▶ 3	4 = 4
3	21 ◀ 16	20 ◀ 12	15 ◀ 11	13 ◀ 9	14 = 14	13 ◀ 9	17 ◀ 11	17 ◀ 9
4	45 ▶ 50	48 ◀ 44	46 ◀ 44	40 ▶ 48	28 ▶ 39	40 ▶ 48	28 ▶ 42	25 ▶ 37
5	9 ◀ 2	7 ◀ 1	3 ◀ 2	4 ◀ 3	2 ◀ 0	4 ◀ 3	2 ◀ 1	2 = 2
6	1 ◀ 0	1 ◀ 0	1 ◀ 0	1 ◀ 0	1 ◀ 0	1 ◀ 0	1 ◀ 0	1 = 1
7	10 ▶ 14	10 ▶ 15	10 ▶ 14	9 ▶ 14	9 ▶ 14	9 ▶ 14	9 ▶ 14	9 ▶ 14
8	2 = 2	2 = 2	2 = 2	2 = 2	2 = 2	2 = 2	2 = 2	2 = 2
9	13 ▶ 15	14 ▶ 15	18 ◀ 16	17 ◀ 16	7 ▶ 13	17 ◀ 16	9 ▶ 12	7 ▶ 12
10	1 ◀ 0	1 ◀ 0	1 ◀ 0	1 ◀ 0	1 ◀ 0	1 ◀ 0	1 ◀ 0	1 ◀ 0
11	0 = 0	0 = 0	0 = 0	0 = 0	0 = 0	0 = 0	0 = 0	0 = 0
12	23 ◀ 13	22 ◀ 15	18 ◀ 11	13 = 13	19 ◀ 10	13 = 13	13 ◀ 9	13 ◀ 12
13	5 = 5	5 = 5	6 ◀ 5	4 ▶ 6	4 ▶ 5	4 ▶ 6	4 ▶ 5	4 ▶ 6
14	2 = 2	2 ◀ 1	2 ◀ 1	3 ◀ 0	2 = 2	3 ◀ 0	2 = 2	2 ◀ 1
15	5 ◀ 1	5 ◀ 2	5 ◀ 2	6 ◀ 3	6 ◀ 3	6 ◀ 3	7 ◀ 3	7 ◀ 3
16	10 ◀ 8	9 = 9	9 ▶ 10	15 ◀ 9	5 ▶ 7	15 ◀ 9	5 ▶ 7	5 ▶ 7
17	52 ▶ 53	52 ▶ 57	51 ▶ 54	57 ◀ 56	49 ▶ 55	57 ◀ 56	49 ▶ 51	49 ▶ 51
18	21 ◀ 14	20 ◀ 12	16 ◀ 13	25 ◀ 18	17 ◀ 11	25 ◀ 18	14 = 14	14 ◀ 13
19	35 ◀ 26	33 ◀ 26	29 ◀ 26	46 ◀ 35	44 ◀ 21	46 ◀ 35	29 ◀ 26	29 ◀ 24
20	8 ◀ 2	8 ◀ 2	6 ◀ 2	9 ◀ 3	3 ◀ 2	9 ◀ 3	2 ◀ 0	2 ◀ 0
21	2 ◀ 1	2 = 2	2 ◀ 1	1 = 1	2 ◀ 1	1 = 1	2 ◀ 1	2 ◀ 1
22	95 ◀ 91	95 ◀ 91	90 ◀ 89	91 ▶ 95	93 = 93	91 ▶ 95	90 ▶ 91	90 ◀ 89
23	11 ◀ 4	9 ◀ 4	5 ◀ 2	10 ◀ 7	10 ◀ 1	10 ◀ 7	12 ◀ 11	12 ◀ 11
24	4 = 4	5 ◀ 4	5 ◀ 4	3 = 3	4 = 4	3 = 3	3 ▶ 4	3 ▶ 4
25	36 ◀ 26	35 ◀ 22	17 ▶ 19	7 ▶ 8	19 ◀ 17	7 ▶ 8	6 ▶ 13	6 ▶ 9
26	14 ▶ 17	11 ▶ 16	6 ▶ 19	15 ▶ 18	13 = 13	15 ▶ 18	15 ▶ 18	15 ▶ 19
27	2 = 2	2 = 2	1 ▶ 2	1 ▶ 2	3 ◀ 1	1 ▶ 2	1 ▶ 2	1 ▶ 2
28	6 ◀ 2	6 ◀ 2	5 ◀ 2	8 ◀ 2	3 ◀ 2	8 ◀ 2	4 ◀ 2	3 ◀ 2
29	22 ◀ 18	23 ◀ 16	19 = 19	28 ◀ 24	24 ◀ 18	28 ◀ 24	21 ◀ 18	21 ◀ 19
30	14 ▶ 15	10 ▶ 16	13 ▶ 14	21 ◀ 16	11 = 11	21 ◀ 16	6 ▶ 14	5 ▶ 13
31	1 ◀ 0	1 = 1	0 ▶ 1	0 ▶ 1	2 ◀ 1	0 ▶ 1	2 ◀ 0	2 ◀ 0
32	220 ◀ 83	210 ◀ 79	110 ◀ 76	77 ◀ 70	193 ◀ 80	77 ◀ 71	107 ◀ 65	103 ◀ 69
33	5 ◀ 1	3 ◀ 1	2 ◀ 1	4 ◀ 2	6 ◀ 0	4 ◀ 2	2 ◀ 0	2 ◀ 1
34	12 ◀ 9	12 ▶ 13	9 ▶ 10	68 ◀ 63	9 ▶ 10	68 ◀ 63	10 ▶ 13	11 = 11
35	59 ◀ 27	56 ◀ 29	45 ◀ 33	56 ◀ 50	66 ◀ 32	56 ◀ 50	55 ◀ 36	55 ◀ 34
36	67 ▶ 76	68 ▶ 77	69 ▶ 79	66 ▶ 78	68 ▶ 76	66 ▶ 78	63 ▶ 73	63 ▶ 76
37	13 ◀ 2	13 ◀ 3	8 ◀ 2	5 = 5	10 ◀ 1	5 = 5	10 ◀ 1	8 ◀ 1
38	3 ▶ 4	3 ▶ 4	4 ◀ 3	2 = 2	0 ▶ 4	2 = 2	0 ▶ 3	0 ▶ 3
39	0 = 0	0 = 0	0 = 0	0 = 0	1 ◀ 0	0 = 0	1 ◀ 0	1 ◀ 0
40	26 ◀ 10	23 ◀ 11	18 ◀ 9	16 ◀ 8	20 ◀ 10	16 ◀ 8	20 ◀ 7	15 ◀ 10
41	21 ▶ 22	21 = 21	22 = 22	24 = 24	18 ▶ 22	24 = 24	21 ▶ 22	21 ▶ 23
42	13 ▶ 17	14 = 14	13 ▶ 16	15 ▶ 18	9 ▶ 14	15 ▶ 18	12 ▶ 14	10 ▶ 14
43	10 ◀ 9	10 ▶ 11	9 ▶ 11	10 ▶ 12	8 ◀ 7	10 ▶ 12	7 = 7	7 ▶ 10
44	2 = 2	2 = 2	2 = 2	2 ▶ 3	3 ◀ 2	2 ▶ 3	3 = 3	3 = 3
45	53 ◀ 46	53 ◀ 48	42 ▶ 44	52 ◀ 50	33 ▶ 35	52 ◀ 50	32 ◀ 29	32 ◀ 29
46	7 ◀ 3	7 ◀ 3	5 ◀ 3	7 = 7	8 ◀ 3	7 = 7	6 ◀ 3	8 ◀ 4
47	7 ◀ 4	7 ◀ 3	6 ◀ 4	8 ◀ 4	8 ◀ 4	8 ◀ 4	7 ◀ 4	7 ◀ 4
48	3 ▶ 6	3 = 3	8 ◀ 7	10 ◀ 7	7 ▶ 8	10 ◀ 7	8 ◀ 2	8 ◀ 2
49	34 ▶ 37	35 ◀ 34	32 ▶ 37	37 ▶ 38	26 ◀ 22	37 ▶ 38	17 ▶ 24	17 ▶ 23
50	26 = 26	28 = 28	27 = 27	34 ◀ 28	5 ▶ 13	34 ◀ 28	8 ▶ 11	17 ◀ 13
51	8 ◀ 5	6 ◀ 5	6 ◀ 5	13 ◀ 3	10 ◀ 5	13 ◀ 3	12 ◀ 4	12 ◀ 4
52	15 ◀ 9	14 ◀ 11	10 = 10	15 ◀ 9	18 ◀ 9	15 ◀ 9	17 ◀ 10	18 ◀ 11
53	32 ◀ 9	34 ◀ 8	23 ◀ 9	37 ◀ 16	41 ◀ 11	37 ◀ 16	33 ◀ 13	33 ◀ 14
54	11 ◀ 5	11 ◀ 5	11 ◀ 5	18 ◀ 11	12 ◀ 5	18 ◀ 11	10 ◀ 9	10 ◀ 8
55	5 ◀ 4	5 ◀ 4	5 ◀ 4	5 ◀ 4	5 ◀ 4	5 ◀ 4	5 ◀ 4	5 ◀ 4
56	10 ◀ 4	10 ◀ 3	7 ◀ 3	8 ◀ 3	8 ◀ 4	8 ◀ 3	9 ◀ 4	9 ◀ 5
57	8 ◀ 6	7 ◀ 5	4 ◀ 3	6 ◀ 4	6 ◀ 3	6 ◀ 4	12 ◀ 5	10 ◀ 3
58	14 ◀ 1	13 ◀ 2	6 ◀ 3	4 ◀ 3	13 ◀ 2	4 ◀ 3	5 ◀ 2	7 ◀ 4
59	5 ◀ 4	4 = 4	3 ▶ 4	4 = 4	0 ▶ 4	4 = 4	0 ▶ 1	0 = 0
60	8 ◀ 7	8 = 8	8 ▶ 9	9 = 9	7 ▶ 8	9 = 9	6 ▶ 9	6 ▶ 9
61	34 ◀ 30	31 ▶ 32	20 ▶ 29	43 = 43	40 ◀ 21	43 = 43	31 ◀ 23	36 ◀ 25
62	37 ◀ 25	33 ◀ 24	19 ▶ 25	48 ◀ 46	32 ◀ 31	48 ◀ 46	28 = 28	28 ▶ 30
#best	13:40 (9)	11:36 (15)	19:34 (9)	14:34 (14)	17:37 (8)	14:34 (14)	22:33 (7)	19:35 (8)
mean	20.4:14.9	19.8:14.7	15.8:14.6	18.9:17.1	18.0:13.4	18.9:17.1	15.1:**13.3**	15.1:13.5

Table 4: The number of reinitialisations of the FoT on 62 sequences. For details, see text.

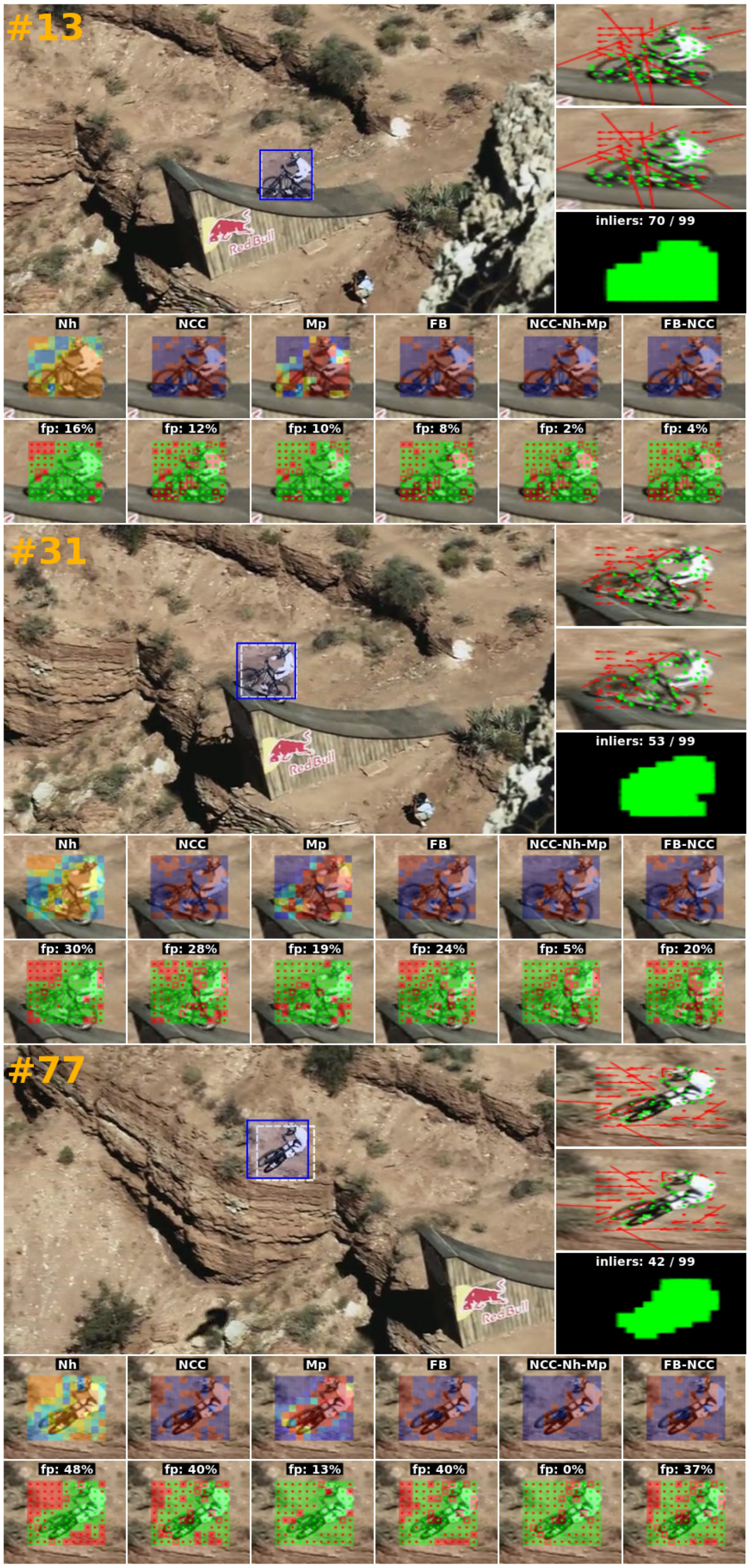


Fig. 10: Visualization of predictors performance on sequence *mountain-bike*. For details, see text.

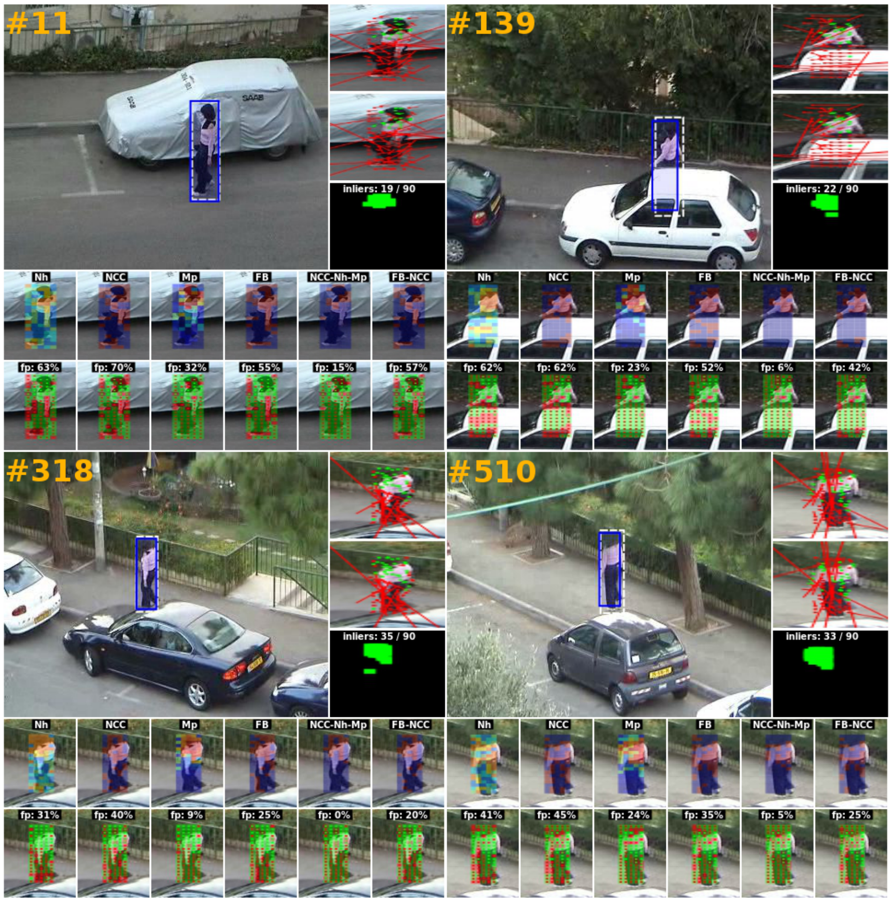


Fig. 11: Visualization of predictors performance on sequence *woman*. For details, see text.

P / Seq.	$\emptyset$	ρ	FB	FB $\circ$ ρ	$N \circ M$	Σ
	m ◇ r	m ◇ r	m ◇ r	m ◇ r	m ◇ r	m ◇ r
Time [ms]	1.53 ▶ 1.55	2.44 ▶ 2.87	2.52 ▶ 2.89	3.43 ▶ 3.58	1.58 ▶ 1.72	2.43 ▶ 2.52

Table 5: A comparison of the speed of tracking reliability prediction methods. All times are in milliseconds. The values are averaged over all sequences.

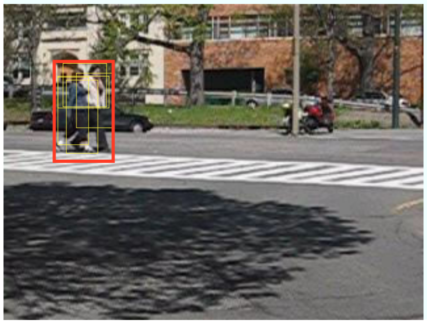
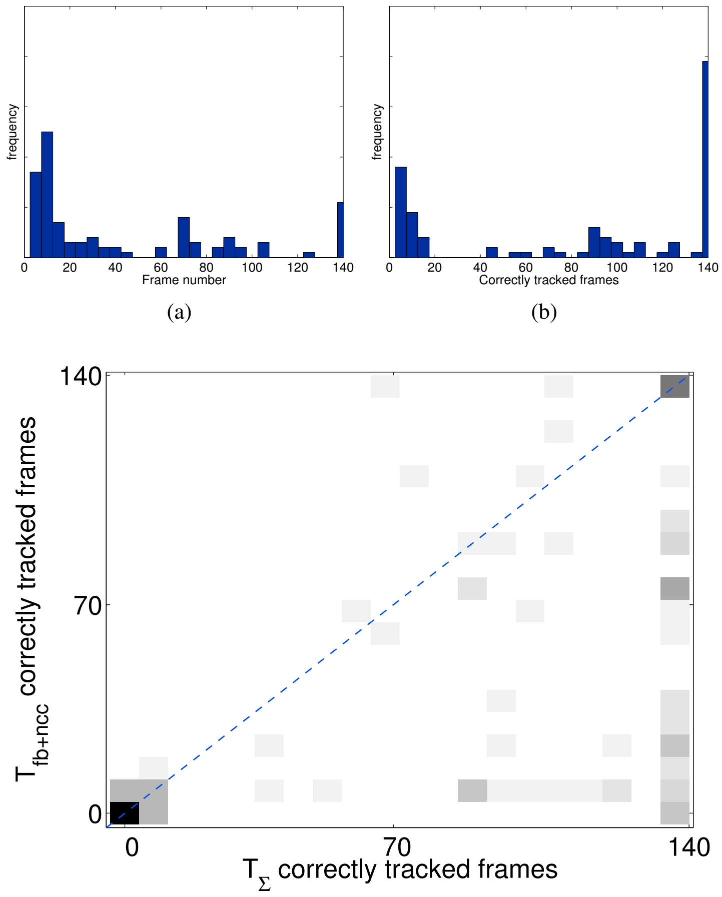


Fig. 12: Examples of randomly generated initial bounding boxes (yellow) randomly generated within the red rectangle.

Method	Score	mean (median)
$P_{FB\circ\rho}$ [ref]	4493	45 (21)
P_{Σ}	8438	84.4 (99.5)

Table 6: Evaluation of filtering methods in terms of the number of correctly tracked frames with randomly initialized bounding box (see. Fig. 12). The “score” is the total number of correctly tracked frames, the mean and the median of the same quantity are presented in the right column.



(a)

(b)

(c)

Fig. 13: Histograms of the number of correctly tracked frames for tracker with (a) $P_{FB \circ \rho}$ and (b) P_{Σ}. (c) The 2D histogram of the number of correctly tracked frames by $P_{FB \circ \rho}$ and P_{Σ} initialized with the same random bounding box.