Real-time 3D Model-Based Gesture Tracking for Multimedia Control

Shih-Yao Lin[1], Yun-Chien Lai, Li-Wei Chan and Yi-Ping Hung[2]
Graduate Institute of Networking and Multimedia, National Taiwan University
{[1]d97944010, [2]hung}@csie.ntu.edu.tw

Abstract

This paper presents a new 3D model-based gesture tracking system for controlling multimedia player in an intuitive way. The motivation of this paper is to make home appliance aware of user's intention. This 3D model-based gesture tracking system adopts a Bayesian framework to track the user's 3D hand position and to recognize meaning of these postures for controlling 3D player interactively. To avoid the high dimensionality of the whole 3D upper body model, which may complicate the gesture tracking problem, our system applies a novel hierarchical tracking algorithm to improve the system performance. Moreover, this system applies multiple cues for improving the accuracy of tracking results. Based on the above idea, we have implemented a 3D hand gesture interface for controlling multimedia players. Experimental results have shown that the proposed system robustly tracks the 3D position of the hand and has high potential for controlling the multimedia player.

1. Introduction

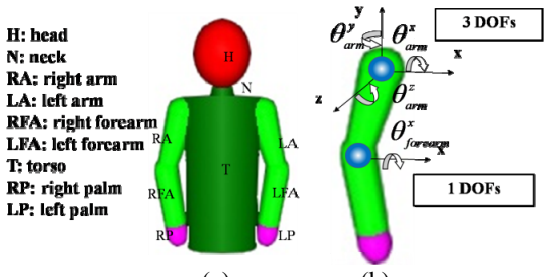
Gesture-based user interface is becoming a crucial issue in digital home applications and interactive games. Traditionally, people employ hand-held remote controllers or wired devices such as keyboard, mouse and joystick for home appliance control. The above mentioned user interfaces are restricted and not intuitive due to requiring users to wear or hold particular devices. Marker-less gesture controlling has been investigated for many years.

W. T. Freeman et al. [4] propose a hand recognition system using 2D template matching technique, allowing users employ the hand gesture to control graphical user interface components on the television. 3D model-based human pose tracking technique has been popular in recent years. However, it is a challenging task due to high dimensionality of the whole body motion parameter. Many researchers [2][3][7] have investigated solution to mark-less human motion capture. K. Tollmar et al. [5] develops a system based on Iterative Closest Point (ICP) which matches 3D model points to the observed 3D data from a stereo camera system, and estimates motions of body parts. Their system allows users to navigate virtual environment by using their body postures. The joint constrains are added while tracking the poses.

In this paper, we propose a real-time model-based gesture tracking system for multimedia control. The system uses 3D information from a stereo camera and adopts a modified hierarchical tracking approach to improve the performance. Moreover, a gesture interface is introduced to control multimedia player interactively.

2. 3D Upper Body Model

We used a 3D upper body model, as shown in **Figure 1**(a), to simulate upper body motion. The entire 3D upper body model consists of 9 rigid body parts: head, neck, torso, two arms, two forearms and two palms. As shown in **Figure 1** (a), the proposed model contains 14 degrees of freedom (DOFs) where 3 DOFs are for global body orientation, 3 DOFs for body location and 8 DOFs for two arms. **Figure 1** (b) shows that each arm of model has 3DOFs ($\theta^x_{arm}, \theta^y_{arm}, \theta^z_{arm}$) and each forearm of model has 1 DOFs ($\theta^x_{forearm}$).



(a) (b)

Figure 1: Upper body model and its motion parameters: (a) 3D articulated upper body, (b) arm motion parameter.

3. 3D Model-based Gesture Tracking Algorithm

This paper proposes a perceptual interface for multimedia control using 3D model-based gesture tracking algorithm.

3.1. Particle Filter

Particle filter [1], which based on Bayesian framework, is a useful technique for 3D body motion tracking because it provides multiple predictions for complex human motions. The posterior Bayesian formulation of the particle filter is defined as

$$p(x_t \mid Z_t) \propto p(z_t \mid x_t) \cdot p(x_t \mid Z_{t-1}) \quad (1)$$

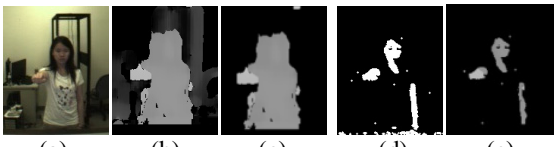
where x_t denotes the state vector at time t and z_t expresses the observation. The history of observations from 1 to t is indicated as $Z_t = \{z_1,...,z_t\}$. The pdf $p(x_t \mid Z_{t-1})$ is the prediction probability distribution at time t-1 and can be expressed as:

$$p(x_t \mid Z_{t-1}) = \int p(x_t \mid x_{t-1}) \cdot p(x_{t-1} \mid Z_{t-1}) d_{x_{t-1}} \quad (2)$$

The particle filter is a popular and useful algorithm for 3D body tracking. However, it suffers from the degeneracy problem and huge computational burden.

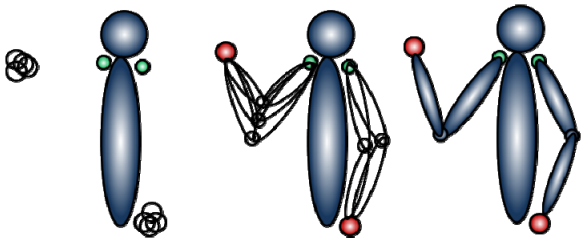
3.2. Feature Extraction

Our stereo camera provides a *depth map*: O_{depth} (**Figure 2** (b)) and a *skin color map*: O_{skin} (**Figure 2** (d)). In order to filter out unnecessary background information, which might interfere with the tracking result, we take the depth value of user's face as the baseline and only retain the depth data in front of this baseline (**Figure 2** (c) *user depth map* O_{ud}). *Palm possibility map*: O_{pp} (**Figure 2** (e)) is the product map of O_{skin} and O_{ud} .



(a) (b) (c) (d) (e)

Figure 2: Feature Extraction: (a) original image, (b) depth map, (c) user depth map, (d) skin color map and (e) palm possibility map.



(a) (b) (c)

Figure 3: Modified hierarchical searching procedure: (a) palm tracking process, (b) arm tracking process and (d) tracking result.

3.3. Gesture Tracking Algorithm

In our system, we employ a modified progressive particle filter [2][7] which allows us to perform an effective gesture tracking procedure in the high dimensional space. The motion state vector X_t of upper body is denoted as $X_t = \{x_t^L, x_t^p, x_t^A\}$. (x_t^L is global body motion state, x_t^P is the palm position state and x_t^A as the arm motion state at time *t*). The posterior Bayesian formulation is redefined as:

$$p(x_t \mid Z_t) \propto p(x_t^L, x_t^P, x_t^A \mid Z_t) \quad (3)$$

In the global motion tracking process, we use Adaboost face detection [6] to locate the face coordinate in the image plane. We then obtain 3D face position by finding corresponding coordinate on depth map. We set the head position of our 3D upper body model to user's face position to locate global body motion state.

In the arm tracking process, since the shoulder locations are inferred from head position, we use the connection between the shoulder position and palm to recover the motion parameters of whole arm. Unlike original hierarchical searching of the progressive particle filter, we first track the palm position and then estimate the entire motion parameter of arms. The modified hierarchical searching process is shown in **Figure 3**.

We modify the hierarchical searching framework of the progressive particle filter and divide the original posterior density into three parts, using the modified progressive particle filter as following:

$$\begin{aligned} & p(x_t^L, x_t^P, x_t^A \mid Z_t) \\ & \propto p(x_t^L \mid Z_t^L) p(x_t^P, x_t^A \mid Z_t^p, z_t^A) \\ & \propto p(x_t^L \mid Z_t^L) p(x_t^P \mid x_t^L, Z_t^P) p(x_t^A \mid x_t^L, x_t^P, Z_t^A) \end{aligned} \quad (4)$$

In the the posterior density of palm tracking $p(x_t^P \mid x_t^L, Z_t^P)$, x_t^P is the 3D position of palm at time t and x_t^L is the 3D head location. The posterior of palm tracking can be decomposed as:

$$p(x_t^P \mid x_t^L, Z_t^P) \propto p(z_t^P \mid x_t^L, x_t^P) p(x_t^P \mid x_t^L, Z_{t-1}^P) \quad (5)$$

where the likelihood $p(z_t^P \mid x_t^P)$ estimates the probability value by three different meansure processes.
1. *End-point measure process*: $p_{end}(z_t^P \mid x_t^P)$ is employed to verify that the prediected particle is drawn at the end-point of the arms, since that is the position of the palms.
2. *Skin measure process*: $p_{skin}(z_t^P \mid x_t^P)$ is used to calculate the number of pixels with skin color which are occupied by the predicted particle.
3. *Depth measure process*: $p_{depth}(z_t^P \mid x_t^L, x_t^P)$ estimates the distance from the camera to the palm. We assume that the palm is usually in the farest position from the body, when user controls the home appliance. Therefore the likelihood $p(z_t^P \mid x_t^L, x_t^P)$ is expressed as follows:

$$\begin{aligned} & p(z_t^P \mid x_t^L, x_t^P) \\ & \propto p_{depth}(z_t^P \mid, x_t^P) p_{skin}(z_t^P \mid, x_t^P) p_{depth}(z_t^P \mid, x_t^P) \end{aligned} \quad (6)$$

In the arm tracking process, we employ the inverse kinematics estimation. There are four motion parameters of entire arm. The shoulder position is given through global location tracking. The palm position is obtained after palm tracking process. As shown in **Figure 4**, the shoulder position is the green sphere, the joint of forearm is the blue sphere and the palm is the red sphere. The lengths of model's arm and forearm are given. Therefore we can calculate three angles: $\theta_{arm}^x, \theta_{arm}^y, \theta_{forearm}^x$ of the arm by inverse kinematics computation. In this process, we only need to estimate the probability angle of θ_{arm}^z. *The posterior of arm tracking* $p(x_t^A \mid x_t^L, x_t^P, Z_t^A)$ *is defined as:*

$$\begin{aligned} & p(x_t^A \mid x_t^L, x_t^P, Z_t^A) \\ & \propto p(z_t^A \mid x_t^L, x_t^P, x_t^A) p(x_t^A \mid x_t^L, x_t^P, Z_{t-1}^A) \end{aligned} \quad (7)$$

We sample the particles around all the angle θ_{arm}^z, the likelihood $p(z_t^A \mid x_t^L, x_t^P, x_t^A)$ is calculated by the overlapped region between the projection area of predicted model's arm and the skin area in O_{pp}.

3.4. Multimedia Gesture Controlling Interface

Our gesture controlling interface is used to control a multimedia player. This interface contains several virtual buttons (Play, Pause and Stop), time slide bar (to control the progress of video) and a volume tuner (to adjust the volume) in the 3D space. The estimated palm position is used to push the virtual button. The virtual time slide bar and volume modification are controlled in the same way. We regard "waving hand action" as the "start" command for the proposed multimedia player.

4. Experimental Result

Our system runs on the Intel Core2 Quad CPU 2.66GHz PC with Ram 3.25G. The proposed method is implemented using C++ language. The stereo camera is developed by Videre Design. We employ 300 particles for the palm tracking process and 15 samples for the arm tracking. The frame per second (fps) of entire upper body tracking is 16.

Figure 5 shows the proposed multimedia gesture interface. The yellow sphere in **Figure 5** is the cursor obtained by the palm position. User can manipulate this multimedia intuitionally.

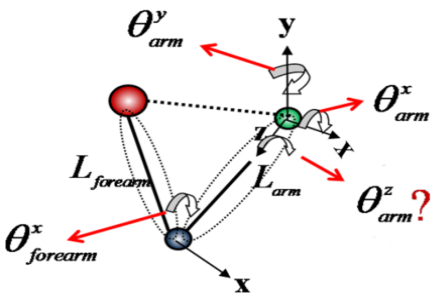


Figure 4: kinematics computation of entire arm reconstruction: we can obtain $\theta_{arm}^x, \theta_{arm}^y, \theta_{forearm}^x$ after palm tracking process and kinematic computation.

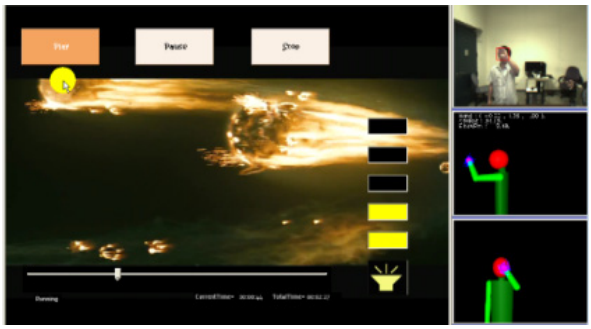
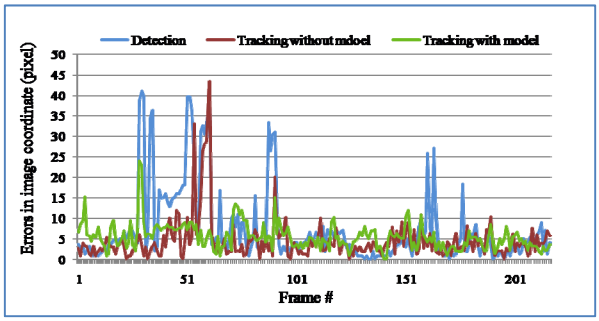


Figure 5: The proposed gesture recognition interface for multimedia control.

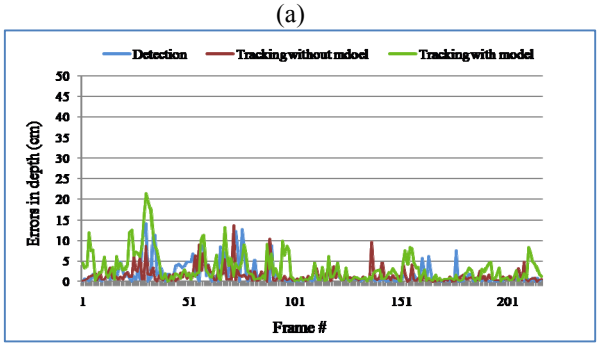
We evaluate the proposed framework by comparing it with palm detection technique and 3D palm tracking approach. The palm detection and 3D palm tracking are taking the highest value in the *palm probability map* (**Figure 2** (e)).

As shown in **Figure 6**, the result shows "tracking with model" is better than "tracking without model". The reason is that "tracking without model" has no information of body structure.

In **Figure 7**, we find that palm detection (**Figure 7**(c)) can easily fail when the area of forearm contains the same depth information and skin color information. **Figure 8** presents the tracking result in the sequence of pushing virtual button.

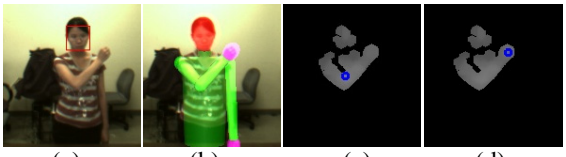


(a)



(b)

Figure 6: Comparison between palm detection, palm tracking and palm tracking with 3D model. (a) Errors (pixels) in 2D image coordinate. (b) Errors (cm) in depth. Green curve is the error of palm detection result, red curve is the error of 3D palm tracking result and green curve is the error of our approach tracking result.



(a) (b) (c) (d)

Figure 7: Comparison between palm tracking result: (a) original image, (b) palm tracking with 3D model, (c) palm detection and (d) palm tracking (note: the blue sphere in (c)(d) denotes palm tracking position).

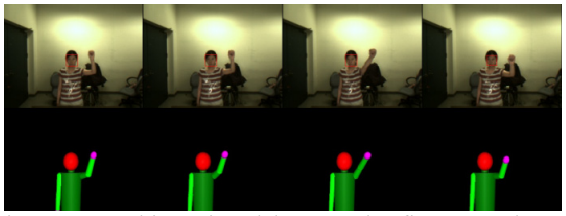


Figure 8: Pushing virtual button: the first row shows the original images and the second row presents the body tracking results.

5. Conclusion

This paper presents a new user interface for controlling multimedia player based on marker-less gesture tracking techniques. The proposed hierarchical tracking method with Bayesian framework runs the entire upper body tracking procedure in real-time. Moreover, the proposed method fuses various features to track the upper body posture and provides accurate results. We also design a 3D gesture interface to easily control a multimedia player. Experimental results have demonstrated that the proposed 3D model-based tracking approach is more robust and accurate than the approaches only applying tracking. In the future, we will apply our method to control of the other home appliances.

Acknowledgements

This work was supported in part by National Taiwan University, under grant 98R0062-04 and by the National Science Council, Taiwan, under grant NSC 98-2221-E-002-128-MY3

References

[1] M. Isard and A. Blake, "Condensation conditional density propagation for visual tracking," In *IJCV*, Vol. 1, pp. 5-28, 1998.

[2] S.-Y. Lin and I.-C. Chang, "3D Human Motion Tracking Using Progressive Particle Filter," In *ISVC*, 2008.

[3] J. Deutscher, A. Blake, and L. Reid, "Articulated Body Motion Capture by Annealed Particle Filtering," In *CVPR*, 2000.

[4] W. T. Freeman, "Television control by hand gestures," Workshop on *AFGR*, 1995.

[5] K. Tollmar, D. Demirdjian, T. Darrell, "Navigating in virtual environments using a vision-based interface," In *NordicCHI*, 2004.

[6] Paul Viola, Michael Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," In *CVPR*, 2001.

[7] S.-Y. Lin and I.-C. Chang, "Dynamic Kernel-based Progressive Particle Filter for 3D Human Motion Tracking," In *ACCV*, 2009.