Marker-less Detection of Virtual Objects using Augmented Reality

Prakhar Kapoor
University of Massachusetts Lowell, USA
prakhar_kapoor@student.uml.edu

Usama Ghufran
Dreamworks Dedicated Unit, India
usamazeb@gmail.com

Manan Gupta
Accenture Services Pvt Ltd, India
Manan.gupta@accenture.com

Alok Agarrwal
Dept. of CSE, JPIET, Meerut, India
alok289@yahoo.com

Abstract
We present marker less camera tracking and user interface methodology for readily inspecting augmented reality (AR) objects in wearable computing applications. Instead of markers, human hand is used as a distinctive pattern that almost all wearable computer users have readily available. We present a robust real-time algorithm that recognizes fingertips to reconstruct the six-degree-of-freedom camera pose relative to the user's outstretched hand. A hand pose model is constructed in a one time calibration step by measuring the fingertip positions in presence of ground-truth scale information. Through frame-by-fame reconstruction of the camera pose relative to the hand, we stabilize 3D graphics annotations on top of the hand, allowing the user to inspect such virtual objects conveniently from different viewing angles in AR. We evaluate our approach with regard to speed and accuracy and compare it to state-of-art marker based AR systems. We demonstrate the robustness and usefulness of our approach in an example AR application for selecting and inspecting world-stabilized virtual objects.

Key words: augmented reality, marker-less detection, marker-based detection

1. Introduction

Human motion capture can be done using two ways namely marker-based and marker-free or marker-less method. In marker-based methods, actor wears marker on each joint so as to identify the motion by the positions or angles between the markers. Markers are of various types like acoustic, inertial, LED, colored markers, magnetic or reflective markers etc. which are tracked optimally at least two times the rate of the desired motion to sub-millimeter positions. The motion capture computer software records the positions, angles, velocities, accelerations and impulses, providing an accurate digital representation of the motion. Marker-based capture systems are quite popular due to efficiency and accuracy but are highly costly, require laboratory setup and restrict the movement of the actor. Another kind of motion capture systems employ marker-free methods which do not require markers of any kind and are built upon the concepts of computer vision offering high degree of freedom to the actor.

Augmented reality is a field of computer research which aims at supplementing reality by mixing computer-generated data and real world environments. Structural coherence is needed to deceive the human eyes and make virtual objects appear and behave as they would really exist in the scene. The contextual coherence helps in providing selected information according to the user needs, without flooding him plenty of undesired data. The real-world elements are the football field and the players while the virtual elements are the score numbers and the team flags which are drawn over the image by computers in real time, stretched as they would appear if laid on the court. Most of the AR applications can be dramatically boosted using see-through display glasses or other special visual devices. For instance, driving assistance applications, inside cars or airplanes, make up for this by deploying head-up displays integrated into the windshield.

In [1] authors have presented a framework for tracking human motion in an indoor environment from sequences of monocular grayscale images obtained from multiple fixed cameras. Multivariate Gaussian models are applied to find the most likely matches of human subjects between consecutive frames taken by cameras mounted at various locations. Experimental results from real data show the robustness of the algorithm and its potential for real time applications. In [2] a robust camera pose estimation method is proposed based on tracking calibrated 2D fiducials in a known 3D environment. To efficiently compute the camera pose associated with the current image, results of the fiducials are combined with the Orthogonal Iteration (OI) Algorithm. In [3] authors have discussed face detection algorithms on the basis of skin colour. Colour

© 2013. The authors - Published by Atlantis Press

spaces RGB, YCbCr and HIS have been used by their algorithm which detects human faces on an average with a 95.18% accuracy. In [4] human skin regions have been detected using Bayes rule in color images. To avoid the effect of brightness included in the RGB color space, skin color in the chromatic and pure color space YCrCb which separates luminance and chrominance components have been proposed. In [5] authors have proposed a closed-form solution to calibrate a camera followed by a nonlinear refinement based on the maximum likelihood criterion. Compared with classical techniques which use expensive equipment such as two or three orthogonal planes, the proposed technique is easy to use and flexible and advances 3D computer vision one more step from laboratory environments to real world use. In [6] authors have developed a homography-based adaptive visual servo controller to enable robot end-effectors to track a desired Euclidean trajectory as determined by a sequence of images for both the camera-in-hand and fixed-camera configurations. In [7] a closed-form least-squares solutions have been proposed to the overconstrained 2D-2D and 3D-3D pose estimation problems. A globally convergent iterative technique is given for the 2D-perspective-projective-projection-3D pose estimation problem. The experimental results show that the robust technique can suppress the blunder data which come from the outliers or mismatched points. In [8] authors have presented a novel robust camera pose estimation algorithm based on real-time 3D model tracking. A non-linear optimization method is used to estimate the camera pose parameters. Robustness is obtained by integrating a M-estimator into the optimization process. In [9] tracking is addressed and a real-time, robust, and efficient 3D model based tracking algorithm is proposed for a "video see through" monocular vision system. The proposed method has been validated on several complex image sequences including outdoor environments.

Rest of the paper is organized as follows. Image processing, analysis & design used in the proposed work are discussed in Section 2 and 3 respectively. Future prospects of the work are discussed in Section4. Finally Section 5 concludes the work of the paper.

2. Image Processing

The goal of virtual keyboard is to detect the movements of hands or fingers and handle the keyboard not through other external devices but with the help of captured movements of hands such that the response time is very small and handling is less complicated. With the external or internal webcam the hands are displayed on the screen and corresponding movements are detected and the buttons are pressed.

However as soon as the camera is turned on, the background is set pressing the 'B' key which examines the background conditions and sets it and after that the movements of hands are brought in front of the camera and with corresponding changes in the frame the movements are detected and background subtraction takes place. Lighting conditions play an integral part in this concept.

A) Background Subtraction

The background is set and initially the frame is fixed and then the hand is brought in front of the virtual keyboard i.e. in the frame to detect the movements. Then the current frame incorporating the detected movements of the hands is compared with the previous frame.

B) Motion Detection

Motion can be detected by measuring changes in speed or vector of an object or objects in the field of view. This can be achieved either by mechanical devices that physically interact with the field or by electronic devices that quantify and measure changes in the given environment.With the changes from the initial frame the movements have been detected and captured. The image captured every time is overwritten and the latest image is displayed every time.

C) Contour Making

When the hands are brought in front of the camera the corresponding contours are drawn by detecting the RGB values. The contours are made for any object in the frame and by detecting the longest edge in the contour the graph is plotted.

D) Centroid

The centroid is calculated by calculating the distances between the points on the hand (x, y) and (x_1, y_1) that is the maximum points.

$$Z_n = \sqrt{(x_1-x)^2 + (y_1-y)^2}$$

where Z_n is the distance calculated.
For the five fingers we get the corresponding co-ordinates and calculate the centroid.

Centroid= $(z_1+z_2+z_3+z_4+z_5)$ / total number of pixels.

E) Skin Detection

Skin Detection takes place for the particular value of RGB. By doing skin detection we made contours across the hands and fingers which was previously done for the whole body. By doing skin detection accuracy was enhanced as only that part was detected for which movements were needed.

F) Training or pose estimation
Initially the hands are kept in a certain position and this gives us our starting point, With the change in the position of the hand the x, y and z coordinate changes and so many more poses can be trained for the movement of the hand. The object tends to revert if the hand is reversed and tends to disappear if that pose is made. The virtual object also increases and decreases if it is brought close or taken far simultaneously.

G) Hand Model Construction
The hand pose is estimated and a corresponding homogeneous matrix is constructed which has poses of the objects stored in it and it changes with the changes in the position of the object. The matrix we get is a model matrix multiplied with a view matrix.

H) Ellipse Fitting
With the largest finger and the thumb we approximate an ellipse by taking the distance relation of them. The object is approximated on top of the ellipse.

3. Analysis and Design

Considerably less information is to be provided about the hand. Some features such as the finger against a background of skin would be very hard to distinguish since no depth information would be recoverable. Essentially only "silhouette" information could be accurately extracted. The silhouette data would be relatively noise free given a background sufficiently distinguishable from the hand and would require considerably less processor time to compute than either multiple camera system. It is possible to detect a large subset of gestures using silhouette information alone and the single camera system is less noisy, expensive and processor hungry. Although the system exhibits more ambiguity than other systems, this advantage is more than outweighed by the advantages mentioned above. The output of the camera system comprises of a 2D array of RGB pixels provided at regular time intervals. To detect silhouette information it is necessary to differentiate skin from background pixels.

The task of differentiation of the skin pixels from those of the background and markers is made considerably easier by a careful choice of lighting. If the lighting is constant across the camera then the effects of self-shadowing can be reduced to a minimum. The intensity should also be set to provide sufficient light for the CCD in the camera. An attempt is made to extract the hand and marker information using standard room lighting, in this case a 100 watt bulb and shade mounted on the ceiling. This permits the system to be used in a non-specialist environment.

The two realistic options of camera orientation are to point the camera towards a wall or towards the floor (or desktop). Light intensity would be higher and shadowing effects least if the camera is pointed downwards. It is also desirable that the colour of the background differs as much as possible from that of the skin. The floor color in this work is a dull brown.

A low cost computer vision system that can be executed in a common PC equipped with an USB web cam is one of the main objectives of the proposed work. The system is able to work under different degrees of scene background complexity and illumination conditions, which shouldn't change during the execution. Steps used are briefly discussed below:

Initialization: the recognizable postures are stored in a visual memory which is created in a start-up step. In order to configure this memory, different ways are proposed.

Acquisition: a frame from the webcam is captured.

Segmentation: each frame is processed separately before its analysis; the image is smoothed, skin pixels are labeled, noise is removed and small gaps are filled. Image edges are found, and finally, after a blob analysis, the blob which represents the user's hand is segmented, a new image is created which contains the portion of the original one where the user's hand was placed.

Pattern Recognition:
Executing Action: finally, the system carries out the corresponding action according to the recognized hand posture.
Figure 1 shows the overview of the system.

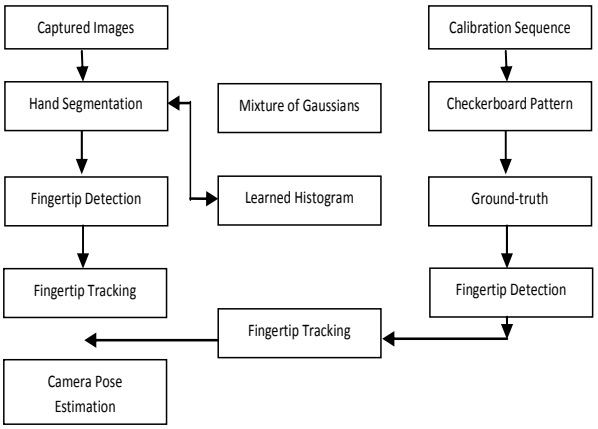

Figure 1: Overview of the system

Hand segmentation using skin Detection:

Modeling skin colour requires the selection of an appropriate colour space and identifying the cluster associated with skin colour in this space. Skin detection is applied so as to allow only the use of the human hand. One simple and classical RGB color space-based classifier is used. It takes two different conditions (involving strict thresholds) into account: uniform daylight and flash or lateral illumination, as presented in sets of equations in (1) and (2).

Uniform daylight illumination: (1)

R> 95, G> 40, B> 20,

Max {R,G,B} – Min {R,G,B} > 15,

[R – G] > 15, R>G, R>B.

Flashlight or daylight lateral illumination : (2)

R> 220, G> 210, B> 170,

[R-G] = 15, B<R, B.G.

Fingertip Detection and Tracking:

Using Convex Hull fingertips from the contour are obtained. In mathematics, the convex bull or convex for a set of points X in a real vector space V is the minimal convex set containing X. In Computational geometry, a basic problem is finding the convex hull for a given finite nonempty set of points in the plane. The convex hull is then typically represented by a sequence of the vertices of the line segments forming the boundary of the polygon, ordered along that boundary. For planar objects, i.e., lying in the plane, the convex hull may be easily visualized by imagining an elastic band stretched open to encompass the given object; when released, it will assume the shape of the required convex hull.

Every objects location can be depicted by making a matrix plotting its x and y co-ordinates. The marker has a pose captured on it. That pose is same as that of the object. In short, marker has an objects pose captured on it. That object's matrix is made considering its initial pose and its movements is the x,y and z plane. The camera tracks the objects pose and displays it. For tracking the pose, Open GL has been used and its inbuilt functions have been utilized.

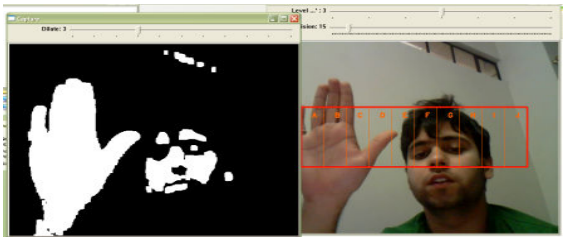

Figure 2(a): Skin detection of face and hand

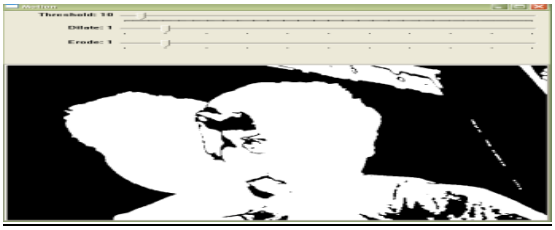

Figure 2(b): Motion Detection with changed background

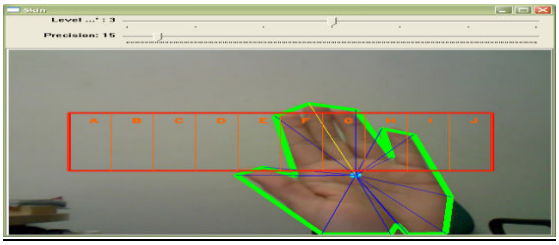

Figure 2(c): Perfect Skin detection on hand with virtual keyboard

Figure 2(d): Perfect Skin detection on hand with virtual keyboard

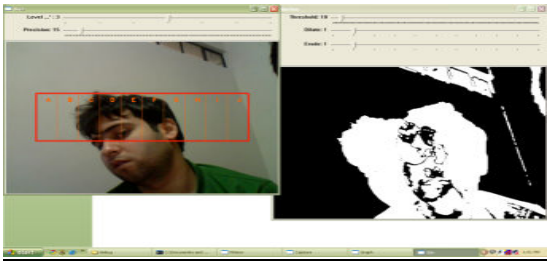

Figure 2(e): Motion Detection displayed on the right window with grayscale image

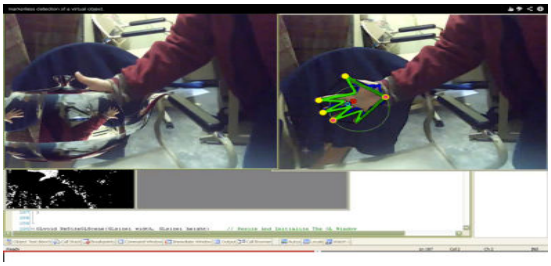

Figure 2(f): Moving the kettle to various positions with the hand

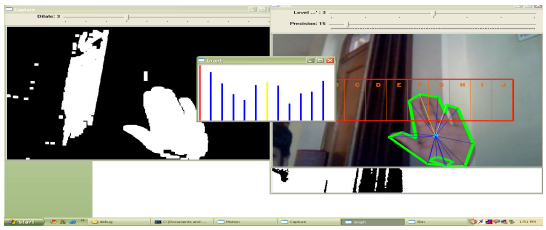

Figure 2(g): Keyboard and displaying the graph for movement

4. Future work

HMDs can be used that would enhance the user interface and make Augmented Realty more realistic. It could also be made more collaborative as multiple kids/users could collaboratively play and build in the same work/play space. As an enhancement we can make the detection marker less in the future. Better hand gesture recognition techniques can be developed. For Multi-stage gestures, it could be possible to represent a much larger number of labels if each label consisted of two or more gestures combined with hand position changes. Development of better user interfaced for the Augmented World Objects.

5. Conclusion

Our tests and experiments indicate that the hand segmentation is somewhat sensitive to changes in illumination, even though our adaptively learned color model helps robustness a great deal. In outdoor scenes, hand color can change quite drastically over short periods of time and even spatially within one frame due to more pronounced surface normal shading and shadowing. Also the hand color can get saturated to white, which does not distinguish well from other bright using a high quality camera featuring rapid auto-gain control.

Since we are using only fingertips as point correspondences for camera pose estimation due to possible self occlusions. When fingertips are not visible, our system determines that it fingertips again as in. while this recovery happens promptly enough to not be overly disruptive in wearable applications, we have started to investigate use of more features on the hand and silhouette-based approaches such as active shape models or smart snakes in order to deal with more articulated hand poses and self occlusions.

References

1. Cai, Q., Aggarwal, J.K., "Tracking human motion in structured environments using a distributed – camera system," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* pp. 1241-1247, vol. 21, no. 11, Nov 1999.
2. Fakhr-eddine Ababsa, Malik Mallem, "Robust camera pose estimation using 2d fiducials tracking for real – time augmented reality system," *Pro. of the 2004 ACM SIGGRAPH international Conference on Virtual Reality continuum and its applications in industry*, *Singapore,* June 16-18, 2004.
3. Sanjay Kr. Singh, D.S. Chauhan, Mayank Vatsa, Richa Singh, "A Robust Color Based Face Detecion Algorithm," *Tamkang Journal of Science and Engineering,* vol. 6, no. 4, pp. 227-234, 2003.
4. Aoutif Amine, Sanaa Ghouali, Mohammed Rziza, "Face Detection in still Color Image Using Skin Color Information", *Proceedings ISCCSP2006,* UC Berkley, 2009.
5. Zhang, Z, "A Flexible new technique for camera calibration", *Proc. IEEE TPAMI,* 22(11), pp. 1330-1334, 2000.
6. Jian Chen Dawson, DM, Dixon, W.E. Behal,A "Adaptive homography - based visual servo tracking for a fixed camera configuration with a camera- in hand extension," *Control Systems Technology, IEEE Transactions on,* pp. 814-825, volume 13, no.5, Sept. 2005.
7. R. Haralick, H. Joo, C. Lee, X Zhuang, V. Vaidya and M. Kim, "Pose Estimation form Corresponding pointData", *IEEE Trans. Systems, Man and Cybernetics,* vol. 19, no 6,pp. 1426-1446.
8. Ababsa, F Mallem, M., "Robust camera pose estimation combining 2D/3D point and lines tracking", *Proc. IEEE International Symposium on Industrial Electronic, 2008, (ISIE-2008),* pp. 774-779, June 30-July 2 2008.
9. A.I. Comport, E. Marchand, M Pressigout and F. Chaumette, "Realtime markrless tracking for augmented reality: virtual servoing Framework." *IEEE Trans. on Visualization and Computer Graphics,* 12(6), pp. 615-628, July/ August 2006.