ROBUST HAND TRACKING BASED ON ONLINE LEARNING AND MULTI-CUE FLOCKS OF FEATURES

Hong Liu, Xing Liu

Engineering Lab on Intelligent Perception for Internet of Things(ELIP), Shenzhen Graduate School,
Key Laboratory of Machine Perception(Ministry of Education), Peking University, China
Email: hongliu@pku.edu.cn, liuxing@sz.pku.edu.cn (Corresponding Author)

ABSTRACT

Robust hand tracking is in increasing demand from areas such as natural Human Robot/Computer interaction (HRI/HCI) and surveillance systems, while it is still a great challenge due to human hand's drastic appearance change. In recent years, online learning techniques have shown great potential in learning appearance of objects and tackling occlusion. This paper extends an online learning framework called Tracking-Learning-Detection (TLD) to track human hand. The main extensions are: 1) original tracker is replaced by a hybrid multi-cue base tracker combining Median-Flow tracker and Flock-of-Features (FoF) tracker, 2) skin color cue is integrated into cascade detector and PN online learning for more efficiency. Extensive experiments show that the new framework works with more robustness compared with state-of-the-art hand trackers and also original TLD.

Index Terms— Hand Tracking; FoF; Multi-cue; Skin Color; Online Learning

1. INTRODUCTION

Bare hands are probably the most natural tools for interactions for its dexterity and hand tracking has shown great potential in applications for sign language [1], gesture recognition [2] and HCI based applications [3]. However robust hand tracking is difficult because of the fact that human hand is an articulated object with complicated shape and appearance change, which makes many excellent methods for tracking rigid objects not easily applicable and robust.

In the past decades plenty of methods have been developed to tackle the hand tracking problems. The methods can generally be grouped into model-based and appearance-based ones [4, 5, 6]. Model-based methods try to build the mapping from model configuration space to image feature space, transforming the task into searching in a higher dimensional space. Full or partial Degrees-of-Freedom hand model can be built to track hand articulations with a rather high accuracy [7, 8]. Reviews can be found in [5]. On the other hand, appearance based methods try to build the mapping from image feature space to the hand appearance space. Skin color feature is often chosen for its simplicity [9]. Contour features have been used in particle filter based methods [10]. Maximally Stable Extremal Regions (MSER) is a new stable region feature that has been used for hand tracking [11]. These methods are robust to appearance change to a certain degree, but cannot resolve or recover from total occlusion and tracking failures, where hand detection should take over.

Recently tracking by detection method has become popular [12]. These systems usually perform one-shot learning of the detector at the first frame. Surrounding patches are labeled negative [13]. The detectors are usually classifiers trained off-line, and the training process is laborious and inefficient [14, 15], what's more the detectors are not adaptive to the change of non-rigid object. An online learning framework would be more desirable, and should be feasible, as demonstrated by researches in rigid object tracking [13]. However, the problem with online learning methods is that the updating process will also drift without a monitor.

To our best knowledge, little work has been done to apply online learning methods to tracking articulated objects. In [16], a Tracking-Learning-Detection (TLD) framework is built and successfully learns a multi-view face appearance with large appearance change. And the online learning component is monitored by a so-called P-N experts to limit errors caused by drift. However, when applied to hand tracking, it is observed that the low performance of the online classifier and the sub-tracker on articulated objects contributed to the poor result in our evaluations.

In this paper, based on analysis of the structure of the framework, a new hybrid base tracker is proposed to improve the performance of the original one, color cue is integrated to improve cascade detection and added for online learning. The extended framework shows high robustness against background clutter, face distracter and could recover from full occlusion and tracking failures.

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),Scientific and Technical Innovation Commission of Shenzhen Municipality (No.JCYJ20120614152234873, CXC201104210010A).

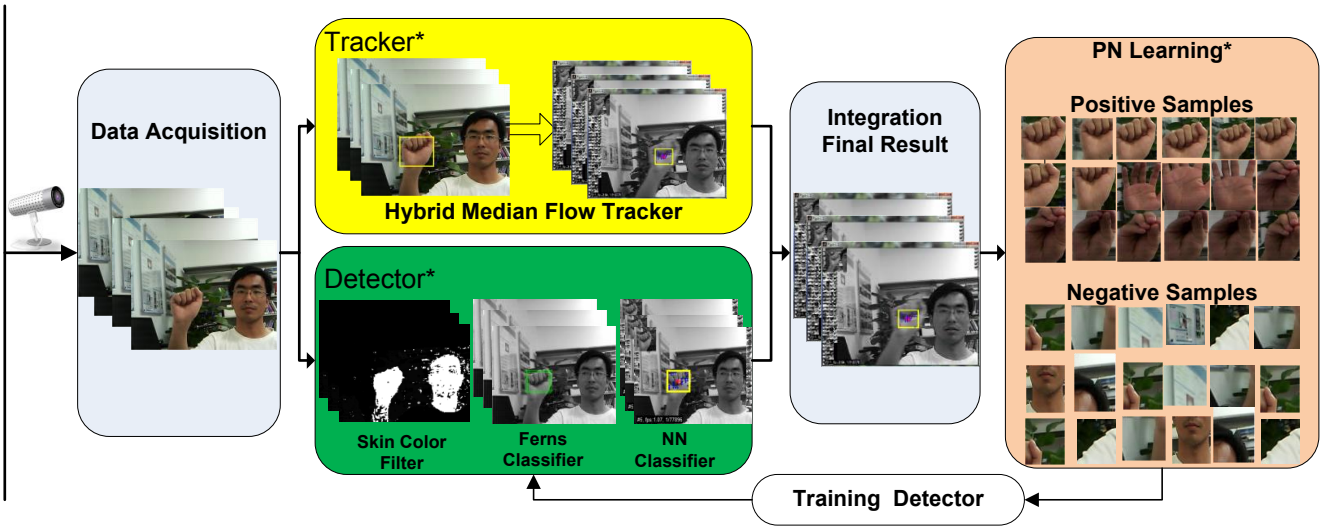

Fig. 1. Work-flow of the hand tracking framework. Tracker*, Detector* and PN Learning* are extended for hand tracking.

2. EXTENDED TLD FOR HAND TRACKING

2.1. Framework of TLD

TLD is a new proposed framework designed for long-term tracking of unknown objects in consecutive frames [17]. The TLD framework combines tracking, online learning and detection based on monitoring tracking errors. Its direct target of online learning is to build an adaptive object appearance model which is a collection of positive and negative example patches. But when it is applied to hand tracking, poor result is observed, as tracker, detector and PN online learning are not designed to track articulated object like hand. Another observation is that the original TLD is a single-cue system. For hand tracking, skin color cue is usually a indispensable information for tracking robustness. New framework tuned for hand tracking is illustrated in Fig.1.

2.2. Prerequisites

In original TLD, a hand is represented by a bounding box, parameterized by its location and scale, manually initialed in the first frame. Hand appearance model is a structure consists of a collection of positive and negative patches, which grows as new patches coming into the model, $M = \{p_1^+, p_2^+, ..., p_m^+, p_1^-, p_2^-, ..., p_n^-.\}$, in an order of the time they are added in. In the realization of this hand model, we limit the maximum number of patches to a threshold. If the number of positive or negative patches exceeds the threshold, patches will be randomly removed.

Skin color has been widely studied and proven to be useful in face detection, localization and tracking [18]. It has been integrated in TLD for hand is a region consists of skin color pixels while multi-cue can gain more robustness as explored in our previous work [19]. As described in [18], skin color distribution can be modeled by an elliptical Gaussian joint probability density function (*pdf*), defined as:

$$p(c|skin) = \frac{1}{2\pi|\sum_s|^{1/2}} \cdot e^{-\frac{1}{2}(c-\mu_s)^T \sum_s^{-1}(c-\mu_s)} \quad (1)$$

here, c is a color vector, μ_s and $\sum_s$ are the distribution parameters, estimated from training data:

$$\begin{aligned} \mu_s &= \frac{1}{n}\sum_{j=1}^{n} c_j \\ \sum_s &= \frac{1}{n-1}\sum_{j=1}^{n} n(c_j-\mu_s)(c_j-\mu_s)^T \end{aligned} \quad (2)$$

2.3. Hybrid Base Tracker

Original tracker in TLD is based on KLT optical flow tracking [20], which is good for texture objects but not for articulated ones. The Flock-of-Features tracker was originally proposed for tracking articulated object, and has shown superior performance on hand tracking over CAMShift [21]. The major power of FoF tracker lies in that the feature points tracked are constrained by a "Flock Rule", which states that the points p_i in the flock $\mathcal{F} = \{\{p_i\}_{i=1}^{N_f}\}$ should satisfy:

$$\begin{aligned} &\text{MINDist} < |p_i - p_j|, \forall i,j \in \{1,2,...,N_f\}, \text{and} \\ &\text{MAXDist} > |p_i - m|, \forall i \in \{1,2,...,N_f\}. \\ &m = \text{median}(\mathcal{F}) \quad or \quad \text{centroid}(\mathcal{F}). \end{aligned} \quad (3)$$

where $|p_i - p_j|$ is the distance between p_i and p_j. The MINDist and MAXDist are parameters the flock holds, which constrain the points in the flock so that they can cover a certain area without being too converged or too scattered. This "Flock Rule" is quite useful for tracking articulated objects since each of its articulated parts can be covered by some feature points, so it is applied in our hybrid tracker.

Algorithm 1 The modified Flocks-of-Features tracker

INPUT: Last bounding box B_{pre}, search window W, last feature points $\mathcal{F}_{pre}$, confidence map M
OUTPUT: New bounding box B_{new} (may be empty)
BEGIN:
$\{\mathcal{F}_{\text{tracked}}, \mathcal{F}_{\text{lost}}\} \leftarrow \text{KLT}(\mathcal{F}_{pre}, B_{pre}, I)$
$\mathcal{F}_{\text{lost}} \leftarrow \text{getSkinPoints}(\text{M}, \text{W})$
$S_p = \sum_{i=1}^{N_f} \text{M}(\mathcal{F}_{pre}(i)) \times \mathcal{F}_{pre}(i)$
for $i = 1 \rightarrow N_f$ **do**
  "Center": $c_i = \frac{S_p - \mathcal{F}_{pre}(i)}{N_f - 1}$
  "Positive Driving": $d_p = c_i - \mathcal{F}_{pre}(i)$
  "Negative Driving": $d_n = 0$
  for $j = 1 \rightarrow N_f$ **do**
    if $j \neq i \wedge ||\mathcal{F}_{pre}(i) - \mathcal{F}_{pre}(j)||_2 \leq d_{\text{min}}$ **then**
      $d_n = d_n + d$
    end if
  end for
  $\mathcal{F}(i) = \mathcal{F}_{pre}(i) + \alpha \times d_p + \beta \times d_n$
end for
Estimate shift and scale change, $B_{new} \leftarrow \{B_{pre}, \mathcal{F}\}$
END

To integrate skin color cue into the base tracker, color points are added to the feature flock by initially sampling from the palm region and then tracked. LK feature points are tracked on grey image by KLT, and color points are tracked on the skin color map. Formula (3) constrains the tracked points. Points lost during tracking are replaced by sampling at proper regions. The whole process is illustrated in Algorithm 1. It can be observed that skin color cue is applied to force the flocks moving to area in which skin color points crowded.

Time complexity of pyramid optical flow is $O(I+F\times L\times K \times W)$, where I is the number of pixels in an image, W is number of pixels in a search window, L is level of pyramid, F is the number of features used and K is iteration time of every level. Therefore, time complexity of hybrid tracker is:

$$T_{tracker} = O(I + F \times L \times K \times W + F \times F). \quad (4)$$

which is a liner function of size of image I and quadratic function of feature points, so it can be realized in real-time.

2.4. Cascade Detector

The original TLD uses variance filter as the first stage of the classifier cascade, suitable for tracking textured objects, but not working well for tracking hand. Skin color cue is applied as first stage in our method. A cost function that measures how well a hand candidate matches the model is defined as:

$$\phi(p_i) \sim -\ln p(p_i|\mu_s, \sum_s) \quad (5)$$

Ferns classifier is the second stage in detection, adopted for its high speed, accuracy and possibility of incremental update. At last, a patch is classified as positive only if it is similar enough to the existing patches in online model using *NN* classifier. The procession of Detection can be described as:

$$\begin{aligned} &Candidate\ Patches \quad P = \{p_1, p_2, p_3, ..., , p_n\}. \\ &\Rightarrow P_1 = \{p_i | p_i \in P \wedge Skin\{p_i\}\}. \\ &\Rightarrow P_2 = \{p_i | p_i \in P_1 \wedge Fern\{p_i\}\}. \\ &\Rightarrow P_3 = \{p_i | p_i \in P_2 \wedge NN\{p_i\}\}. \\ &\Rightarrow B_{detector} = \frac{\sum_1^N p_i * w_i}{N}, \quad p_i \in P_3. \end{aligned} \quad (6)$$

$B_{detector}$ is result by detector, and w_i is assigned in [0,1] according to the distance between p_i and B_{pre}.

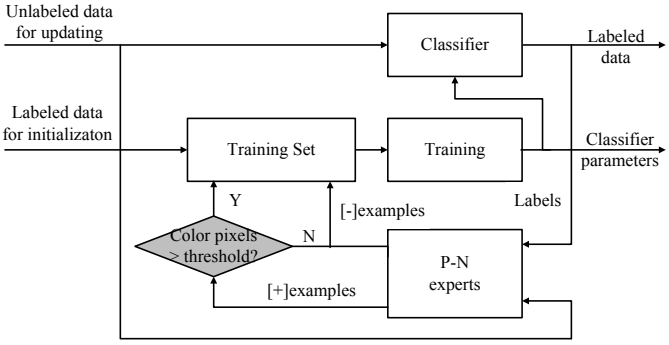

Fig. 2. PN learning initials the classifier with labeled examples from first frame and then iterates in following frames.

2.5. Online Learning with Skin Color

In the original P-N Learning [22], only temporal and spatial constraints are used to correct the labeling of the patches. P-expert requires the patches that are close to the trajectory be labeled positive based on temporal constraints. On the other hand, N-expert requires the patches surrounding the current object be labeled negative based on spatial constraints. But this will surely decrease the learning efficiency in hand tracking context. Therefore, skin color constraints is integrated into this module as in Fig.2. This constraint is based on the assumption that patch labeled positive should contain color pixels in a proportion. The patches labeled as positive with less than a threshold skin color pixels are labeled back to negative patches, which decreases the false alarm.

3. EXPERIMENTS AND DISCUSSIONS

3.1. Experiment Settings

The proposed algorithm is trained with about 1000 frames and tested on 6 real-scene sequences with in total more than 5000 frames as described in Table 1, whose resolution is 640×480 pixels. There is no overlap between training data and testing data. Our method is reimplemented in Matlab 2010b on Windows xp professional with speed of about 10 fps, while original TLD runs at about 15 fps.

Table 1. Comparison on tracking performance, **bold** fonts indicate the best performance.

Dataset sequences		Seq1	Seq2	Seq3	Seq4	Seq5	Seq6
Characteristic		face distraction	large motion	background clutter	occlusion	out-of-view motion	3D changing postures
Num. of Testing Frames		1006	1360	830	527	540	763
Proposed	Recall	**91**%	**92**%	**90**%	**89**%	**91**%	**92**%
	Precision	**86**%	87%	**85**%	**90**%	**90**%	**89**%
CAMShift	Recall	70%	84%	85%	75%	80%	80%
	Precision	60%	84%	80%	76%	65%	70%
FoF	Recall	80%	80%	84%	80%	86%	82%
	Precision	76%	**88**%	**85**%	80%	79%	80 %
Original TLD	Recall	89%	**92**%	88%	86%	89%	85%
	Precision	81%	**88**%	**85**%	87%	**90**%	87%

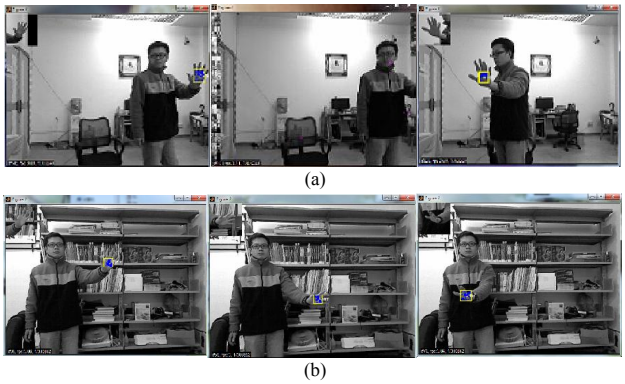

Fig. 3. Proposed method dealing with out-of-view problem in (a) and clutter background in (b).

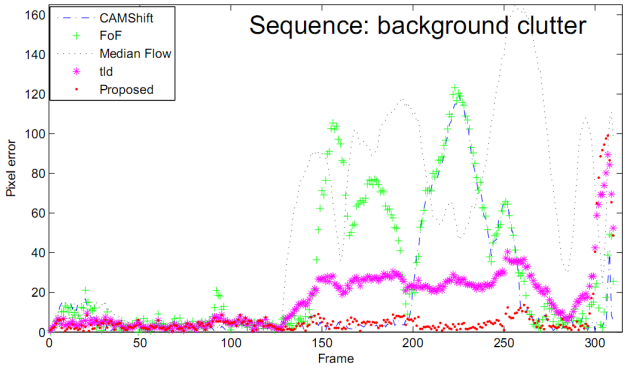

Fig. 4. Comparison of methods about pixels errors against manually labeled ground truth of 300 frames in Sequence 3.

The performance is evaluated with recall and precision, whose definition is:

$$\begin{aligned} Recall &= \frac{\text{Num. of } P^+}{\text{Num. of } P^+ + \text{Num. of } N^-} \\ Precision &= \frac{\text{Num. of } P^+}{\text{Num. of } P^+ + \text{Num. of } P^-} \end{aligned} \tag{7}$$

where P^+ are successfully tracked targets (overlap with ground truth $> 60\%$), P^- are wrongly tracked targets, and N^- are tracking failure reports when the hand is in the image. From the definition, it can be learned that the recall reports the robustness and continues tracking ability of the testing method and the precision shows the ability of discriminate distracter and the target.

3.2. General Performance and Comparison

An overview of the results from different video sequences is shown in Table 1. It can be observed that, the proposed method is more robust than the original TLD, which is the effect of the base hybrid tracker. In sequence 1, face distraction is an unsolved problem of CAMShift for it's based on the color information, which is easy be distracted by face. In sequence 3, the distracter is clutter background, with color cue integrated, the CAMShift and FoF also show robustness in tracking hand. Another great improvement is in out-of-view sequence. With online learning component, the out-of-view and occlusion problems can be tackled, which is nearly impossible for CAMShift or FoF. From Fig.3, it can be observed that the out-of-view problem can be well tackled, which is quite difficult for other methods. And hand can be well tracked under clutter background.

Additionally, our method is compared with other state-of-arts about the pixels errors against ground truth of 300 frames selected from testing sequence 3. Sequence 3 includes face distraction along with background clutter, which is 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 our method achieved the longest stable tracking result.

4. CONCLUSIONS

This paper proposes a novel multi-cue hand tracking algorithm based on an online learning framework called Tracking-Learning-Detection. Main extensions are: firstly, Flock of Features tracker is combined with Median Flow tracker to build a hybrid base tracker for the framework to work more effectively in hand tracking context; secondly, skin color cue is integrated into the framework and increased its robustness against textured background; thirdly, skin color is integrated in PN learning to decrease the false alarm. Extensive experiments show the robustness of our method compared with state-of-art methods. However the main limitation is the speed, which needs to be improved.

5. REFERENCES

[1] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman, "Long term arm and hand tracking for continuous sign language tv broadcasts," *BMVC*, pp. 1105–1114, 2008.

[2] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 1685–1699, Sept. 2009.

[3] J. Rehg and T. Kanade., "Visual tracking of high d-of articulated structures: an application to human hand tracking," in *European Conference on Computer Vision*, 1994, pp. 35–46.

[4] F. Mahmoudi and M. Parviz, "Visual hand tracking algorithms," in *Proceedings of the Geometric Modeling and Imaging-New Trends*, 2006, pp. 228–232.

[5] A. Erola, G. Bebisa, M. Nicolescua, R. D. Boyleb, and X. Twomblyb, "Vision-based hand pose estimation: A review," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 52–73, Oct. 2007.

[6] I. Oikonomidis, N. Kyriazis, and A. Argyros., "Efficient model-based 3d tracking of hand articulations using kinect," *BMVC*, vol. 5, pp. 3, Feb. 2011.

[7] J. Rehg and T. Kanade, "Digiteyes: vision-based hand tracking for human-computer interaction," in *Workshop on Motion of Non-Rigid and Articulated Bodies*, 1994, pp. 16–22.

[8] B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla, "Model-based hand tracking using a hierarchical bayesian filter," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1372–1384, 2006.

[9] G. Bradski, "Computer vision face tracking for use in perceptual user interface," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

[10] M. Isard and A. Blake, "Condensation-conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[11] Michael Donoser and Horst Bischof, "Real time appearance based hand tracking," in *International Conference on Pattern Recognition*, 2008, pp. 1–4.

[12] J. Leistner, C. Saffari, A. Pock, T. Bischof, and H. Bichof, "Prost: Parallel robust online simple tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 723–730.

[13] S. Avidan, "Ensemble tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 261–271, Feb. 2007.

[14] Eng-Jon Ong and R. Bowden, "A boosted classifier tree for hand shape detection," in *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, May 2004, pp. 889–894.

[15] L. Ballan, A. Taneja, J. Gall, L. Van Gool, and Marc P., "3d tracking = classification + interpolation," in *International Conference on Computer Vision*, Oct. 2003, pp. 1441–1448.

[16] Z. Kalal, K. Mikolajczyk, and J. Matas, "Face-tld:tracking-learning-detectioin applied to faces," in *IEEE International Conference on Image Processing*, Sept. 2010, pp. 3789–3792.

[17] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1409–1422, 2010.

[18] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques," in *ICCGV*, 2003, pp. 85–92.

[19] H. Liu, Z. Yu, H. Zha, and Y. Zou, "Robust human tracking based on multi-cue integration and mean shift," *Pattern Recognition Letters*, , no. 30, pp. 827–837, 2009.

[20] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *DARPA Image Understanding Workshop*, 1981, pp. 674–679.

[21] M. Kolsch and M. Turk, "Fast 2d hand tracking with flocks of features and multi-cue intergration," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop)*, June 2004, p. 158.

[22] Z. Kalal, K. Mikolajczyk, and J. Matas, "P-n learning: Boostrapping binary classifiers by structural constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 49–56.