MAXIMALLY STABLE CURVATURE REGIONS FOR 3D HAND TRACKING

Can Wang, Hong Liu and Xing Liu

Engineering Lab on Intelligent Perception for Internet of Things(ELIP),
Shenzhen Graduate School, Peking University, China
Email: canicvol@163.com, hongliu@pku.edu.cn, liuxing@sz.pku.edu.cn

ABSTRACT

Fast and robust hand detections and tracking is in increasing demand from areas such as natural Human Robot interaction(HRI) and surveillance systems. Previous works always use skin color or contour model to detection hand. However, they always fail for hands always exhibits drastic appearance change due to illumination change, non-rigid nature and hands are hard to discriminate from clutter background. Actually, the hand region has a specific nature that its curvature is relatively higher than other body parts and keeps stable whatever its poses and locations are, but none of pervious works exploit this nature for hand detection. In this work, a novel algorithm MSCR (Maximally Stable Curvature Regions) based on curvature nature to detect hands. It does not require manually initialization in the first frame, the hands are located by MSCR and skin color detector in the global image. 3D optical flow integrated Kalman Filter works to estimate the next location for local detector. Extensive experiments demonstrate that robust 3D tracking of hand articulations can be achieved in real-time with accurate results.

Index Terms— Hand Tracking with Detection, Skin Color, MSCR, Spatial Constraints, 3D Optical Flow

1. INTRODUCTION

Bare hands are probably the most natural tools for interactions for its dexterity and have shown great potential in applications for sign language [1], gesture recognition [2] and HCI based applications [3]. Plenty of vision-based methods have been developed to tackle the hand tracking problem in the past decades. Its major difficulty stems from 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. Fast and robust hand tracking has not yet achieved, and global hand tracking is still of great challenge in complex scenarios. In the past decades, plenty of methods have been developed to tackle the problem of vision-based hand pose recovery. Hand tracking methods can be grouped into model-based and appearance-based ones [4, 5, 6, 7]. 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 [8, 9]. 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 [10]. Contour features have been used in particle filter based methods [11]. Maximally Stable Extremal Regions (MSER) is a new stable region feature and has been used for hand tracking [6].

These methods are robust to appearance change to a certain degree, but they always fail for hands always exhibits drastic appearance change due to illumination change, non-rigid nature and hands are hard to discriminate from clutter background. At the same time, most of the above researches assume restrictions on user's clothing, the scene and hand motion speed, the main reason is lack of more robust information. Recent years, depth sensors like Kinect are becoming affordable, and depth information has been widely studied and applied to hand tracking [7]. Depth information is robust to the illumination change. With depth information, body contour can be easily and accurately obtained, which can be applied to extract more robust features.

Actually, the hand region has a specific nature that its curvature is relatively higher than other body parts and keeps stable whatever its poses and locations are, but none of pervious works exploit this nature for hand detection. Works apply curvature feature to hand are always confined to the curvature of fingers and palm [12, 13], while our work focus on the curvature of hand among body contour. [12] proposes a curvature based hand shape recognition system for a virtual wheelchair control interface, which studies curvature of fingers and palm to recognize hand shape. CSS (Curvature scale space) is used in [13] to recognize hand pose, which also limit curvature to the local hand. In our work, curvature is applied to detect hand among human body contour, which is not used ever. It is observed that the curvature of hand keeps the same no matter how the appearance changes. Another observation is that hand region has the most stable and highest curvature among human body parts. With these two observations, hand region can be extracted from human body parts by calculating curvature of the body contour.

In this paper, color information and depth information are applied to detect hands, which are fused by the spatial constraints of human body. Hands are automatically detected in global image of first frame by the detector. Skin color is an indispensable feature of human hand which has been applied to detect hands. A novel method MSCR is a novel algorithm to find convex region with maximally stable curvature, given an enclosed contour. MSCR can be used to separate hands from human body based on the observation that hand is a maximally stable curvature region of human body. Then a tracker that combines a fast 3D optical flow of hands region and Kalman Filer is proposed to estimate next position of hands. And detector is activated to localize hands in the local region by result of tracker, which is much faster than global detector.

To the best of our knowledge, our method is the first that (a) proposes MSCR (Maximally Stable Curvature Regions) method to detect hands from depth image, (b) combines a fast 3D optical flow and Kalman Filter to estimate hands motion to predict next location. Tracking with detection framework used in our method achieves robust result in real-time, at about 20 fps.

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.JC201005280682A, CXC201104210010A).

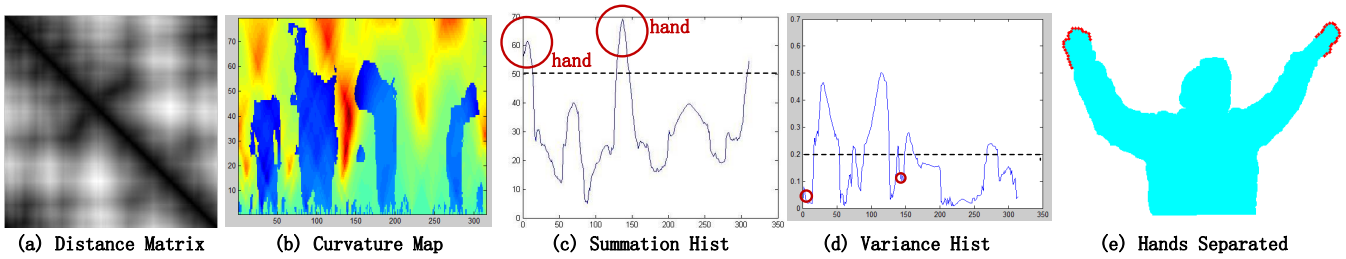

Fig. 2. Work-Flow of MSCR Algorithm

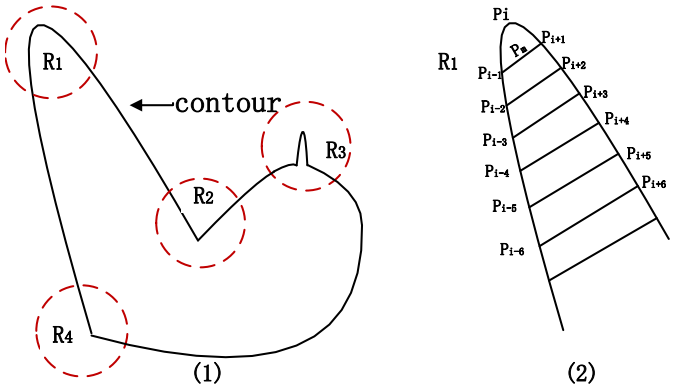

Fig. 1. An example of MSCR. (1) is a enclosed contour with candidate regions R_1, R_2, R_3, R_4. (2) is a curvature calculation procession of pixel p_i

Algorithm 1 Maximally Stable Curvature Regions

Input: Contour C , Distance Marix D.
Output: Maximally Stable Curvature Region R.
Begin:
Initial region R and step s.
for $\forall\ p_i \in C$ **do**
  for $s = 1 \rightarrow \frac{n}{4}$ **do**
    if p_m is in inner region of **C**
      R is concave, $C_s \leftarrow 0$
      else R is convex
        $C_s \leftarrow s\ /\ d(p_{i-s},p_{i+s})$
      end else
    end if
  Curvature Map $M \leftarrow \{C_{p_i} | 1 < i < n\}$
  end for
  if $Sum\{C_{p_i}\} > \tau_{sum}$ and $Variance\{C_{p_i}\} < \tau_{var}$
    p_i is stable, $R \leftarrow \{R, p_i\}$
end for
End

2. MAXIMALLY STABLE CURVATURE REGIONS

MSCR (Maximally Stable Curvature Regions) is an algorithm to extract a convex region of enclosed contour, which has the highest and most stable curvature, illustrated in Algorithm 1.

An example is shown in Fig. 1 to make MSCR algorithm easier to understand. (1) is an enclosed contour with candidate corner regions R_1, R_2, R_3, R_4. According to our definition, R_1 is the MSCR region in the contour. R_2 is not MSCR because it is a concave corner, although it has stable and high curvature. R_3 is not a stable curvature region, for it only has a high curvature when step size is small, which is likely to be noise. R_4 is a smooth region corner which always has a low curvature. In (2), the calculation procession of pixel p_i in R_1 is illustrated. When step is t, the curvature is calculated as : $c = t/d(p_{i-t}, p_{i+t})$, where $d(p_{i-t}, p_{i+t})$ can be inquired in distance Matrix M.

MSCR is applied to detect hands based on the fact that human hand is the only MSCR region among human body parts, as hand has the highest and stable curvature among human body. At first, depth information can be obtained from the Kinect, which is robust to the illumination change and color distractions. Then we have to translate depth data captured from the Kinect in image coordinates to the world coordinates in order to avoid prospective effective. Suppose a point coordinates in depth data is (u, v, d), where (u, v) is coordinates in color image and d is coordinates of depth. The world coordinates is (x_c, y_c, z_c). The translation equation is defined as follows:

$$x_c = \frac{(u - u_0)d}{f_u},\ y_c = \frac{(v - v_0)d}{f_v},\ z_c = d. \qquad (1)$$

After translating depth data to the world coordinate, we project the depth image to front view and top view to get the contour of human body termed as C_f, C_t. A distance matrix M can be obtained from calculating the distance between every two pixels in the contour C. Then curvature distribution map D is obtained by calculating curvatures of a pixel by different steps. The summation and variance of curvatures are applied to decide whether curvatures of a pixel is stable. The MSCR Region consists of stable pixels is termed as R in the contour C, which is most likely to be hands.

Procession of real test frame of human hands is illustrated in Fig. 2. (a) is the distance matrix map between pixels on the enclosed contour. (b) is the curvature distribution map obtained by calculating the curvatures in different step of each point. Curvature of different step is plotted as a function of pixel index. Curvature is growing higher from blue to red. (c) is a summation hist map of the curvature of every pixel. Peaks with red circle are regions whose curvature exceeds a threshold. These regions are candidates of hand. (d) is a variance map represents the curvatures of a pixel under different steps. We get the summation hist and variance hist of every pixel. If the summation of a pixel is higher than a threshold, it demonstrates that this pixel has high curvature. Then the variance hist is applied to judge the stability of that pixel. If the curvature variance of different steps is lower than a threshold, the pixel is stable. Only region consists of pixels with high summation of curvature and low variance is MSCR region. (e) is the result of separating hands from human body after projecting the MSCR region back to the depth contour, plotted in red dots. A confidence map that measures how well a MSCR region r to be hand is defined as: $M_2 = \{p(l_i^t | r_i^t)\}$, which is the probability of seeing the hand at l_i^t, given MSCR region r.

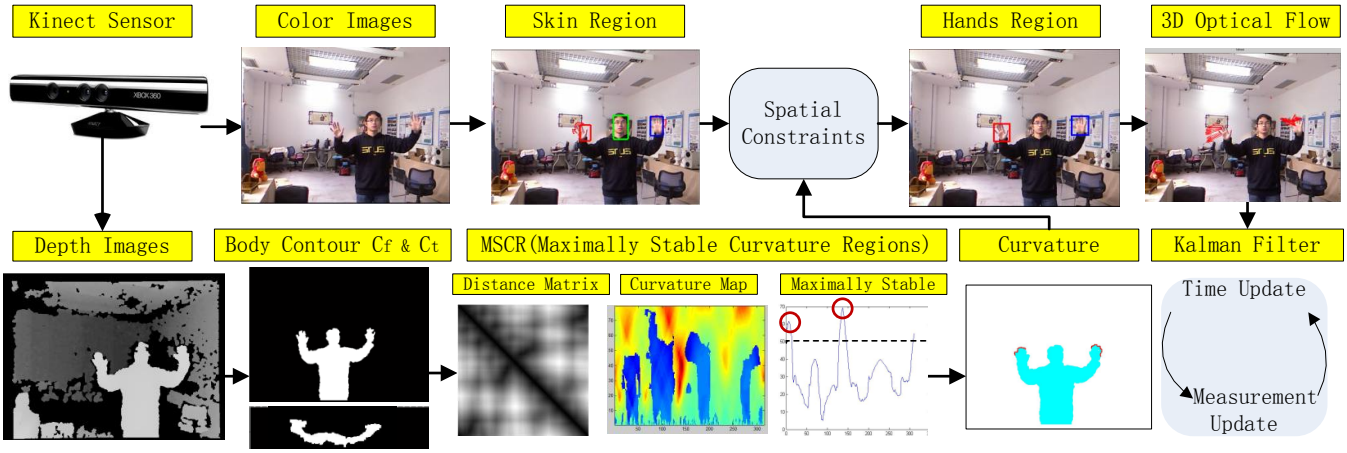

Fig. 3. A brief illustration of hand tracking method based on MSCR.

3. HAND TRACKING BASED ON MSCR

3.1. Problem Setting and Overview

The task is to extract information about human hands from color images and depth images obtained from the Kinect sensor. At first, skin color regions as face and hand are extracted from the image. However the skin color cue is quite sensitive to the illumination change, which is not very robust. So the depth information is applied to increase the robustness. With the novel method MSCR, we can obtain the corner with stable high curvature. MSCR is a novel way to locate maximally stable curvature region from depth images. A spatial constraints can be applied to combine the skin color and MSCR result to separate hands from human body. After hands are detected, a fast 3D optical flow is used to describe the motion information of hands. And then, Kalman filter is applied to track hands with 3D optical flow. After hands are tracked, the local detector of skin color and MSCR is activated to detect hands in local area. The framework of our method is a combination of tracking and detection. More details of the work-flow of our method is illustrated in Fig. 3.

3.2. Skin Color Model

Skin color feature has been widely studied and proven to be useful in face detection, localization and tracking [14, 15]. As described in [14], skin color distribution can be modeled by an elliptical Gaussian joint probability density function (*pdf*), defined as:

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

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

$$\begin{aligned} \mu_s &= \frac{1}{n}\sum_{j=1}^{n} C_j \\ \sum_s &= \frac{1}{n-1}\sum_{j=1} n(C_j-\mu_s)(C_j-\mu_s)^T \end{aligned} \quad (3)$$

A confidence map that measures how well a candidate patch matches the skin color model can be defined: $M_1 = \{p(l_i^t|\mu_s, \sum_s)\}$, the probability of presence of hand at location l_i^t, computed by the normalized sum of pixel likelihoods in patch p at location l_i^t.

3.3. Hand Spatial Constraints

There is strong spatial constraints between the two hands and head as illustrated in [16], which can be applied to combine color detector and MSCR detector. A model that constraints hands and head spatial distribution is designed using a cost function $\varphi(l_i^t, l_j^t, l_h^t)$, where l_i^t is the location of left hand in frame t, l_j^t for right hand and l_h^t for head, defined as:

$$\varphi(l_i^t, l_j^t, l_h^t) \sim -\ln p(l_{left} = l_i^t, l_{right} = l_j^t, l_{head} = l_h^t) \quad (4)$$

where the $p(l_{left} = l_i^t, l_{right} = l_j^t, l_{head} = l_h^t)$ is the joint probability of seeing left hand at l_i^t, right at l_j^t and head at l_h^t. And when head is not detected, the distribution model is modeled as: $\varphi(l_i^t, l_j^t) \sim -\ln p(l_{left} = l_i^t, l_{right} = l_j^t)$. If one hand is occluded, the distribution model is modeled as: $\varphi(l_i^t) \sim -\ln p(l = l_i^t)$.

3.4. Hand motion model

So far, we have located hands region in the image, the motion of hand is model as 3D optical flow described in our previous work. Suppose the hand region candidate in frame *t* is termed as X_t, a fast 3D optical flow vector is calculated to update the Kalman Filter. A parallel implementation of LucasKanade(LK) method [17] is utilized to estimate 2D dense optical flow between points set in intensity image candidate patch $I(X_t^i)$ and $I(X_{t-1}^i)$. Suppose the 2D optical flow of point set **s** in i^{th} candidate is $(v_x^i(\mathbf{s}), v_y^i(\mathbf{s}))$, its depth component v_z is calculated as:

$$v_z^i(\mathbf{s}) = D_t(\mathbf{s} + (v_x^i(\mathbf{s}), v_y^i(\mathbf{s}))) - D_{t-1}(\mathbf{s}) \quad (5)$$

Kalman Filter model is iteratively updated with the 3D optical motion to estimate next location of hands. And then skin color and MSCR are applied to detect hands in the local region output by tracker.

4. EXPERIMENTS AND DISCUSSIONS

4.1. Experiments Setting and Data set

To demonstrate effectiveness of our method, two groups of experiments are conducted on a RGB-D video data set recorded via Kinect. And all experiments are conducted on a Pentium i3-2410M 2.3 GHZ PC with 2.0 Gb RAM. Ground truth square bounding boxes for left hand, right hand and head are annotated manually. We use a training data set of about 1000 frames and testing data set of about 8000 frames, whose resolution is 640 $\times$ 480. Our method runs nearly real-time, at about 20 fps.

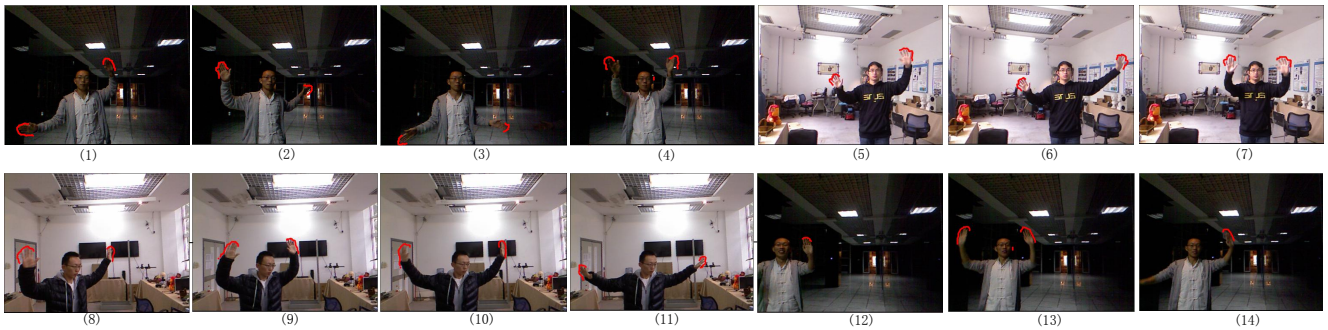

Fig. 5. A brief illustration of results in challenge testing sequences by MSCR algorithm.

Fig. 4. Result of Color detector and MSCR detector in our tracking framework. Bounding boxes are results of Color detector and red dots are results of MSCR detector

4.2. General performance of our method

As shown in Table 1, we test our method with four challenging sequences. The complex conditions include illumination change, occlusion, drastic change and fast move. In sequence 1, hands move at night in outdoor environment. In sequence 2, hands are making circle in front of clutter background. In sequence 3, the face of tester is not always facing the Kinect camera, which generates difficulties for skin color detector and spatial constraints. In sequence 4, one hand sometimes moves out of the view, which is considered to be full occluded or lost. The hands are considered to be successfully tracked if the bounding box overlaps with ground truth $> 60\%$.

General performance of correct rates of our method dealing with challenging sequences can be seen in Table 1. The reason our method achieves good performance of sequence 1 is that MSCR is not sensitive to illumination change. The result on sequence 3 is not that good as head detection is important in spatial constraints to fuse the skin cue and MSCR. The result of sequence 4 demonstrates that our method is mainly based on MSCR, which is robust to occlusion. If a hand is occluded, the other one will be extracted as before. Tracking result of our method in sequence 2 is illustrated in Fig. 4. Head and hands are located by skin color detector shown in colored bounding boxes. Pixels of MSCR result is plotted in red. It can be observed that when the left hand moves by the panel on the wall, skin color detect fails while MSCR still works well. Because color of the panel is similar to human hand, then skin color failed. On the other hand, MSCR is based on depth information, not sensitive to illumination and color distraction, so MSCR detector works well.

4.3. Analysis on MSCR

The second group of experiments is to evaluate the performance of MSCR in real sequences of complex conditions. Four challenging testing sequence are used to test MSCR, shown in Fig. 5, in which pixels in MSCR are plot in red. (1) - (4) is the result of MSCR on sequence 1 (Out door and Night). Tester waves his hands from left to right. (5) - (7) is the result of sequence 2 (Out door and Night). Though hands move in front of clutter background, MSCR works well to detect hands. (8) - (11) in second row is the result of sequence 3 (Not always facing Camera). When tester dose not face the camera, head is difficult to detect. (12) - (14) is the result of sequence 4 (Occlusion at outdoor in Night). Right hand moves into the view and then moves out. It can be seen that if one hand is fully occluded, the other one can still be detected.

Table 1. General Performance of Our method against 4 real challenging sequences

Seq. NO.	Characteristics	Tracking Frames	Correct Rates
Seq.1	Night, out door	1580	0.90
Seq.2	Day, circling	1478	0.93
Seq.3	Not always facing camera	2547	0.82
Seq.4	Night, occlusion	2632	0.88

From this experiment on testing sequences, high accuracy of MSCR has been achieved, which is important for the hand detector. It can be observed that MSCR is robust to out-of-plane rotation, great appearance change, illumination change and fully occlusion with high speed. [18] proposes a method to locate body part by calculating path of candidate points by Dijkstra's algorithm, which requires high time complexity, while our method is fast for it only calculates curvature of each point in contour. Based on the MSCR, orientation of hands can also be calculated, which will be much useful for HCI/HRI and gesture recognition. It has to be pointed out that if hand is put on human body or hand is hold something, the algorithm will fail to detect hands. So skin color detector is added into our framework to make up for MSCR. With skin color and MSCR, hands can be detected in most cases.

5. CONCLUSIONS

This paper proposes a novel hand tracking algorithm combing color cue and depth information obtained by Kinect sensor. The good results of MSCR demonstrate that curvature nature of hand is appropriate for hand detection. Firstly, an algorithm extract stable region with high curvature called MSCR is proposed to detect the hands. Secondly, a new framework combines detector and tracker is proposed to tracking hands, which gains good result. MSCR shows good performance in detecting hands, while there are still several parts that should be improved. First, the way to calculate contour curvature should be improved to be faster. Second, orientation of MSCR regions should be calculated, which will be useful in our future work about gesture recognition and robot control.

6. 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 dof articulated structures: an application to human hand tracking," in *ECCV*, 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] Ali Erola, George Bebisa, Mircea Nicolescua, Richard D. Boyleb, and Xander Twomblyb, "Vision-based hand pose estimation: A review," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 52–73, Oct. 2007.

[6] Michael Donoser and Horst Bischof, "Real time appearance based hand tracking.," in *ICPR*, 2008, pp. 1–4.

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

[8] 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. 16–24.

[9] 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.

[10] G. Bradski, "Computer video face tracking for use in perceptual user interface," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 705–740, 1998.

[11] 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.

[12] Seong-Pal Kang, Michal Tordon, and Jayantha Katupitiya, "Curvature based hand shape recognition for a virtual wheelchair control interface," in *ICRA*, 2004, pp. 2049–2054.

[13] C. Chang, I. Chen, and Y. Huang, "Hand pose recognition using curvature scale space," in *ICPR*, 2002, pp. 386–389.

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

[15] Charles Bibby and Ian Reid, "Robust real-time visual tracking using pixel-wise posteriors," in *Proceedings of the European Conf on Computer Vision, Marseille*, 2008, pp. 831–844.

[16] Trinh H., Quanfu Fan, Gabbur P., and Panksanti S., "Hand tracking by binary quadratic programming and its application to reail activity recognition," in *CVPR*, 2012, pp. 1902–1909.

[17] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 1981 DARPA Image Understanding Workshop*, 1981, pp. 121–130.

[18] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun, "Real-time identification and localization of body parts from depth images," in *ICRA*, 2010, pp. 3108–3113.