Stereo Vision Based Video Real-Time 3D Pointing Gesture Recognition

Yepeng Guan
School of Communication and Information Engineering, Shanghai University
149 Yanchang Rd., Shanghai 200072, China
E-mail: ypguan@shu.edu.cn
Tel: +86 21 56331967; fax: +86 21 56336908

Keywords: Pointing gesture recognition; stereo matching; segmentation; tracking; location

Abstract

Among natural human gestures occurring during non-verbal communication, pointing gesture can be easily recognized and taken as more natural new human computer interfaces. An approach to pointing gestures recognition is proposed based on binocular stereo vision. A robust automatic foreground extraction algorithm is developed based on wavelet multi-scale transformation across background subtraction without the need to determine an optimal threshold for each image with different clutter background. The method provides 3D information regarding the tracked hands and head. During solving the correspondence of fingertip of the pointing arm, a stereo matching strategy is presented based on geometric constraints between fingertip points, arm axis and epipolar line. Continuity constraint is adopted to insure the validity of the processed correspondence. Experimental results show that the system works in real time on a low cost hardware platform. It is fairly accurate and independent of user characteristics and environmental changes.

1 Introduction

Hand gestures play a natural and intuitive communication mode for all human dialogs. The ability for computers to visually recognize hand gestures is essential for future human-computer interaction (HCI). In current application, keyboards, mice, wands, and joysticks are still the most popular and dominant devices. However, they are inconvenient and unnatural. The use of human movements, especially hand gestures, has become an important part of HCI in recent years. Among natural human gestures occurring during non-verbal communication, pointing gesture can be easily recognized and included in more natural new human computer interfaces, which does not require any a priori skills or training, and is a perfect candidate for the design of a natural interaction device based on computer [3].

Researchers have been working on the problem of a gesture based interface for several years. A simple example of vision-based hand pointing interface is the "digital desk" [13], and re-proposed with diverse enhancements in the "virtual touch-screen" [9] and "finger paint" [4] systems described respectively. These systems share the same elements, a video projector, a camera and a planar surface (screen). The camera is located so as to have the screen in view. System output is displayed by the projector onto the screen, whose locations can be pointed at by the direct contact of the index finger. Baudel [1] presented glove-based systems which need the user to wear a tethered glove. Kahn [6] introduced an early method to detect pointing gestures by using the Perseus architecture. Kolesnik [8] used an overhead camera where they segment the observed silhouette in two parts for the body and one for the arm. Nickel [11] presented pointing detection for human machine interactions that uses Hidden Markov Models trained on different sample pointing gestures to detect the occurrence of a pointing gesture. Segen [12] use a stereo pair of calibration cameras to recognize pointing gestures. Yamamoto [10] used stereo cameras mounted in the corners of a ceiling that look down at an oblique angle to capture entire bodies and faces. Kehl [7] presented a multi-view approach to measuring 3D directions of one or both arms in 3D. These methods assumed that eyes and the fingertip of pointing arm were found in all frames.

Unlike these existing approaches, we propose our approach to pointing gestures recognition based on binocular stereo vision. Our pointing direction is based on the line of sight connecting the point of pseudo-eyes below a certain distance of head top vertically and the fingertip of the pointing arm. Based on the biometric constraint that the eyes lie at some distance below the head top vertically if the user is standing upright, we propose to use the position below the head top instead of that of eyes, which makes our approach different from other approaches proposed recently [2,3,5,11,14]. Our model does not constrain the flat surface pointed to by user to be visible by the cameras. The only condition on camera placement is that both mounted overhead cameras which look down at an oblique angle to capture head and pointing arm. The unique feature of our system is that cameras do not need to capture entire bodies and faces since the positions of user's eyes are estimated by that of user's head top.

The organization of the rest of the paper is as follows. In the next section, we discuss foreground segmentation based on wavelet multi-scale transformation. In section 3, pointing gesture analysis is introduced. Experimental results are given in section 4 and followed by conclusions in section 5.

2 Foreground segmentation

Background subtraction is a commonly used method for segmenting out objects of interest in a scene by comparing the intensity of the background and observed images. It is well-known that optimal selection of threshold value is difficult but very important for the precision of moving object detection during carrying background subtraction. In this paper, we propose a novel automatic foreground extraction algorithm by wavelet multi-scale transformation based on background subtraction. The proposed algorithm produces better foreground segmentation without the need to determine an optimum threshold for each image with different clutter background by experiments.

The 2-D dyadic wavelet transformation (WT) of an image $I(x,y)$ at scale 2^j and in orientation k is defined as [10]:

$$W_{2^j}^k f(x,y) = I * \psi_{2^j}^k (x,y), k = 1,2 \qquad (1)$$

If we want to locate the positions of rapid variation of an image I, we should consider the local maxima of the gradient magnitude at various scales which is given by

$$M_{2^j} I(x,y) \equiv \left\| \nabla_{2^j} I(x,y) \right\| = \sqrt{\left(W_{2^j}^1 I(x,y)\right)^2 + \left(W_{2^j}^2 I(x,y)\right)^2} \quad (2)$$

A point (x, y) is a multi-scale edge point at scale 2^j if the magnitude of the gradient $M_2^j I$ attains a local maximum there along the gradient direction $A_2^j I$.

For each scale, we can collect the edge points together with the corresponding values of the gradient. The resulting local gradient maxima set at scale 2^j is

$$P_{2^j}(I) = \left\{ p_{2^j,i} = (x_i, y_i); \nabla_{2^j} I(x_i, y_i)) \right\} \qquad (3)$$

where $M_2^j I(x_i, y_i)$ has local maximum at $P_2^j{}_{,i}$=(x_i, y_i) along the direction $A_2^j I(x_i, y_i)$. For a J-level two-dimensional dyadic WT, the set

$$\rho(I) = \left\{ S_{2^J} I(x,y), \left\lfloor P_{2^j}(I) \right\rfloor_{1 \le j \le J} \right\} \qquad (4)$$

is called a multi-scale edge representation of the image $I(x, y)$. Here $S_2^j I(x, Y_i)$ is the low-pass approximation of $I(x, y)$ at the coarsest scale 2^J. The result of foreground segmentation by proposed algorithm is shown as Figure 1.

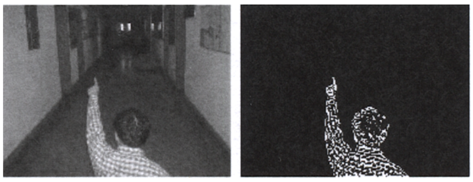

Figure 1 Foreground segmentation. The current original image and segmented foreground from left to right, respectively.

In some cases, obtained foreground has misses or splits. In order to obtain the correct foreground region, we use an algorithm called as filter close operation. The close mask region is performed close operation. The final foreground extraction is shown in Fig. 2

Figure 2 Final segmented binary foreground with close mask.

3 Pointing gesture analysis

In order to detect pointing gesture we need to detect and track the 3D positions of user's head and the pointing fingertip. Based on the extracted object regions, we adopt coarse to fine strategy to segment part of pointing hand and head. This feature makes our approach different from other approaches proposed recently [3,5,11,14]. We do not constrain the user's hands to be exposed for detecting by skin color, which is necessary for some hand pointing recognition system as mentioned in [5,11,14].

Our goal is to recognize human arm-pointing gestures in an actual environment. Based on the position of the user pre-defined with respect to the camera set-up, which implies that the user's head would be captured by the cameras, we take the centriods of user's head top in different views as corresponding points directly. The choice of centroids for correspondence strategy implies the low computation complex. The key problem is how to solve the correspondence between the pointing fingertips in different views. We propose a stereo matching strategy to solve the correspondence of fingertips based on multi-geometrical constraint relationship between fingertips, the axis of pointing arm, and epipolar line.

Since the extracted pointing hand is quite salient, we can determine the axis of arm and extract fingertip point easily. Assuming that the axis of arm is extracted, based on the extracted fingertip points and calibrated stereo rig, we associate *the axis of arm* with epipolar line to obtain the correspondence points. According to continuity constraint, we can determine correct correspondences.

Given a calibrated stereo rig and two corresponding points of pointing fingertip and centroids of head top, it is straightforward to reconstruction 3D positions of the fingertip and head top in the world frame.

Since we are not really interested in the position of the head top, but rather that of the eyes, we estimate the latter as that of the former lower by an amount of 8 cm vertically.

4 Experimental results

Our system is implemented in C++ and runs in real-time (50Hz) on a Pentium Ⅳ PC 600 MHz running windows. Two image capture cards at a resolution of 320×240 are used for image acquisition. The experimental setup includes two projects.

Twelve intended panels located at rough 3.5m to 12m away from cameras with a maximum size of 25cm×20cm are placed on our laboratory corridor ground. We take normal lighting environments for this project. The test for the project was carried out by 36 volunteers with different physical characteristics and pointing styles, altering during interaction. All users point their left, and right hands. There are a total of 864 pointing gestures (12panels × 2 hands × 36 users). The result of recognition is shown in Table 1.

Table 1 Results of recognition

Panel	Location (mm)			Recognition rate (%)		
	Xc	Yc	Zc	Right hand	Left hand	Total
1	-1400	125	3500	97.5	97.1	97.3
2	0	125	3500	97.7	97.3	97.5
3	1400	125	3500	97.6	97.2	97.4
4	-1400	125	6000	96.5	96.1	96.3
5	0	125	6000	96.8	96.3	96.5
6	1400	125	6000	96.6	96.1	96.3
7	-1400	125	8500	95.8	95.4	95.6
8	0	125	8500	95.2	94.8	95.0
9	1400	125	8500	95.8	95.5	95.6
10	-1400	125	11000	93.4	93.2	93.3
11	0	125	11000	93.8	93.4	93.6
12	1400	125	11000	93.5	93.2	93.3

One may notice that the precision decreases as users aim target from long-distance to short-distance. The error of camera calibration is enlarged as distance augmenting; on the other hand, user's eyesight becomes dim as distance increasing, which makes the precision of the line of sight connecting the eyes and the fingertip decrease.

Another project is an example in a multimedia exhibition hall with clutter background. The user points his arm at one of eleven intended models with sizes of from 50cm×50cm to 2m×1.5m locating at rough 3m to 12m away from user. When user moves his arm to another model, the power of previous model is turn off and the pointing model is turning on. There are 85 volunteers to test the accuracy of pointing gesture recognition. The visual feedback of light seems to be accurate to the user. The accuracy is 95.6% on an average. One view of experimental setup is shown in Figure 4.

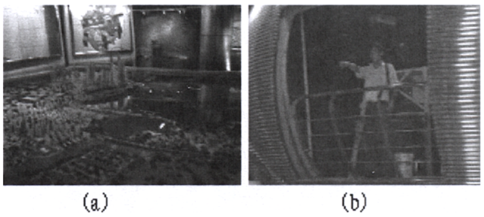

Figure 4 Example of display in a multimedia exhibition hall. (a) Portion of the scene and its selectable regions. (b) Interactive controlling a frontal model by pointing.

The whole running time for the two projects takes about

150ms on an average on a PⅣ600 MHz 512RAM.

5 Conclusions and further work

We described a fast and robust algorithm to recognize arm pointing gesture in a real-time from binocular views, which is not necessary to know beforehand the dominant hand of user. It is both suited for right-handed or left-handed users. Any user can start performing actions by pointing at any direction without being forced to wear special clothes or markers. Furthermore, it is obvious that the tracking system is usable without user specific training phase. The user is able to interact with the display without reading operating instruction first. In the first experiment, we evaluate the performance of methods developed for multi-objects with relatively constant background. In the second experiment, we show an actual application of pointing gesture to control model power with clutter background. The system works in real time on a low cost hardware platform. It is fairly accurate and independent of user characteristics and environmental changes. Further work will be especially devoted to extend system operation to the management of several users. To this end, more sophisticated tracking algorithms capable of dealing with severe occlusion conditions will be developed, allowing two or more users to share the same interaction space and also to interact together through a computer support based on natural gestures.

References

[1] Baudel T, Lafon M.B. "Charade: remote control of objects using free-hand gestures", *Communications of the ACM*, Volume 36, No.7, pp.28–35, (1993).

[2] Carbini S, Viallet J.E, Bernier O. "Pointing gesture visual recognition by body feature detection and tracking", *Proc. International Conf. Computer Vision and Graphics*, Varsovie, Poland, pp.203-208, (2004).

[3] Colombo C, Bimbo A.D, Valli A. "Visual capture and understanding of hand pointing action in a 3-D environment", *IEEE Trans. System, Man, and Cybernetics–part B: Cybernetics*, Volume 33, No.4, pp. 677-686, (2003).

[4] Crowley J.L. "Vision for man-machine interaction", *Robotics and Autonomous Systems*, Volume 19, No.3, pp.347–358, (1997).

[5] Hild M, Hashimoto M, Yoshida K. "Object recognition via recognition of finger pointing actions", *Proc. 12th International Conf. on Image Analysis and Processing*, Mantova, Itlay, pp.88-93, (2003).

[6] Kahn R.E, Swain M.J. "Understanding people pointing: the Perseus system", *Proc. International Symposium on Computer Vision*, Coral Gables, USA, pp.569-574, (1995).

[7] Kehl R, Gool L.V. "Real-time pointing gesture recognition for an immersive environment", *Proc, 6th IEEE International Conf. Automatic Face and Gesture Recognition*, Seoul, South Korea, pp.577- 582, (2004).

[8] Kolesnik M, Kulessa T. "Detecting, tracking and interpretation of a pointing gesture by an overhead view camera", *Proc. DAGM Symposium*, Munich, Germany, Volume 2191, pp.429-436, (2001).

[9] Maggioni C, Röttger H. "Virtual touchscreen – a novel user interfac e made of light – principles, metaphors and experiences", *Proc. 8th International conf. Human-Computer Interaction*, Munich, Germany, September, Volume 1, pp. 301-305, (1999).

[10] Mallat S, Hwang W.L. "Singularity detection and processing with wavelets", *IEEE Trans. Information Theory*, Volume 38, No. 2, pp.617-643, (1992).

[11] Nickel K, Stiefelhagen R. "Real-time person tracking and pointing gesture recognition for human-robot-interaction", *Proc. HCI* 2004, Prague, Czech Republic, pp.28-38, (2004).

[12] Segen J, Kumar S. "Human-computer interaction using gesture recognition and 3D hand tracking", *Proc. International Conf. Image Processing*, Chicago, US, pp.188-192, (1998).

[13] Wellner P. "Interacting with paper on the DigitalDesk", *Communications of the ACM*, Volume 36, No.7, pp.86–96, (1993).

[14] Yamamoto Y, Yoda I, Sakaue K. "Arm-pointing gesture interface using surrounded stereo cameras system", *Proc. 17th International Conf. Pattern Recognition*, Cambridge, UK, Volume 4, pp.965-970, (2004).