# Robust Hand Tracking with Hough Forest and Multi-cue Flocks of Features

Hong Liu, Wenhuan Cui, Runwei Ding

Key Laboratory of Machine Perception and Intelligence,
Shenzhen Graduate School, Peking University, P.R.China.
hongliu@pku.edu.cn cwh@sz.pku.edu.cn dingrunwei@pkusz.edu.cn

**Abstract.** Robust hand tracking is highly demanded for many real-world applications relevant to human machine interface, however, current methods achieves no satisfactory robustness in real environments. In this paper a novel hand tracking method was proposed integrating online Hough Forest and Flocks-of-Features tracking. Skin color was integrated in the Hough Forest framework, gaining more robustness against drastic hand appearance and pose changes, eapecially against partial occlusions. Also a novel multi-cue Flocks-of-Features tracking algorithm based on computer graphics was integrated in to enhance the framework's robustness against distractors and background clutter. Additionally, recovery from tracking failure was addressed. Experiments were carried out to compare our method with CAMShift, Hough Forest tracker, and the original Flocks-of-Features Tracker, and showed the effectiveness of our method.

## 1 Introduction

Markerless vision-based hand tracking has been under research for decades, because of its great potential for natural human computer interface, such as robotics, intelligent surveillance, and so on. However, due to the complicated shape and appearance changes of human hand in motion, and the change of environment, there is not yet one single method that can give satisfactory robustness.

Hand tracking methods are usually classified as model-based and appearance-based [1]. Hand models are most often used for 3D hand tracking, as in [2] and [3]. On the other hand, appearance-based methods extract image features as appearance. Skin color feature is often chosen for its simplicity and discriminability. It was used in CAMShift tracker for fast face tracking [4], and for simple hand tracking. Contour feature was used to represent the shape of hand in a particle-filter framework to track hand deformation against clutter [5]. Maximally Stable Extremal Regions (MSER) is a new *stable* region feature and was used for hand tracking [6]. Combination of these features can increase a hand tracker's robustness. However, difficult real-world situations, such as occlusion, background clutter, drastic environment change cannot be easily overcome with these methods. Especially a robust hand tracker should be able to recover tracking failure through re-detection. Off-line-trained tree classifier was used in [7] and combined with interpolation for 3D hand tracking. In [8], clustering based on shape context and boosted tree classifier were used for hand detection. However, to build a classifier off-line is a great challenge due to hand's huge space of possible shape and appearance variation. On the other hand, classifier built online can learn more of the appearance and shape space. Online-learning methods are gaining increasing interest in recent years, and has demonstrated great power in rigid object tracking-by-detection [9]. Especially, part-based online-learning methods were used for learning representations for articulated objects [10][11]. In [10] an Implicit Shape Model was proposed to integrate votings from patches of object parts in a *Generalized Hough Transform* manner, and the patches were clustered by their similarity of appearance. In [11] random forest was used to cluster these patches and votings. And in [12] the Hough forest tracking framework is extended in an on-line fashion, with additional steps of backprojection and segmentation, bringing greater adaptivity to drastic pose and shape changes.

This work was built upon the Hough forest tracking-by-detection framework described in [12]. The *stability-plasticity dilemma* [13] of the approach in [12] was tackled by integrating the Hough forest classifier with a modified multi-cue Flocks-of-Features tracker [14]. The modification is based on a computer graphics technique to simulate flock behavior [16]. The resulting framework showed good robustness against difficult situations such as clutter, distraction, occlusion and tracking failure. A system overview of our approach is shown in Fig. 1.

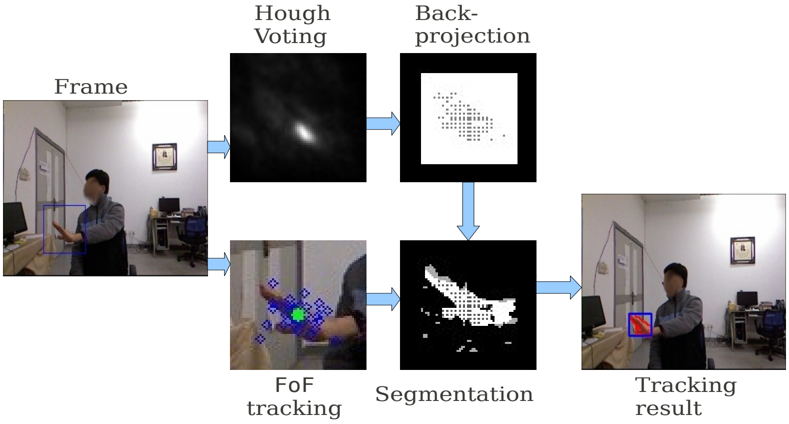

**Fig. 1.** Overview of our approach

The rest of the paper is organized as follows. Section 2 gives a brief introduction to the Hough forest tracking framework, and shows its limitations. Section 3 describes in detail the Flocks-of-Features tracker and its combination with the Hough forest tracking. Experiments and discussions are given in section 4. Finally conclusions and future directions are given in section 5.

# 2 Hough Forest for Hand Tracking

## 2.1 Hough Forest: Hough Voting and Random Forest

The Implicit Shape Model (ISM) is a part-based model of appearance and configuration of constituting parts of an object [10]. The Hough forest learns the appearance and their configurations by random forest [11]. An illustration of the working scheme of Hough forest detection is given in Fig. 2. Patches around densely sampled pixels and their offset from the hand center are fed to the random forest for training and testing. Details are given in the following.

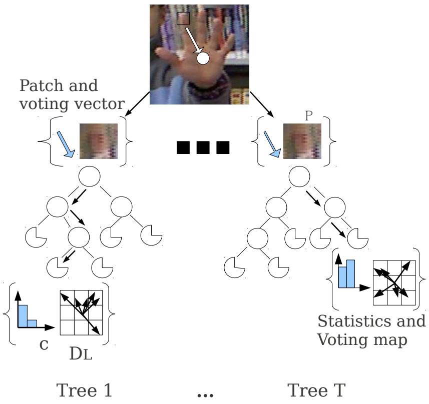

**Fig. 2.** Hough forest detection

Let the class label be $c$, and $c \in \{1 = \text{Object}, 0 = \text{Background}\}$, and image patch be $\mathcal{P}$. With densely sampled $N$ image patches $\{\mathcal{P}_i = (\mathcal{I}_i, c_i, \mathbf{d}_i)\}_{i=1}^{N}$, where $\mathbf{d}_i$ is the offset vector, each tree $\mathcal{T}$ is trained on $M(< N)$ samples randomly selected from the $N$ samples. Each node $j$ in the tree $\mathcal{T}$ is assigned a binary test $t(\mathcal{I}) \to \{1, 2\}$, defined on a patch's appearance $\mathcal{I}_i$. Random locations, $(p, q)$ and $(r, s)$, are chosen for comparison, and of all $C$ feature channels, a random channel $a \in \{1, 2, ..., C\}$ is chosen. The test at node $j$ is defined as:

$$t_j = \begin{cases} 0, \text{ if } I^a(p,q) < I^a(r,s) + \theta \\ 1, \text{ otherwise} \end{cases} \tag{1}$$

where $\theta$ is a threshold value chosen based on some optimization criteria. In Hough forest, the patch class statistics and voting map are both stored in leaves, forming a voting map $D_L = \{\mathbf{d}_i\}$. In such a way an implicit discriminative *codebook* is constructed at each leaf. All votes from each tree are collected into the Hough space, and the maximum location is determined as the object's center.

### 2.2 Hough Forest for Tracking: Increasing Stability and Adaptivity

In [12], the Hough forest was formulated in an online learning manner for tracking non-rigid objects. The Grabcut Segmentation algorithm [15] was used to generate ample training samples online. The seeds were generated by backprojection/support of the Hough voting. The support $S(\mathbf{m}, \rho)$ of a specific point $\mathbf{m}$ in Hough space is the pixel locations that give votings to points within a range of radius $\rho$ to $\mathbf{m}$. Furthermore, an forgetting scheme was used in [12] to reduce the weights of formerly learned samples and increase the online learner's adaptivity, which was applied on both statistics and voting map in the leaves. Skin color was used in our work to refine the segmentation, which further increase the accuracy of classification.

## 3 Hough Forest Tracking with Flocks of Features

### 3.1 Tracking Flocks of Features

The Flocks-of-Features (FoF) tracker first proposed in [14] exploited a biological phenomenon, the *Flock Behavior*, which states that the members $p_i$ in the flock $\mathcal{F} = \{\{p_i\}_{i=1}^{N_f}\}$ should satisfy:

$$\begin{aligned} & d_{\text{min}} < |p_i - p_j|, \forall i, j \in \{1, 2, ..., N_f\}, \text{and} \\ & d_{\text{max}} > |p_i - m|, \forall i \in \{1, 2, ..., N_f\}. \\ & m = \text{median}(\mathcal{F}) \text{ or centroid}(\mathcal{F}). \end{aligned} \tag{2}$$

where $|p_i - p_j|$ is the distance between $p_i$ and $p_j$. The $d_{\text{min}}$ and $d_{\text{max}}$ are maximum and minimum distances which are parameters the flock holds. Unlike [14], in our work techniques in computer graphics for simulating flocking behaviors are considered. We chose the well-known *Boids* algorithm [16], which is widely used to simulate flock behavior in computer graphics. Call the members of the flock as *boids*, the flocking behavior is maintained by the following rules:

**Separation** boids try to keep a distance away from other boids.
**Cohesion** boids try to fly towards the center of mass of neighbours.
**Allignment** boids try to match velocity with near boids.

Based on the *Boids* algorithm, we proposed a new realization of FoF tracker, which showed better *flock behavior* than that of [14]. The algorithm is summarised in Alg. 1.

The key steps are KLT optical flow tracking [17] of feature points and a modified Boids algorithm. Major modification is the use of a confidence map to weight the feature points. The intuition behind it is that, for the flock to track a target, members near to the target should have more importance. In our current implementation, only skin color was used to build the confidence map, which is the backprojection of hue histogram of the hand, similar to the CAMShift tracking [4]. Better confidence map is possible.

**Algorithm 1** The modified Flocks-of-Features tracker

```
INPUT: Current search window $B$, last feature points $\mathcal{F}_{pre}$, image frame $I$, confidence map W
OUTPUT: New feature points $\mathcal{F}$
BEGIN:
$\mathcal{F}_t \leftarrow \{\mathcal{F}_{\text{tracked}}, \mathcal{F}_{\text{lost}}\} \leftarrow \text{KLT}(\mathcal{F}_{pre}, I)$
$\mathcal{F}_{\text{lost}} \leftarrow \text{getNewPoints}(\text{W}, \text{B})$
for $i = 1 \rightarrow N_f$ do
    $S_p \leftarrow \sum_{i=1}^{N_f} \text{W}(\mathcal{F}_{pre}(i)) * \mathcal{F}_{pre}(i)$
    "Perceived center": $s_i \leftarrow \frac{S_p - \mathcal{F}_{pre}(i)}{N_f - 1}$
    "Positive Driving Force": $f_p \leftarrow s_i - \mathcal{F}_{pre}(i)$
    "Negagive Driving Force": $f_n \leftarrow 0$
    for $j = 1 \rightarrow N_f$ do
        if $j \neq i$ AND $||\mathcal{F}_{pre}(i) - \mathcal{F}_{pre}(j)||_2 \leq d_{\text{min}}$ then
            $f_n \leftarrow f_n + d$
        end if
    end for
    $\mathcal{F}_t(i) \leftarrow \mathcal{F}_{pre}(i) + \alpha * f_p + \beta * f_n$
end for
END
```

### 3.2 Integration of FoF Tracker with Hough Forest Framework

Sole tracking-by-detection has the problem of stability-plasticity dilemma, which can be alleviated by integrating different trackers, as shown in [18] and [19]. In the FROST framework [18], trackers occupying different stability-adaptivity scope were combined to compensate each other. Alternatively, in [19] a self-verifying tracker was devised to supervise the on-line *P-N learning.* These integration methods have shown impressive results on rigid object tracking, however, poor results for tracking human hand were observed in our evaluations. It was observed that the low performance of the online classifier and the sub-tracker on articulated objects contributed to the poor result.

In this work the Hough forest classifier was chosen to be integrated with FoF tracker, both of which have shown good tracking results on articulated objects. The interactions between relevant modules are shown in Fig. 3.

The *tracking validity* is an essential value for integration, because supervision of learning is based on it. For FoF tracking, since it is KLT-based, its tracking validity can be staightforwardly defined as:

$$V_f = \frac{N_t}{N_a} \tag{3}$$

where $N_t$ is the number of feature points that are confidently tracked, and $N_a$ is the number of all feature points. The final tracking result was given by the Hough forest tracker, while the update of the Hough forest was influenced by the FoF tracker. When the validity of FoF tracking $V_f$ is beyond a threshold ( 0.3 in our implementation), and the bounding box of the feature points is large

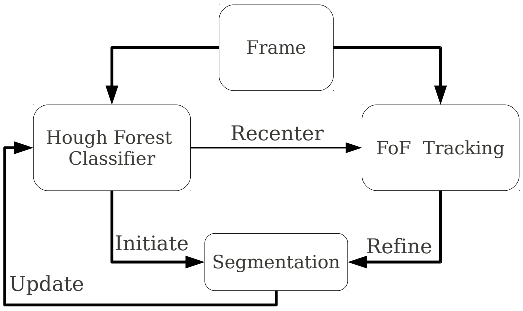

**Fig. 3.** Integration of Hough Forest classifier and FoF tracker

enough, the Grabcut segmentation result was pruned using this bounding box to eliminate highly-possible distracting patches.

## 4 Experiments and Discussions

The proposed approach was tested on six real-world sequences with more than 3500 frames taken by an ordinary RGB camera, whose resolution is $640 \times 480$ in pixel. No assumptions of still camera or still body were made. In all sequences illumination change was unconstrained, and face was a constant distractor. An overall performance of the proposed approach is shown in Table. 1. Each of the sequences has specific difficulties designed to test different aspects of tracking robustness, depicted in second column of Table. 1. The performance was measured with *recall* and *precision*, whose definition were:

$$\begin{aligned} \text{recall} &= \frac{\text{Num. of True Positives}}{\text{Num. of True Positives} + \text{Num. of False Negatives}} \\ \text{precision} &= \frac{\text{Num. of True Positives}}{\text{Num. of True Positives} + \text{Num. of False Positives}} \end{aligned} \tag{4}$$

where True Positives are tracked targets being real hand, False Positives are wrongly tracked target, and False Negatives are tracking failure report when the hand is in the image.

Most of the recalls in the tests are 100 percent because the hand is always in the scene and tracking failure never happened, so there is no false negatives. In sequence one and three, most false tracking were caused by face' distraction, which can become severe when hand changed pose quickly in front of face. The tracking could be recovered if the hand stayed ahead of face for a while to let the classifier learn the hand's appearance and forget that of face. In sequence two and five, as re-detection rate was low, a low recall as 79 percent was obtained for many false negatives were produced. This revealed the lack of stability of the Hough forest part in the framework. In the last two sequences, face detection technique was used to counteract the distraction from face. By removing the face' distraction, much tracking errors were avoided, giving a higher precision. Although the online Hough forest classifier has no long-range memory of the hand's appearance, the short-term memory can guide the tracker toward parts of the hand. Therefore, tracking failure caused by out-of-view motion could still be recovered after some time, as shown in sequence five.

**Table 1.** Hand tracking on our sequences: Recal and Precision

| Sequence | Difficulties | Total Frames | Correctly Tracked | Recall | Precision |
|---|---|---|---|---|---|
| Seq.1 | face distraction | 1006 | 805 | 100% | 80% |
| Seq.2 | background clutter | 310 | 295 | 100% | 95% |
| Seq.3 | large motion | 706 | 693 | **100%** | **98%** |
| Seq.4 | occlusion, distraction | 527 | 384 | 99% | 73% |
| Seq.5 | out-of-view motion | 413 | 249 | **79%** | **87%** |
| Seq.6 | 3D changing postures | 763 | 697 | 100% | 91% |

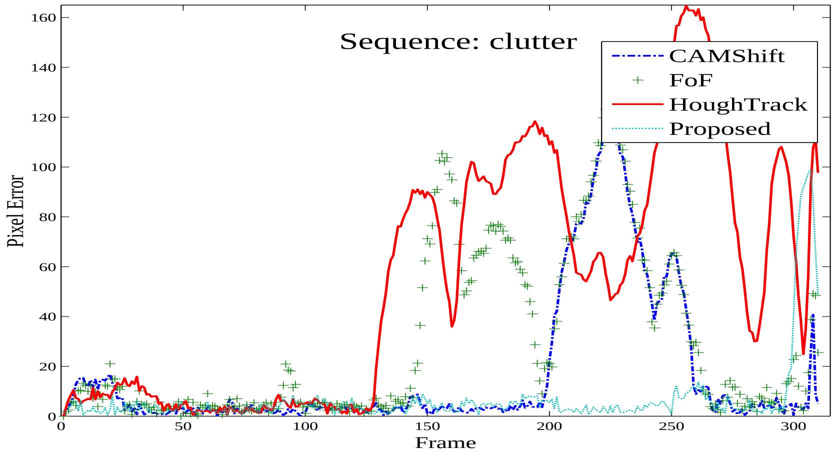

**Fig. 4.** Tracking error of CAMShift, FoF, HoughTrack and proposed, on sequence 2

The proposed approach was also compared with CAMShift, FoF tracker, HoughTrack on sequence two, because it has much clutter along with face' distraction, a good testbench for all methods. The pixel error between the center of each tracked bounding box and the ground truth as a function of frame number is plotted in Fig. 4. It can be seen that the proposed method gave the longest stable tracking with 300 frames, untill the person was leaving to end the sequence. On the contrary, the HoughTrack behaves as the least stable tracker in the test. After less than 150 frames, it got locked on the background bookshelf and never recovered. This can be attributed to the fact that textured objects are easier to track since they have more stable feature points for Hough forest to learn. And once the stable appearance of the background was learned, it could hardly be forgotten. The FoF tracker also showed much fluctuation, which were caused by the distraction of face and the bookshelf. It suffered the same problem as that of Hough forest tracker. On the other hand, the CAMShift tracker is based on skin color, thus is robust against environment clutterness, and it kept stable tracking for nearly 200 frames. However, illumination change and other skin color objects can easily attract CAMShift tracker away, as happened around frame 200. But CAMShift resumed tracking when the hand moved over face a later time since it distinguishs no face and hand.

For hand tracking, the robustness against human face' distraction, occlusion, and tracking failure is most challenging and important. Our method displayed good performance against these difficulties mainly because it's a part-based approach with tracking failure recovery. Fig. 5 shows an example of occlusion. Interestingly the track drifts away with another hand after full occlusion, but get recovered when the false track is lost. In Fig. 6, the ISM with FoF tracking tracking alleviates the face' distraction. Fig. 7 shows the recovery of a tracking failure caused by out-of-view motion.

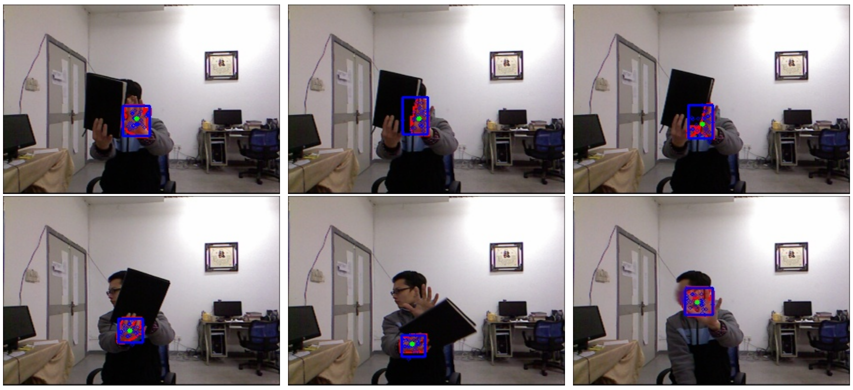

**Fig. 5.** Robustness against occlusion, frames: 60, 61, 62, 63, 69, 88

## 5 Conclusions

This paper proposed a robust hand tracking method based on a Hough forest tracking framework integrated with a Flocks-of-Features tracker. The improved realization of Flocks-of-Features tracker was proposed based on computer graphics techniques. The proposed part-based framework has demonstrated great robustness against partial occlusion, distraction, background clutter, and recovery

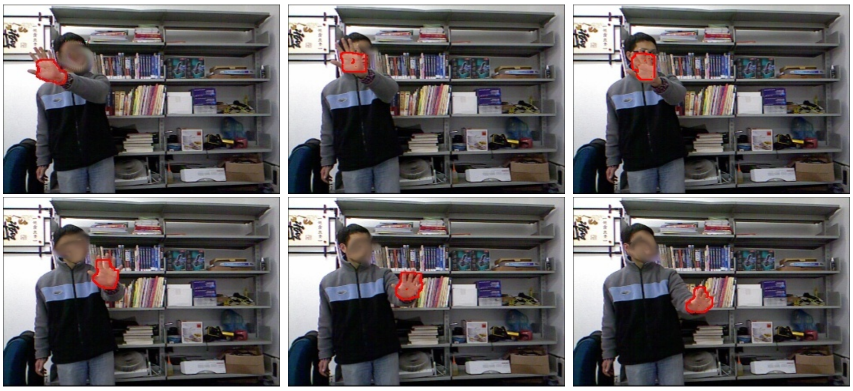

**Fig. 6.** Robustness against face' distraction, frames: 261, 264, 268, 272, 275, 278

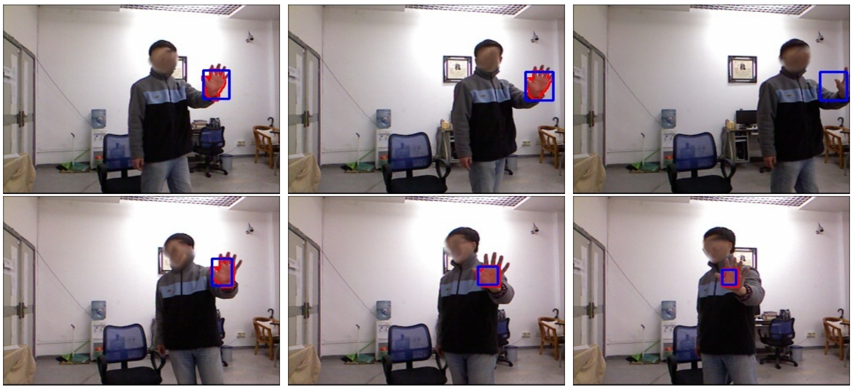

**Fig. 7.** Recovery from tracking failure caused by out-of-view motion, frames: 128, 137, 144, 183, 192, 201

from tracking failure, yet does not rely on still camera or static body. Experiments also revealed the lack of stability of the Hough forest part. Future work would explore temporal information for more meaningful Grabcut segmentation, and generating more sophisticated behavior of flocks of features to increase the stability.

## Acknowledgements

This work is supported by National Natural Science Foundation of China(NSFC, No.60875050, 60675025), National High Technology Research and Development Program of China(863 Program, No.2006AA04Z247), Shenzhen Scientific and Technological Plan and Basic Research program (No.JC200903160369A), Natural Science Foundation of Guangdong (No.9151806001000025).

## References

1. Erol, A., Bebis, G., Nicolescu, M., Boyle, R.D., Twombly, X.: Vision-based hand pose estimation: A review. Computer Vision and Image Understanding **108** (2007) 52–73
2. Rehg, J., Kanade, T.: Digiteyes: Vision-based hand tracking for human-computer interaction. In: Proceedings of the workshop on Motion of Non-Rigid and Articulated Bodies. (1994) 16–22
3. Stenger, B., Thayananthan, A., Torr, P.H.S., Cipolla, R.: Model-based hand tracking using a hierarchical bayesian filter. IEEE Transactions on Pattern Analysis and Machine Intelligence **28** (2006) 1372–1384
4. Bradski, G.R.: Computer vision face tracking for use in a perceptual user interface. Intel Technology Journal (1998)
5. Isard, M., Blake, A.: CONDENSATION - conditional density propagation for visual tracking. International Journal of Computer Vision **29** (1998) 5–28
6. Donoser, M., Bischof, H.: Real time appearance based hand tracking. In: Proceedings of the International Conference on Pattern Recognition. (2008) 1–4
7. Tomasi, C., Petrov, S., Sastry, A.: 3d tracking = classification + interpolation. In: Proceedings of the International Conference on Computer Vision. Volume 2. (2003) 1441–1448
8. Ong, E.J., Bowden, R.: A boosted classifier tree for hand shape detection. In: Proceedings of International Conference on Automatic Face and Gesture Recognition. (2004) 889 – 894
9. Avidan, S.: Ensemble tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **29** (2007) 261 –271
10. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: Proceedings of the Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic (2004)
11. Gall, J., Lempitsky, V.: Class-specific hough forests for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. (2009) 1022–1029
12. Godec, M., Roth, P.M., Bischof, H.: Hough-based tracking of non-rigid objects. In: International Conference on Computer Vision. (2011) 81–88
13. Grossberg, S.: Competitive learning: From interactive activation to adaptive resonance. Cognitive Science **11** (1987) 23–63
14. Kolsch, M., Turk, M.: Fast 2D hand tracking with flocks of features and Multi-Cue integration. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop. (2005)
15. Rother, C., Kolmogorov, V., Blake, A.: ”grabcut”: interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. **23** (2006) 309–314
16. Reynolds, C.W.: Flocks, herds and schools: A distributed behavioral model. SIGGRAPH Comput. Graph. **21** (1987) 25–34
17. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 1981 DARPA Image Understanding Workshop. (1981) 121–130
18. Santner, J., Leistner, C., Saffari, A., Pock, T., Bischof, H.: Prost parallel robust online simple tracking. In: IEEE Conference on Computer Vision and Pattern Recognition. (2010) 720–730
19. Kalal, Z., Matas, J., Mikolajczyk, K.: P-n learning: Bootstrapping binary classifiers by structural constraints. In: IEEE Conference on Computer Vision and Pattern Recognition. (2010)