21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Combining Contrast Saliency and Region Discontinuity for Precise Hand Segmentation in Projector-Camera System

Jingwen Dai and Ronald Chung
Department of Mechanical and Automation Engineering
The Chinese University of Hong Kong, Shatin, NT, Hong Kong
Email: {jwdai, rchung}@mae.cuhk.edu.hk

Abstract

One goal of projector-camera system is let human finger be used like a mouse to click and drag objects in the projected content. It requires segmentation of the human palm and fingers in the image data captured by the camera, which is a challenging task in the presence of the incessant variation of the projected video content and the shadow cast by the palm and fingers. We describe a coarse-to-fine hand segmentation method for projector-camera system. After rough segmentation by contrast saliency detection and mean shift-based discontinuity-preserved smoothing, the refined result is confirmed through confidence evaluation. Extensive experimental results are shown to illustrate the accuracy and efficiency of the approach.

1. Introduction

In recent years a number of so-called pico-projectors launched to the market, which are of extremely small size and low cost, yet with adequate performance, making the projector-camera system widespread in visualization and human-computer interaction (HCI) [8]. Hand segmentation, as the first step for most barehand-based applications, plays an important role in the robustness, accuracy and efficiency of a HCI system.

The approaches for hand segmentation have been studied extensively in computer vision society. Among them, skin color detection [7, 6] is very common for its simpleness and easy implementation, and is very efficient against simple background or in the scene of hand being the only skin-colored object. However, diverse video contents are projected continuously in projector-camera scenario, when some skin-colored objects are projected on the background (Region A in Fig. 1) or non-skin-colored objects are projected on the hand (Region B in Fig. 1), the skin color based methods will be influenced severely. Since the geometrically and radiometrically calibrated projector-camera system can predict where the video contents are projected and how they should appear in the image data, background subtraction [1] is adopted to segment the hand as the set of pixels that are out of expectation on the projection surface, but suffers from separating hand region from the hand-cast shadows (Region C in Fig. 1), let alone calibration procedures and constraints of constant ambient illuminations and fixed projection surface. The graph-based [9, 3] approaches are able to generate good segmentations. However, the time consuming of these approaches and the requirement of user's interaction would weaken their advantage for the HCI application where the speed is an important factor for realtime interaction. Rather than monocular camera, some researchers use additional instruments, such as infrared camera [12], stereo camera [10], depth sensor [4], to distinguish hand region from background, that inevitably increasing the complexity of projector-camera system configuration.

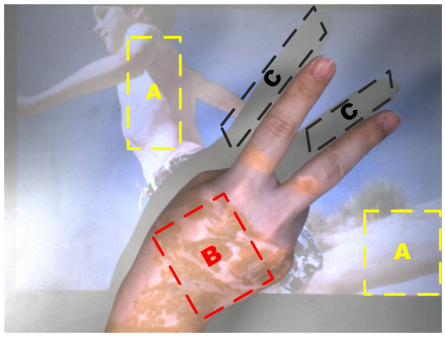

Figure 1. A sample hand image captured by projector-camera system.

In this paper, we introduce a coarse-to-fine approach

to solve the aforementioned problems. The main idea is to combine contrast saliency map with mean-shift based smoothing and segmentation by a confidence function. Low-level contrast saliency detection enables the hand region to be highlighted roughly, and mean-shift based smoothing method removes the noises induced by projection contents without demolishing discontinuity information. Moreover, without any pre-training and pre-calibration procedures, the robust, precise and also rapid hand segmentation can be derived.

2. Methodology

2.1. Rough Segmentation by Contrast Saliency

Although incessant varied video contents are projected onto the projection surface and the hand operating above, it is obvious that the hand is always the most noticeable object from the human vision system's perspective. Motivated by this biological vision cue, firstly we employed a saliency detector to derive a rough hand region segmentation. Salient region detection as a typical low-level vision approach has been widely studied. According to our special projector-camera scenario, the saliency detector must satisfy the following requirements: (a) Uniformly highlighting the largest salient objects; (b) Accomplishing detection less than $15ms$ for real-time requirement. After compare different saliency detection methods, we chose the histogram-based contrast [5], which best fulfills the aforementioned criterions, to define the saliency values for image pixels.

The saliency of a pixel is defined using its color contrast to all other pixels in the image, pixels with the same color value have the same saliency value, which is defined as

$$S(I_k) = S(c_l) = \sum_{j=1}^{n} f_j D(c_l, c_j), \qquad (1)$$

where c_l is the color value of pixel I_k, n is the number of distinct pixel colors, $D(c_l, c_j)$ is the color distance metric between colors c_l and c_j in the HSV color space, and f_j is the probability of pixel color c_j in image I.

In order to reduce the high dimension of 256^3 true-color space, more frequently emerging 85 colors were selected by building a compact color histogram using color quantization. At the same time, artifacts would be introduced. A smoothing procedure is used to refine the saliency value for each color, which replacing the saliency value of each color by the weighted average of the saliency value of similar colors. Typically, $m = n/4$ nearest color are chosen to refine the saliency value of color c by

$$S'(c) = \frac{1}{(m-1)T} \sum_{i=1}^{m} [T - D(c, c_i)] S(c_i), \qquad (2)$$

where $T = \sum_{i=1}^{m} D(c, c_i)$ is the sum of distances between color c and its m nearest neighbors c_i, and the normalization factor comes from $\sum_{i=1}^{m} [T - D(c, c_i)] = (m-1)T$. More implementation issues are detailed in [5]. The saliency map $S(x, y)$ of image I (Fig. 2-a) is derived as shown in Fig. 2-b.

2.2. Mean-Shift Region Smoothing

Even though the hand region has been highlighted through saliency detection, as illustrated in Fig. 2-b, it is not uniformly emphasized due to the influence of the projection content on the hand and projection surface. Hereby, it is impossible to have precise hand segmentation through traditional threshold methods. We employed mean-shift based smoothing and segmentation approach [2] in the salient regions, which not only eliminates the noises but also preserves the discontinuity by adaptively reduce the mount of smoothing near abrupt changes in the local structure, i.e. boundaries.

An important advantage of mean shift-based segmentation [2] is its modularity which makes the control of segmentation output very simple, just through three parameters: (h_s, h_r, M). The range parameter h_r and the smallest significant feature size M control the number of regions in the segmented piecewise constant model, larger values have to be used for h_r and M to discard the effect of small local variation. The spatial parameter h_s determines the size of spatial window. In our case, (h_s, h_r, M) is set to $(7, 10, 20)$.

It is worth mentioning that the inherent iterative property of mean-shift based method usually invokes the efficiency problem. The rough salient region detection decreases the mean-shift search space which accelerates the convergence speed dramatically.

After mean-shift smoothing and segmentation, the image is divided into L candidate partitions P_k, $i = 1, \ldots, L$, as shown in Fig. 2-c. The contour of the hand is preserved well.

2.3. Precise Segmentation by Fusing

For the sake of acquiring precise hand region segmentation, we proposed a confidence function combining contrast saliency and region discontinuity to evaluate the probability of a candidate partition to be a part of hand region. The value of confidence function $C_F(k)$

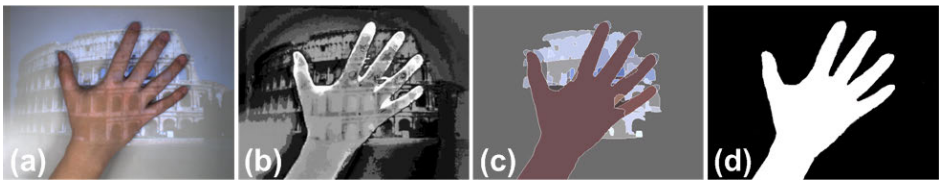

Figure 2. (a) Origin image; (b) histogram contrast saliency map; (c) partitions derived through mean-shift; (d) refined segmentation result.

for partition P_k is formulated as

$$C_F(k) = \frac{1}{e^{(L-1)}}[\alpha\bar{S}(k) + \beta\bar{S}_N(k) + \gamma A(k)], \quad (3)$$

where $\bar{S}(k)$ is the average saliency value of the pixels in P_k, $\bar{S}_N(k)$ is average saliency value of its N neighbor partitions, and $A(k)$ is the partition's area. The three terms above are all scaled to $[0, 1]$. L is the number of image boundary to which the partition attached, when $L \geq 2$, it is indicated that the partition belongs to background that should have low confidence value. α, β, γ are the average weights. If the number of the neighbor partitions N is equal to 1, $\beta = 1/2, \alpha = \gamma = 1/4$, which means that the confidence value is mostly depends on its surround neighborhood, when the partition is an isolated area in hand region or background region; Otherwise, $\alpha = 1/2, \beta = \gamma = 1/4$.

If $C_F(k)$ is greater than a pre-defined threshold Δ, the partition is considered as a part of hand region. Hence, the refined binary segmentation is derived, as shown in Fig. 2-d.

3. Experimental Results

The projector-camera system we used in our experiment consisted of a Pico DLP projector of resolution 640×480 and a CCD camera of resolution 648×488. The system was calibrated geometrically and radiometrically by method detailed in [11] for background subtraction method.

We collected a great diversity of images (e.g. flowers, buildings, celebrities, animals etc.) from Google and projected them to a desk surface. An experimental dataset of 500 images was captured under different projection contents and different hand shapes. The ground-truth is manually annotated with the assistance of GrabCut [3]. Several test images with their ground-truth are shown in Fig. 3-a and 3-b. In order to illustrate the merits of proposed method, we conducted comparison experiments with some related methods. The choice of these methods is motivated by the following reasons: citation in literature (the classic approach of statistical color model-based (SCM) method is widely cited [7]), precision (the background subtraction method (BkSub) has higher precision, since it is on the basis of using pre-calibrated geometric and radiometric information to predict the background image [1]), and recency (the sophisticated graph based method (GB) [9]).

Method	Ours	SCM [7]	BkSub [1]	GB [9]
Time (ms)	29.6	10.9	2.3	115.2

Table 1. Average processing time.

As in [5], we adopted the F-beta score to evaluate the accuracy of segmentation, which considers both the precision p and the recall r to compute the score: $p = N_C/N_R$ and $r = N_C/N_G$, where N_C, N_R, N_G are the number of correct segmented pixels, all segmented pixels and ground-truth pixels respectively. The F-beta score is the harmonic mean of precision and recall, formulated as $F_\beta = (1+\beta^2)pr/(\beta^2 p + r)$, β is set to 0.3 to weight precision more than recall. The visual and quantitative comparisons are shown in 3 and 4 respectively. Among all the methods, our method shows the highest precision, recall and F_β values. It is evident that the skin color-based method (SCM) gets low precision when some projected objects have the color similar to skin, such as human face and yellow flower in the case of Fig. 3-d1 and 3-d2. The background subtraction method (BkSub) shows a high recall but poor precision, verifying that the shadow cast by video projection has great influence (Fig. 3-e3 and 3-e4). The graph-based method (GB) can not reserve smooth boundaries and confuse projected objects with hand region, which are the main reasons for low precision (Fig. 3-f3 and 3-f4).

Table 1 compares the average processing time per frame taken by each method. All the methods are implemented in C++ and executed on a desktop PC with Intel Core 2.8GHz CPU and 2GB RAM. Although our method is not the fastest one, it is sufficiently for real-time applications.

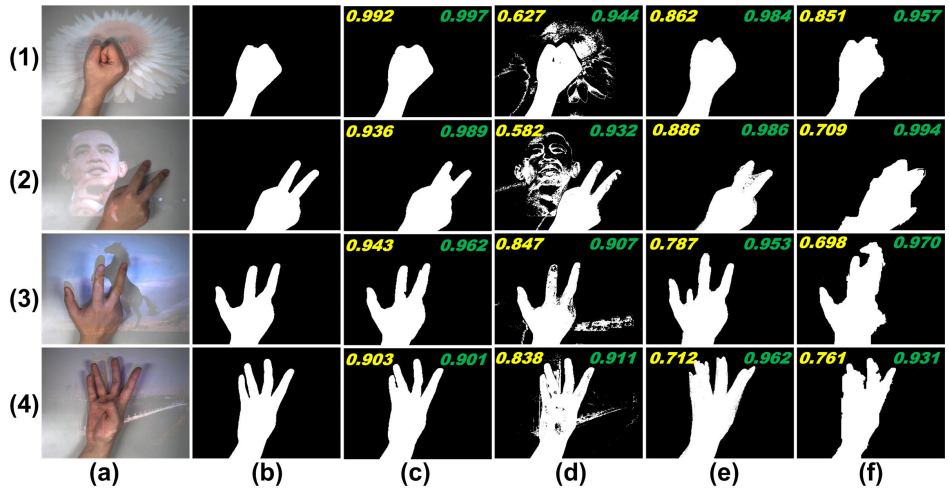

Figure 3. Visual comparison. (a) original image; (b) ground-truth; (c) our method; (d) SCM [7]; (e) BkSub [1]; (f) GB [9]. The yellow (top-left) and green (top-right) numbers in each result image are the corresponding precision p and recall r values, respectively.

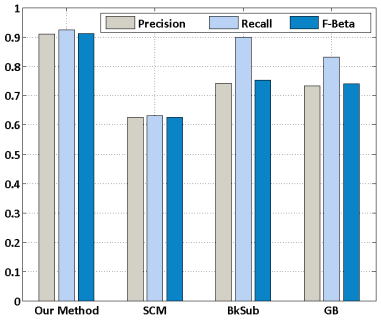

Figure 4. Precision-Recall bars for hand segmentation using different methods. Our method shows high precision, recall and F_β values.

4. Conclusion

We describe a novel coarse-to-fine approach for hand segmentation in projector-camera system, which puts together contrast saliency and region discontinuity information through a confidence function. Experimental results show that the proposed method can segment hand region accurately and rapidly, even if the captured images are interfered by successively projected display content and shadow cast by the moving hand. Future research will almost certainly focus on improving the formulation of confidence function and reducing algorithm complexity to ensure the entire consuming time for hand segmentation is below 15ms/frame.

5. Acknowledgment

This work is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies

References

[1] A. Licsar and T. Sziranyi. Hand gesture recognition in camera-projector system. In *ECCV*, 2004.
[2] C. Christoudias and P. Meer. Mean shift: A robust approach toward feature space analysis. *TPAMI*, 24(5):603–619, 2002.
[3] C. Rother et al. "GrabCut": interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004.
[4] M. Bergh et al. Combining RGB and tof cameras for real-time 3D hand gesture interaction. In *WACV*, 2011.
[5] M. Cheng et al. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011.
[6] M. Donoser and H. Bischof. Real time appearance based hand tracking. In *ICPR*, 2008.
[7] M. Jones and J. Rehg. Statistical color models with application to skin detection. *IJCV*, 46(1):81–96, 2002.
[8] O. Bimber and et al. The visual computing of projector-camera systems. In *SIGGRAPH*, 2008.
[9] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.
[10] Q. Wang et al. Skin color weighted disparity competition for hand segmentation from stereo camera. In *BMVC*, 2010.
[11] X. Chen et al. Color mixing property of a projector-camera system. In *PROCAMS*, 2008.
[12] Y. Sato et al. Fast tracking of hands and fingertips in infrared images for augmented desk interface. In *AFGR*, 2000.