Towards real-time body pose estimation for presenters in meeting environments

Ronald Poppe Dirk Heylen Anton Nijholt Mannes Poel
University of Twente
Department of Computer Science
Human Media Interaction Group
P.O. Box 217, 7500 AE, Enschede, The Netherlands
{r.w.poppe, d.heylen, a.nijholt, m.poel}@ewi.utwente.nl

ABSTRACT

This paper describes a computer vision-based approach to body pose estimation. The algorithm can be executed in real-time and processes low resolution, monocular image sequences. A silhouette is extracted and matched against a projection of a 16 DOF human body model. In addition, skin color is used to locate hands and head. No detailed human body model is needed. We evaluate the approach both quantitatively using synthetic image sequences and qualitatively on video test data of short presentations. The algorithm is developed with the aim of using it in the context of a meeting room where the poses of a presenter have to be estimated. The results can be applied in the domain of virtual environments.

Keywords

3D human pose estimation, blob tracking, silhouette matching, avatar animation

1. INTRODUCTION

Being able to recognize gestures or body poses is a key issue in many applications. Human pose estimation based on computer vision principles is a technology that suits the requirements for an inexpensive, widely applicable approach to human pose estimation. We identify five application areas:

1. **Surveillance**. A subject is tracked over time and monitored for special actions, for examples monitoring a parking lot to detect thieves.
2. **Human-computer interaction**. The captured poses are used to provide controlling functionalities. Examples are the control of a crane or a computer game.
3. **Analysis**. Analysis of captured human poses can help understanding movement in clinical studies.
4. **Annotation**. Automatic annotation of video data is useful for automatic indexing. An example is annotation of meeting video data.
5. **Animation**. Body poses are used to animate avatars in virtual environments.

WSCG SHORT papers proceedings, ISBN 80-903100-9-5
WSCG'2005, January 31-February 4, 2005
Plzen, Czech Republic.

In this paper we describe a real-time human pose estimation algorithm that uses low resolution images from a single camera. The approach is suitable for the estimation of body poses in an indoor setting. We have developed it with the aim of using it in the context of a meeting room where the poses of a presenter have to be estimated. The results of the system can also be used in gesture recognition applications. Tracking of multiple people is not supported. Currently the body must be facing the camera at all times. This limitation is realistic for presenter behavior recognition and for human-computer interaction (HCI) purposes.

Related Work

Human motion tracking has received a significant amount of attention from the computer vision research community in the last decade. Many pose estimation approaches use images from multiple cameras to obtain more accurate 3D data. However, a broader range of applications can benefit from a single-camera approach. An overview of computer vision-based human motion capture approaches has been made by Moeslund and Granum [Moes01].

Features are used to describe the subject in the scene in a convenient way. A variety of features has been used in previous approaches, among others colors, edges, silhouettes and textures. Silhouettes are often used because they are quite robust and do not contain ambiguous information such as inner edges. Silhouettes are used in multi-view settings, for example by Kakadiaris and Metaxas [Kaka96]. They recover all the body parts of a subject by monitoring the changes over time to the shape of the deforming silhouette. Delamarre and Faugeras [Dela99] also use silhouette contours in a multi-camera setting. They use a variation of the Iterative Closest Point algorithm to fit a 3D human model inside the silhouettes. A manual initial estimation is needed. Mittal and others [Mitt03] use silhouette decomposition and are able to estimate the poses of multiple persons in cluttered scenes viewed from multiple angles. Sminchisescu and Telea [Smin02] use monocular image sequences and a similarity measure that uses attraction and area overlap terms on smoothed silhouettes. In recent monocular work from Agarwal and Triggs [Agar04] a sparse Bayesian regression method is used to estimate 3D body poses directly from silhouettes. They are able to estimate angles for a 54 degree of freedom (DOF) human body model. Our algorithm uses monocular video sequences and is able to work in real-time, without a complex body model and without relying on previously learned motion patterns.

The pose estimation process estimates the human pose from these features. The output of a pose estimation process is usually a set of angles for each DOF. A DOF corresponds with a single joint rotation.

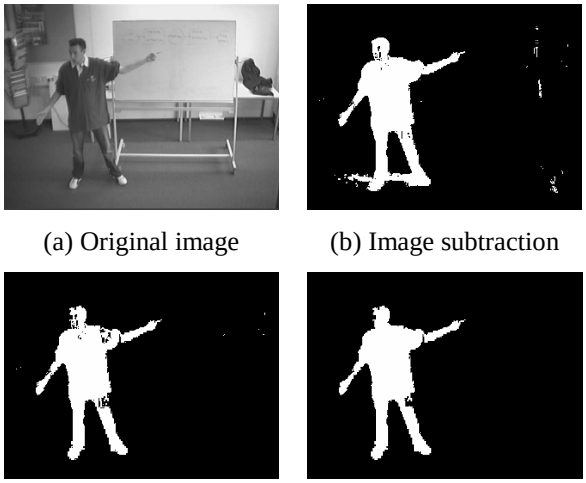
In the remainder of this paper, we discuss feature extraction and pose estimation (Section 2 respectively Section 3), experiment setup and obtained results (Section 4), and conclusions and future work (Section 5 respectively Section 6).

2. FEATURE EXTRACTION

We used silhouettes to represent the subject in the scene. Simple background subtraction was applied, using a pre-recorded image of the background. Lighting compensation was used to correct small variations in light intensity. The algorithm is computationally fast and performs reasonably well for indoor situations. The next step in the extraction process is removal of shadow. Shadow detection was carried out in HSV color space. The saturation channel (S) was ignored since it was not a good predictor.

A novel approach to noise removal using a contour tree was used. All nodes in the contour tree except the root node have a parent which is the containing contour in the image. Since a maximum of one person was allowed in the scene and the human body is rigid, only the largest contour on the first level in the tree needed to be examined. To construct the tree, contours were extracted using an efficient region growing technique. The contour tree algorithm proved to yield better results than traditional methods such as erosion and dilation. The algorithm is computationally fast and can easily be adapted to work with other assumptions.

The results of the image subtraction, shadow detection and noise removal process are shown in Figure 1.



(a) Original image (b) Image subtraction

(c) Shadow removal (d) Noise reduction

Figure 1: Silhouette extraction process example.

Skin Color Detection

Skin color was extracted from the silhouette, using the hue and saturation channel in the HSV color space. An upper and lower threshold were set. On a set of 10 images with hand-labeled skin regions, the precision was nearly 85 % at a recall rate of 95 %. Noise was removed from the results of the skin color detection with the contour tree algorithm described in the previous section. This resulted in several skin color blobs. Results of the skin color detection before and after noise removal are shown in Figure 2. Using color can cause robustness problems. But since only the largest skin color contours are evaluated, very few limitations on clothing color have to be imposed.

(a) Skin color detection (b) Noise removal

Figure 2: Skin detection process example.

3. POSE ESTIMATION

We used the silhouette and skin region features as input for the pose estimation process, which produces a set of joint angles for each degree of freedom in the human body model. This model is described in the next section.

Human Body Model

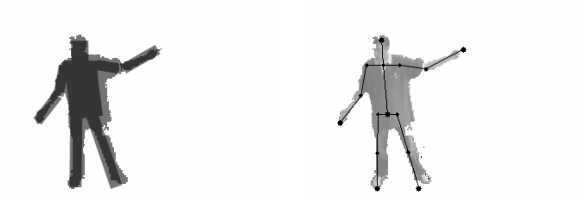
Various body models have been proposed in literature. Some contain a small number of joints or focus on a part of the human body such as the upper body or an arm. The model that is used here is derived from the H-anim standard which models the human body as a hierarchic skeleton composed of joints and rigid segments. The simplified model contains 14 segments that are modeled as cylinders. The 10 joints have a total of 16 DOF, three for each hip and shoulder and one for each knee and elbow. Because of the low resolution of the image frames the rotation of the hands and feet is not modeled. The model contains segment lengths and joint angle limitations.

Calculating 3D Positions

The distance of the person to the camera can be calculated by looking at the bottom of the silhouette. This distance was used to calculate the 3D locations of the head and hands. The skin regions were tracked using a simple and fast tracking algorithm. A minimum frame rate of approximately 10 frames per second is necessary to track fast hand movements. Likely positions of the hands and head were used to get an initial labeling. Since the body always had to face the camera, shoulders and hips were in the same plane. The 3D location of a hand can lie on a line through the camera center and the obtained coordinate in the shoulder-hip plane. Different values for this location were evaluated in the inverse kinematics process.

Inverse Kinematics and Silhouette Matching

An analytical inverse kinematics approach was used to compute the location of the elbows and knees and to calculate the rotations of all DOF. Different values for the hand location and the elbow position were evaluated for the arms and legs. Joint angle limits were applied to reduce the search space. The projection of the human body model was matched against the extracted silhouette. Since the estimation with the best silhouette match is not always the most likely body pose, the previously estimated pose was used to generate a smooth estimation of the movement over time.



(a) Silhouette matching (b) Pose estimation

Figure 3: Silhouette matching and pose estimation example.

4. EXPERIMENTAL RESULTS

Tests with a camera have been carried out to evaluate the system's performance. Since no ground truth about joint angles could be inferred from these data, quantitative experiments were carried out on movie sequences generated by Curious Labs Poser.

System Implementation

The system uses AVI movies or camera streams at a resolution of 320 × 240 pixels. Tests were carried out at a frame rate between 15 and 25 frames per second.

Tests with Poser Movies Sequences

Four movies with a total of 240 frames were generated with frame rate of 15 frames per second. Sample frames of the four movie sequences are depicted in Figure 4. To evaluate the results, the angles for each of the 16 DOF were measured and compared to the estimated angles. Two body models were evaluated, one with default segment lengths and one with 10 % shorter limbs. Shortening the segments yielded better results for the hip and knee rotations but slightly worse results for the shoulder and elbow rotations. The average and standard deviation of the joint angle errors are summarized in Table 1.

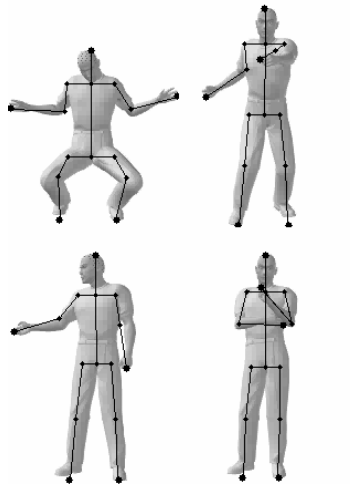


Figure 4: Frame 15 of each of the four synthetic test movies with superimposed skeleton.

The results can be explained by the human body model used. If the segment length in the model does not match the actual segment length, miscalculations occur. One of the causes can be found in the matching method. When arm segment lengths are shortened, the elbow angle increases.

Tests with Camera

Qualitative tests using a camera were carried out on five short presentations by graduate students. The presentations were between two and five minutes long and dealt with graduation topics. The presentations contained typical presenter movements such as pointing to the whiteboard. Few errors were made by the tracker and the pose estimation process. Visualization showed smooth movements. However, dropping the frame rate from 25 to 5 frames per second caused the multiple blob tracker to fail.

Rotation	Default		Shortened	
	Avg	SD	Avg	SD
L. sh. flexion	8.64	6.70	11.44	9.14
L. sh. rotation	0.72	0.61	3.49	0.78
L. sh. adduction	8.95	3.56	4.78	2.27
L. el. flexion	7.17	3.61	7.70	5.85
L. hip flexion	1.05	0.20	0.97	0.21
L. hip rotation	2.11	1.60	1.02	0.34
L. hip adducution	0.97	1.93	0.78	1.19
L. knee flexion	11.97	3.91	8.72	3.99
R. sh. flexion	11.83	8.70	16.37	7.51
R. sh. rotation	4.12	1.61	2.50	1.37
R. sh. adducution	14.58	8.85	14.68	7.23
R. el. flexion	9.56	3.32	13.73	7.25
R. hip flexion	0.92	0.92	0.97	0.87
R. hip rotation	3.78	1.79	1.58	0.34
R. hip adduction	0.83	1.70	0.65	1.24
R. knee flexion	10.51	1.23	7.21	3.44
Average	**6.11**	**3.14**	**6.04**	**3.31**

Table 1: Average and standard deviation of joint angle error distance in degrees for the synthetic test movies with default and shortened limbs.

5. CONCLUSION

In this paper we discussed a pose estimation approach that is computationally fast thanks to efficient algorithms such as a novel noise removal algorithm. Silhouette matching is used to match a projection of a 16 DOF human body model to the extracted silhouette. The system works reasonably well, with an average joint angle error distance below 10 degrees. The system is suitable for the estimation of presenter poses in meeting environments.

6. FUTURE WORK

Future work will aim at allowing multiple persons and removing the limitation that the person has to face the camera. This will make the approach suitable for automatic pose estimation not only for presenters but also for meeting participants. Figure 5 shows a virtual meeting room with participants. Extending the approach with gesture recognition will allow for recognition of meeting acts.

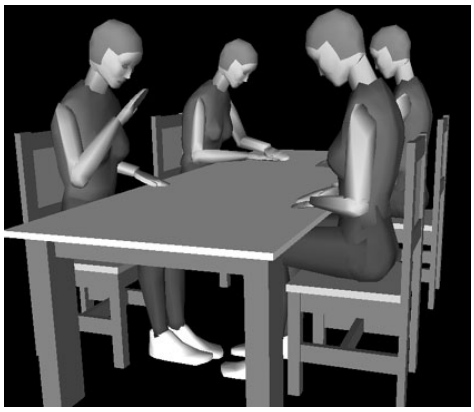


Figure 5: Virtual meeting room.

7. ACKNOWLEDGEMENTS

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-12).

8. REFERENCES

[Agar04] A. Agarwal and B. Triggs, "Learning to track 3D human motion from silhouettes," in Proceedings of the 21st International Conference on Machine Learning (ICML), July 2004.

[Dela99] Q. Delamarre and O. Faugeras. 3D articulated models and multi-view tracking with silhouettes. In International Conference on Computer Vision: ICCV 1999, Corfu, 1999.

[Kaka96] I.A. Kakadiaris and D. Metaxas, "Three-dimensional human body model acquisition from multiple views," IJCV, vol. 30, no. 3, pp. 191–218, December 1998.

[Mitt03] A. Mittal, L. Zhao, and L.S. Davis, "Human body pose estimation using silhouette shape analysis," in Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, July 2003, pp. 263–270.

[Moes01] T.B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," Computer Vision and Image Understanding: CVIU, vol. 81, no. 3, pp. 231–268, 2001.

[Smin02] C. Sminchisescu and A. Telea. Human pose estimation from silhouettes, a consistent approach using distance level sets. In WSCG International Conference on Computer Graphics, Visualization and Computer Vision, 2002.