Real-time Human Motion Sensing based on Vision-based Inverse Kinematics for Interactive Applications

Date, Naoto
Department of Intelligent Systems, Kyushu University

Yoshimoto, Hiromasa
Department of Intelligent Systems, Kyushu University

Arita, Daisaku
Department of Intelligent Systems, Kyushu University

Taniguchi, Rin-ichiro
Department of Intelligent Systems, Kyushu University

出版情報：Proceedings of the 17th International Conference on Pattern Recognition. 3, pp.318-321, 2004-08
バージョン：
権利関係：

Real-time Human Motion Sensing based on Vision-based Inverse Kinematics for Interactive Applications

Naoto Date, Hiromasa Yoshimoto, Daisaku Arita and Rin-ichiro Taniguchi
Department of Intelligent Systems, Kyushu University, Japan
http://limu.is.kyushu-u.ac.jp/

Abstract

Vision-based human motion sensing has a strong merit that it does not impose any physical restrictions on humans, which provides a natural way of measuring human motion. However, its real-time processing is not easy to realize, because a human body has a high degrees of freedom, whose vision-based analysis is not simple and is usually time consuming. Here, we have developed a method in which human postures are analyzed from a limited number of visual cues. It is a combination of numerical analysis of inverse kinematics and visual search. Our method is based on a general framework of inverse kinematics, and, therefore, we can use relatively complex human figure model, which can generates natural human motion. In our experimental studies, we show that our implemented system works in real-time on a PC-cluster.

1. Introduction

Man-machine seamless 3-D interaction is an important tool for various interactive systems such as virtual reality systems, video game consoles, etc. To realize such interaction, the system has to estimate motion parameters of human bodies in real-time. Vision-based motion capturing is a smart and natural approach since it does not impose any physical restrictions on a user. However, it has several problems to be solved.

From the viewpoint of interactive applications, a real-time feature is quite important and, therefore, computation intensive approaches[1, 2] is not realistic even though they provide a general framework. Real-time here means that images are processed at the speed of TV camera signal, .i.e., $20 \sim 30$ frames/sec. To realize such real-time systems, the key issues are as follows:

- robust image features, which are easy to extract
- fast human posture estimation from the image features

Usually, as image features, blobs (coherent region)[3][4] or silhouette contours[5] are employed. However, image features which can be robustly detected are limited, and, therefore, the estimation of 3D human postures from the limited cues are quite essential. To solve this problem, we have introduced vision-based inverse kinematics. In addition, to deal with the view dependency and the self-occlusion problem when a human makes various poses, we have employed an approach of multi-view image analysis. We have implemented our vision-based human motion sensing on a PC-cluster to realize its online and real-time processing.

2. System Overview

The basic algorithm flow of our real-time motion capturing is as follows:

1. Detection of visual cues
   - Silhouette detection, skin color blob detection, face direction detection.
   - Calculation of 3-D positions of features using multi-view fusion.
2. Human Motion Synthesis
   - Generation of human figure full-body motion and rendering in the virtual space including calculation of the interaction.

The key point of our system is *human motion synthesis* referring to a limited number of visual cues, which are acquired by the perception process. Several real-time vision-based human motion sensing have been developed. However, they are based on a rather simple human model and its generated human motion sensing is not natural. For example, the direction of face is not detected, or the number of articulations is limited. Here, we make the human figure model more complex and develop a vision- based algorithm for human motion sensing based on the model (details are discussed in 4).

When a real-time and on-line system is designed, an error recovery mechanism is quite important. If the system is based on feature tracking, once features fail to be tracked, the posture estimation process may reproduce unrealistic human postures, and it can not escape from the erroneous situation. Therefore, in order that we need not reset the system in such a case, we have introduced a simple

0-7695-2128-2/04 $20.00 (C) 2004 IEEE

error recovery process, which executes concurrently with the main human posture estimation process, and which always checks whether the human silhouette makes predefined shapes which are easy to recognize precisely. When the process finds the silhouette makes such shapes, it notifies the human posture estimation process, and the estimation process adopts the recognition result regardless of the estimation result. According to these considerations, we have designed a vision-based real-time human motion sensing as shown in Fig.1.

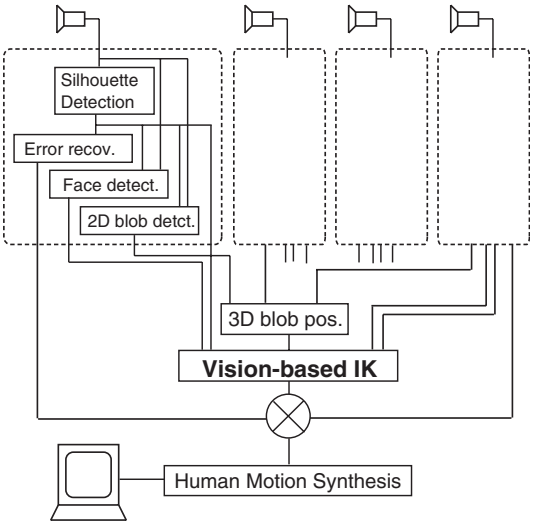


Figure 1. Software structure of real-time vision-based human motion sensing.

3. Acquisition of Perceptual Cues

3.1. Feature detection

We detect color blobs for a head, hands and feet, whose colors are acquired in advance, and calculate 2D positions of the color blobs as the centroids of the blobs. Color blob detection is based on color similarity evaluation, which is done in HUV color space to exclude the influence of lighting condition as much as possible. We also extract a human silhouette region by background subtraction, which is used by an error recovery process and a human posture estimation process.

In principle, the 3D positions of the blobs can be calculated from two views. However, due to self-occlusion, we can not always calculate the 3D positions with only two views, and we have used a certain number of cameras. Therefore, we have to solve a stereo-pair correspondence problem in a multi-view situation. We, first, extract possible stereo-pair candidates of the 2D blobs, and, then, classify their 3D positions into 5 clusters of feature points: head, right hand, left hand, right foot, left foot. In each cluster, we estimate the feature point position as the average position of the 3D position candidates after a dense part of the cluster is selected.

Figure 2. Face direction estimation

3.2. Estimation of face direction

Face direction is an important feature in human posture and is indispensable for interactive application. The problems here is the low resolution of face region, because cameras are arranged to capture a full-body and a face becomes small in the field of view. Therefore, face features such as eyes and mouths can not be clearly detected, and, then, feature-based or structure-based techniques of face direction estimation can not be applied. We, here, have employed a template matching method preparing face templates with multiple aspects. To reduce the computation time, we have employed an eigen-space method with several speed-up tactics[6]. Fig.2 shows an example of the face direction estimation. The estimation accuracy is not very high because of the low resolution images, but in most of interactive applications, we can get certain feedbacks and can modify the face direction based on them. We think the accuracy acquired by this approach is high enough to use.

4. Vision-based Inverse Kinematics

4.1. Inverse Kinematics

Our problem is to estimate human postures from a limited number of perceptual cues, which are blobs corresponding to hands, feet and head. This problem can be explained in a framework of Inverse Kinematics (IK) in the field of robotics. IK is to determine joint angles $\theta_{\mathbf{i}}$ of a manipulator so that the position of an end effector, or a final point, $\mathbf{P}_n$, coincides with a given goal point $\mathbf{G}$: $\mathbf{P}_n(\theta_1, ..., \theta_n) = \mathbf{G}$: where the manipulator has n segments. The difficulty here is that even if the goal is attainable[1], there may be multiple solutions and, thus, the inverse problem is generally ill-posed.

[1] If the distance of the goal to the initial point of the manipulator is larger than the sum of the lengths of the segments, the goal is not attainable.

In our problem, end effectors are hands, feet and a head, and the goals are the blob positions acquired by the perceptual process. The posture estimation, which is to decide the positions of joints of the human model, is achieved by calculating the joint angles in the frame work of IK. In dealing with manipulators, essentially, we can select any of multiple solutions θ_i because the end effector reaches the goal in any cases[2]. However, in human motion sensing, each joint position acquired by IK should be coincide with a joint position of a given human posture, and, therefore, we have to find the unique and correct solution. To this problem, Yonemoto et al employed very simple human model with 14 DOFs and very simple approximation[7]. They approximate each arm or leg as a two-segment manipulator and rotation around the shoulder, or rotation around an axis connecting the initial and the goal point, is fixed to a constant value, and, thus, the IK problem is reduced into a triangle problem in 2D space, which is quite easy to solve. However, because their human model is too simplified, and the assumption that the rotation around the shoulder is constant is not realistic, the estimated human posture is relatively poor. To solve this problem without over simplification, we have employed a human model with relatively complex geometry and have introduced a Vision-based Inverse Kinematics.

4.2. Vision-based Inverse Kinematics

We have used a human model with 23 DOFs shown in Fig.3, which has rich expressiveness of human posture. However, its IK problem becomes more complex, and introduces a large amount of ambiguity for its solution. Our method to solve this problem is divided into two phases: *acquisition of initial solution* and *refinement of initial solution* For simplicity, here, we explain human posture estimation of a upper body.

Acquisition of initial solution

Inverse Kinematics is solved by an ordinary numerical method[8] and initial candidates of 3D positions of shoulders and elbows are calculated. Here, we assume that the lengths of the bones in Fig.3 are given in advance. At time t, a hand position $(x(t), y(t), z(t))$ is represented as

$$(x, y, z) = \mathbf{P}(T_x(t), T_y(t), T_z(t), \theta_1(t), \theta_2(t), \cdots, \theta_N(t)) \quad (1)$$

where

- $T_x(t), T_y(t), T_z(t)$ indicate the head position in the world coordinate,
- $\theta_1(t), \theta_2(t), \theta_3(t)$ indicate rotation angles between the world coordinate and the local coordinate of the head, which are calculated from the face direction,
- $\theta_j(t)(4 \leq j \leq N(= 8))$ indicate rotation angles among connected parts .

[2]With other constraints, solutions should be narrowed down.

We suppose that, at time $t + 1$, the hand position moves to $(x(t + 1), y(t + 1), z(t + 1))$, the head position moves to $(T_x(t + 1), T_y(t + 1), T_z(t + 1))$, and the head direction changes to $(\theta_1(t + 1), \theta_2(t + 1), \theta_3(t + 1))$. Here, we slightly modify $\theta_j(t + 1)$, $(4 \leq j \leq N)$ so as that the hand position, i.e., the position of the end effector, $\mathbf{P}(T_x(t+1), T_y(t+1), T_z(t+1), \theta_1(t+1), \cdots, \theta_N(t+1))$ approaches the goal position $(x(t + 1), y(t + 1), z(t + 1))$. Repeating this process until the end effector position coincides with the goal position, we acquire the positions of a shoulder and an elbow. In order to exclude impossible postures, we have imposed a possible range on each angle θ_j.

Refinement of initial solution

The posture estimated in the previous step is just a solution of inverse kinematics, and it is not guaranteed that it coincides with the actual posture. This is due to ill-poseness of the inverse kinematics. To estimate the posture more accurately, we refine the acquired solution by referring to input image data. The basic idea is simple: if the shoulder and elbow positions acquired by the previous phase are correct, they should be inside of the human region in 3D space. Otherwise, the acquired solutions are not correct and they should be modified so as to be included in the human region. Here, we empirically assume that the shoulder position is acquired by solving the basic inverse kinematics, and we mainly refine the elbow position. Its basic algorithm is as follows:

- We have the shoulder position by solving the inverse kinematics and the hand position by color-blob analysis.
- When the lengths of its upper arm and forearm are given, the position of its elbow is restricted on a circle in 3D space. The circle is indicated by C.
- When the elbow is searched on the circle, we exclude impossible values, with which the arm get stuck in the torso.
- As shown in Fig.4, an elbow detection rectangle is established in a plane which is constructed by the shoulder, an hypothesized elbow and the hand.
- Then, in each view, the rectangle is reversely projected on the image plane and correlation between the projected rectangle and the human silhouette region is calculated.
- Then, by varying the position of the hypothesized elbow, the correlation R can be parameterized by ϕ, which is the angle around the center of the circle C. We search for ϕ giving the maximum R, which indicates the elbow position.

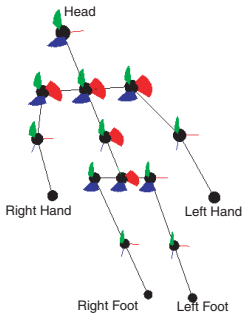


Figure 3. Human model

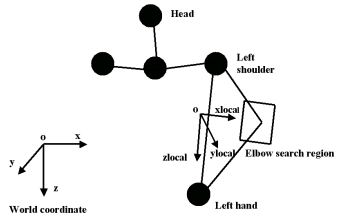
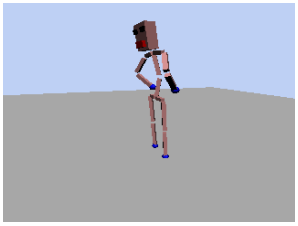
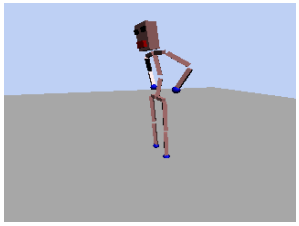


Figure 4. Elbow pos. estimation

Actual posture



estimation by basic IK



estimation by vision-based IK

Figure 5. Result of human posture estimation

5. Experimental Study

In this experiment, we have used 9 sets of IEEE1394-based color cameras (Sony DFW-V500) with f:4mm lenses, which are geometrically calibrated in advance. The images are captured with the size of 640×480 pixels, and the frame rate is 15 fps.

We have implemented our vision-based human motion analysis on PC-cluster with 3.2 GHz PentiumIVs. Computation time required in major vision algorithms of our human motion analysis are as follows:

2D blob detection:	4ms
3D blob calculation:	5ms
face direction calculation:	15ms
IK initial solution:	3ms
IK solution refinement:	27ms
template matching for error recovery:	11ms

The above figures indicate that real-time processing of our vision-based human motion sensing can be achieved on the PC-cluster, where detection of perceptual cues in each view is executed in parallel, and where solving IK and human figure generation are executed in a succeeding PC.

Fig.5 shows a typical example of human motion sensing. In this system, face direction is estimated, the shoulder positions are slightly shifted according to the hand positions, and the upper body bends forward slightly, which can be achieved only with our complex human figure model.

However, with just initial solution of the inverse kinematics, the elbow positions are not correctly estimated, while applying the refinement process, the elbow positions are correctly estimated. Thus, our system can recover human postures accurately in real-time, and can be applied to variety of interactive applications.

6. Conclusions

Here, we have shown a real-time human motion capturing without special marker-sensors. For full body motion analysis, we have adopted a multi-view approach. The key point is that we have established a framework of estimation of full-body motion from a limited number of perceptual cues, which can be stably extracted from input images. Since the system implemented on PC-cluster works in real-time and online, it can be applied to various *real-virtual* applications. We will improve each human posture estimation algorithm to recover human posture more precisely. Robust image feature detection under severe lighting condition is also important, which widen its application fields.

References

[1] T.Nunomaki, et al, "Multi-part Non-rigid Object Tracking Based on Time Model-Space Gradients," *Proc. AMDO*, 2000.

[2] J.Deutshcher, et al, "Automatic Partitioning of High Dimensional Search Spaces Associated with Articulated Body Motion Caputure," *Proc. CVPR*, Vol.2, 2001.

[3] C.Wren, et al, "Pfinder: Real-Time Tracking of the Human Body", *IEEE Trans. PAMI*, Vol.19, No.7, 1997.

[4] M.Etoh, et al, "Segmentation and 2D Motion Estimation by Region Fragments", *Proc. ICCV*, 1993.

[5] K.Takahashi, et al, "Remarks on a Real-Time 3D Human Body Posture Estimation Method using Trinocular Images," *Proc. ICPR*, Vol.4, 2000.

[6] S.A.Nene, et al, "A Simple Algorithm for Nearest Neighbor Search in High Dimensions," *IEEE Trans. PAMI*, Vol.17, No.9, 1997.

[7] S.Yonemoto, et al, "Real-Time Visually Guided Human Figure Control Using IK-based Motion Synthesis," *Proc. WACV*, 2000.

[8] L.Wang, et al, "A Combined Optimization Method for Solving the Inverse Kinematics problem of Mechanical Manipulators", *IEEE Trans. Robotics and Automation*, Vol.7, No.4, 1991.