Binocular Hand Tracking and Reconstruction Based on 2D Shape Matching

Antonis A. Argyros and Manolis I.A. Lourakis
Foundation for Research and Technology – Hellas (FORTH)
Institute of Computer Science (ICS)
Heraklion, Crete, Greece
{argyros, lourakis}@ics.forth.gr

Abstract

This paper presents a method for real-time 3D hand tracking in images acquired by a calibrated, possibly moving stereoscopic rig. The proposed method consists of a collection of techniques that enable the modeling and detection of hands, their temporal association in image sequences, the establishment of hand correspondences between stereo images and the 3D reconstruction of their contours. Building upon our previous research on color-based, 2D skin-color tracking, the 3D hand tracker is developed through the coupling of the results of two 2D skin-color trackers that run independently on the two video streams acquired by a stereoscopic system. The proposed method runs in real time on a conventional Pentium 4 processor when operating on 320x240 images. Representative experimental results are also presented.

1. Introduction

Tracking body parts such as hands and faces of humans engaged in various activities constitutes a key pre-processing step for parsing visual input into constituent behavior elements towards the automatic interpretation of the performed activities. It is the importance and the difficulty of providing robust solutions to this problem without resorting to unrealistic assumptions that justify the volume of past and ongoing related research.

Vision-based methods for tracking hands and reconstructing them in 3D need to provide solutions to subproblems such as the modelling and the detection of hands, the temporal association of detected hands in image sequences and the extraction of 3D information from the inherently 2D observations of the tracked models. As stated in [10], most approaches to 3D hand tracking can be classified as either model-based or as view-based. Model-based approaches make use of articulated 3D hand models. An error function measures the discrepancy between the projection of the hand model onto the image and the image-based evidence for the existence of a hand. The parameters of the model are then appropriately modified to minimize this error measure. Temporal continuity is also exploited by considering previously estimated model parameters as an initial solution for the current frame. As a side effect, model initialization is necessary for the first frame. View-based approaches [7] formulate hand tracking as a classification problem. These methods train a classifier to recognize a limited number of selected hand poses. The training data are formed by associating detected hand features at particular hand poses with these hand poses. Hybrid methods combining both approaches have also been proposed [8]. These methods use 3D models to generate an arbitrarily large training set of 2D hand appearances by projecting a 3D model on the 2D image. View-based classification is then used to solve the inverse problem (i.e. determining the mapping from 2D detected hands to 3D hand poses).

In this paper, we present our approach to 3D hand tracking and reconstruction. The proposed method provides 3D information regarding the tracked hands either in the form of 3D hand positions or in the form of 3D hand contours. Compared to existing approaches, this method for detecting and tracking hands combines several attractive properties:

- *Handling multiple, potentially occluding hands:* Many of the existing techniques deal with the detection and tracking of a single hand. Moreover, only a subset of existing techniques for multiple hand tracking can cope effectively with occluding hands. The proposed method detects and tracks an arbitrary, time varying number of hands that may move in occluding trajectories.
- *Coping with cluttered, dynamic backgrounds and camera motions:* Many of the existing approaches assume a uniform and/or static background to facilitate figure-ground segmentation. The assumption of a static background does not allow for moving cameras, which might be useful for purposefully observing a certain activity. The proposed method may operate in cluttered backgrounds using a moving stereo system under considerable illumination changes.
- *Real time performance:* The complex 3D models employed by certain methods require considerable computational resources to implement tracking.

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

The proposed method computes the 3D trajectories of hands at a rate of 29 Hz, and their 3D contours at a rate of 21 Hz.

The rest of the paper is organized as follows. Section 2 describes the proposed approach in detail. Section 3 provides sample results from the operation of the tracker on binocular image sequences. Finally, section 4 provides the main conclusions of this work.

2. 3D hand tracking and reconstruction

As already mentioned, the proposed method is based on our previous work on 2D tracking of multiple skin colored objects [1]. In the following section, we provide a brief overview of our 2D tracker.

2.1. The 2D hand tracker

The employed 2D tracker [1] encompasses a collection of techniques that enable the modeling and detection of skin-colored objects as well as their temporal association in image sequences. The output of the 2D tracker is a number of blobs corresponding to the tracked hands, together with their 2D contours. A prototype implementation of the 2D tracker achieves real-time performance on a Pentium 4 processor (3.2 MHz, 512Mbytes RAM, Windows XP), limited mainly by the maximum image acquisition rate of the employed cameras which is 30 fps.

2.2. Matching hand blobs in stereo

Each of the video streams delivered by the synchronized cameras of a stereoscopic system is processed by an instance of the previously described 2D hand tracker. We denote by H_L and H_R the sets of hand blobs that have been detected and tracked at time *t* in the left and right images of the stereo pair. Let the cardinality of these sets be N_L and N_R, respectively. Moreover, let h_L^i and h_R^j denote specific hand blobs, with $1 \le i \le N_L$ and $1 \le j \le N_R$. We also designate the 2D centroids of hands h_L^i and h_R^j with m_L^i and m_R^j, respectively.

In order to achieve 3D reconstruction of the hand positions and hand contours, correspondence between hand blobs needs to be established. We formulate this problem as an instance of the *stable marriage problem*, which can be defined as follows. Consider two disjoint sets *A* and *B*. Assume that each of the members of *A* ranks the members of *B* in order of decreasing preference, and that the members of *B* do likewise. A set of pairings of the members of *A* with the members of *B* is said to constitute a stable marriage if and only if there exist no elements $a \in A$ and $b \in B$ which are not assigned to each other but would both prefer each other to their present partners. Gale and Shapley [5] showed that a stable marriage exists for any choice of rankings. The original formulation of the problem assumes that sets *A* and *B* have equal cardinalities. In our case, the roles of sets *A* and *B* are assumed by the sets H_L and H_R. Since it may well hold that $N_L \ne N_R$, we have extended [5] to handle the case of sets with unequal cardinalities.

The required preferences among the members of sets H_L and H_R are formed by employing the epipolar constraint [11]. More specifically, m_R^j is constrained to lie on the epipolar line l_R=$F\,m_L^i$, where *F* denotes the fundamental matrix of the stereo system. Similarly, m_L^i is constrained to lie on the epipolar line l_L=$F^T m_R^j$. A distance measure $D_{LR}(h_L^i, h_R^j)$ of h_L^i from h_R^j can be defined as $D_{LR}(h_L^i, h_R^j) = d\left(Fm_L^i, m_R^j\right)$ where $d\left(l, p\right)$ denotes the Euclidean distance of point *p* from line *l*. Similarly, $D_{RL}(h_R^j, h_L^i) = d\left(F^T m_R^j, m_L^i\right)$. By sorting these distances in ascending order, we may define preferences of left hand blobs to right hand blobs and vice versa. In other words, the larger the distance $D_{LR}(h_L^i, h_R^j)$, the lower is the preference of h_L^i to h_R^j. The preferences of all h_R^j s to all h_L^i s are defined similarly based on $D_{RL}(h_R^j, h_L^i)$. Furthermore, our extension to [5] for solving the stable marriage problem rejects a candidate pair if $D_{RL}(h_R^j, h_L^i)$ or $D_{LR}(h_L^i, h_R^j)$ is above a certain threshold.

Still, there may be cases where the solution to the stable marriage problem with preferences as defined above fails to result in correct matches. As an illustrative example, assume two different hands, each of which is visible in only one of the two images of the stereo pair. Assume also that incidentally, hands are located in such a way that eqs.(1) and (2) give rather small distances (i.e. high preferences). The solution to the stable marriage problem will then provide a wrong match between these two hands. To cope with such situations, we reject matches that yield an impossible or a highly unlikely 3D hand position after 3D reconstruction (i.e. negative depth value or depth value outside reasonable bounds).

The computation of the distances D_{LR} and D_{RL} requires that the fundamental matrix *F* of the stereo configuration is known. Assuming a fully calibrated stereo camera system, *F* can be computed analytically as $\mathbf{F} = A_r^{-\mathbf{T}} R S A_l^{-1}$. In this equation, A_L and A_R are the intrinsic calibration matrices of the two cameras, R is the rotation matrix associated with the relative positions of the cameras of the stereo rig and S is a rank deficient matrix dependent on the relative translation between the two cameras.

2.3. Shape matching through contour alignment

The algorithm described in the previous section matches hands between the images of a calibrated stereo-pair. To recover the 3D contour of a hand correspondences of contour pixels are also required.

The image of a hand is typically a small, textureless area in front of a relatively distant background. The lack of texture and the presence of considerable depth discontinuities are conditions that do not favor correlation-based approaches towards solving the correspondence problem. Instead, we compute correspondences through a top-down approach. The basic idea in this approach is that if two matching hand contours[1] can be aligned, then the necessary correspondences for 3D reconstruction can easily be extracted. To perform this type of alignment, we employ a robust variant of the *Iterative Closest Point* (ICP) algorithm. Given two point sets, *P* and *M*, the task of ICP is to find the affine motion that brings *P* into the best possible alignment with *M*. The original ICP algorithm [3] consists of the following three steps: (1) pair each point of *P* to the closest point in M, (2) compute the motion that minimizes the mean square error (MSE) between the paired points and, (3) apply the motion to *P* and update the MSE. These three steps are iterated and have been proved to converge in terms of the MSE.

Several robust variants of the ICP algorithm have been proposed that can solve the problem in the presence of measurement outliers and, possibly, shape defects. In our hand tracking scenario, such outliers and shape defects can be due to inaccuracies in skin color detection. Filtering them out is very important because it safeguards the later process of 3D reconstruction against gross errors due to wrong point matches. The robust variant of ICP that we employ is similar in spirit with the one described in [4]; the major difference is that we use the Least Median of Squares (LMedS) robust estimator [9] in all steps of the general ICP algorithm, instead of the Least Trimmed Squares (LTS) estimator of [4].

The initial contour alignment that is necessary for bootstrapping the ICP algorithm is easily achieved through information already available from the 2D hand tracker. More specifically, assume that the pixels belonging to hand h_L^i have a covariance matrix Σ_i and that the pixels of the corresponding hand h_R^j in the right image have a covariance matrix Σ_j. Considering the two ellipses defined by equations $p^T \Sigma_i p = 1$ and $q^T \Sigma_j q = 1$, the first can be transformed to the second through a linear transformation $q = Vp$, where V is such that $\Sigma_j = V \Sigma_i V^T$. The recovery of the transformation V and its application to the contour points of hand h_L^i results in the approximate alignment of the contour points of h_L^i with those of hand h_R^j.

2.4. 3D reconstruction

The recovery of the 3D coordinates of the matched hand centroids and the matched hand contour points can be achieved through standard triangulation. Assume a 3D point *M* projecting onto points m_L^i and m_R^j on the left and right images of the stereo-pair. *M* is constrained to lie on the 3D line defined by C_L and M_L where C_L is the 3D location of the optical center of the left camera and M_L is the 3D location of point m_L^i. Similarly, *M* is constrained to lie on the 3D line defined by C_R and M_R where C_R is the 3D location of the optical center of the right camera and M_R is the 3D location of point m_R^i. In the case of noiseless measurements these two lines intersect at *M*. However, noise in the estimation of m_L^i and m_R^j and inaccuracies in camera calibration will almost certainly result in these two lines being skew. For this reason, *M* is taken to be the midpoint of the minimum-length line segment whose endpoints lie on the two 3D lines $M_L C_L$ and $M_R C_R$ [6].

3. Experiments

The developed method for binocular hand tracking and reconstruction has served as a building block in a number of applications and has been tested extensively in various environments. More specifically, it has been employed as a component of a cognitive vision system whose goal is the interpretation of the activities of people handling tools. Fig. 1(a) shows a snapshot from a related experiment in which a person operates a CD player while a pair of cameras is observing the scene. Detected skin-colored pixels are illustrated in white. The contour of the hand is shown in light blue. The index finger has also been detected based on a method

[1] Note that the hand correspondence algorithm of section 2.2 permits us to know which contours to align.

proposed in [2] and is marked with a black dot. Figure 1(b) shows the computed 3D trajectory of the hand centroid. As it can be verified in this figure, the hand moves towards the CD player, opens the tray, moves towards the CD, takes the CD, puts it on the open tray, closes the tray and retracts to its rest position. For improving the readability of the 3D plot, the CD player has also been reconstructed in 3D. Reporting the 3D position of the hand is achieved at 29 fps on a conventional, Pentium 4 processor.

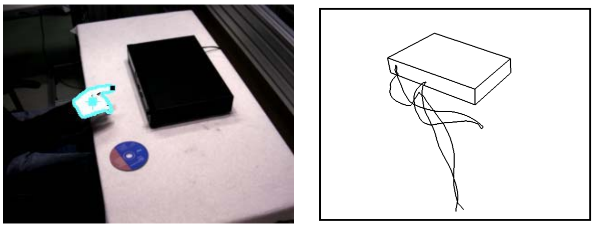

(a) (b)

Figure 1: (a) Left camera snapshot from a 3D hand tracking experiment, (b) the computed hand trajectory in 3D.

Figures 2(a),(b) show the left and right images of another stereo pair. In this experiment, several hands are successfully detected and tracked among images. Figure 2(c) shows the 3D reconstruction of the contours of these hands.

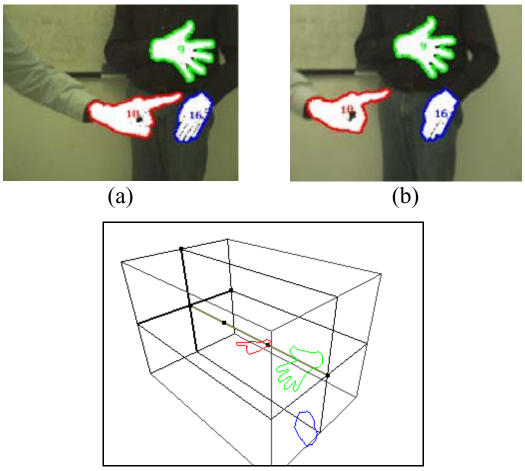
The developed method has also been used in the context of a system for vision-based human computer interaction [2]. Simple gesture recognition techniques applied to the output of the 3D hand tracker have resulted in a system that permits a human to remotely control the mouse pointer of a computer. Videos from related experiments can be found at http://www.ics.forth.gr/~argyros/research/virtualmouse.htm.

4. Summary

In this paper, we have proposed a new method for binocular hand tracking and reconstruction. The proposed tracker is able to report the 3D positions and the 3D contours of several hands that move, possibly occluding each other, in the field of view of a potentially moving binocular camera system. The developed method has several attractive features, such as robustness in illumination changes and real time performance. Because of those, it has served as a building block in many applications.

Acknowledgements

This work has been partially supported by EU IST-2001-32184 project ActIPret and EU-IST NoE MUSCLE (FP6-507752).



(a) (b)

(c)

Figure 2: (a), (b) a stereo-pair from a hand tracking experiment, (c) the estimated 3D hand contours. The origin of the coordinate system is at the far left face of the parallelepiped.

References

[1] A.A. Argyros, M.I.A. Lourakis, “Real time Tracking of Multiple Skin-Colored Objects with a Possibly Moving Camera”, in proc. ECCV’04, Springer-Verlag, vol. 3, pp. 368-379, Prague, Czech Republic, May 11-14, 2004

[2] A.A. Argyros, M.I.A. Lourakis, “Vision-based Interpretation of Hand Gestures for Remote Control of a Computer Mouse”, in proc. HCI’06 workshop, Springer-Verlag, pp.40-51, Graz, Austria, May 13, 2006

[3] P. Besl, N. McKay, “A Method for Registration of 3-D Shapes”, IEEE Trans. PAMI, Vol. 14, No. 2, 1992.

[4] D. Chetverikov, D. Svirko, D. Stepanov, P. Krsek, “The Trimmed Iterative Closest Point Algorithm”, In Proc. ICPR’02, Quebec, Canada, 2002.

[5] D. Gale, L.S. Shapley, “College Admissions and the Stability of Marriage”, American Mathematical Monthly, 69:9, 1962.

[6] R. Goldman, “Intersection of Two Lines in Three-Space”, In Graphics Gems I (Ed. A. S. Glassner), San Diego, Academic Press, 1990, pp. 304.

[7] V. Pavlovic, R. Sharma, T. Huang, “Visual interpretation of hand gestures for human computer interaction: A review”, IEEE T-PAMI, 19(7), 1997.

[8] R. Rosales, V. Athitsos, L. Sigal, S. Sclaroff, “3D hand pose reconstruction using specialized mappings”, In Proc. ICCV’01, vol. I, pp. 378.385, Canada, July 2001.

[9] P. J. Rousseeuw, “Least median of squares regression”, Journal of American Stat. Assoc., 79: 871–880, 1984.

[10] B. Stenger, “Model-Based Hand Tracking Using A Hierarchical Bayesian Filter”, PhD Thesis, Department of Engineering, University of Cambridge, March 2004.

[11] Z. Zhang, “Determining the Epipolar Geometry and its Uncertainty: a Review” Int. Journal of Computer Vision 27, 161–195, 1998.