Human pose estimation

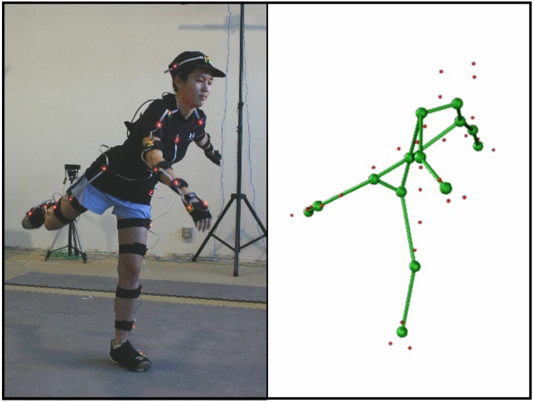
Samir Azrour and Marc Van Droogenbroeck

Academic year: 2015-2016

What is human pose estimation ?

Definition (Human pose estimation)

In computer vision, it is the study of algorithms and systems that recover the pose of a human body, which consists of joints and rigid parts.



Applications of human pose estimation: entertainment

Video games with the camera Kinect of Microsoft.

Analyze the motion of athletes to optimize it.

Applications of human pose estimation: medical

It can be used for the rehabilitation of injured persons or walking analysis of neurologically diseased persons.

It can be used to animate 3D characters.

Types of camera-based pose estimation systems

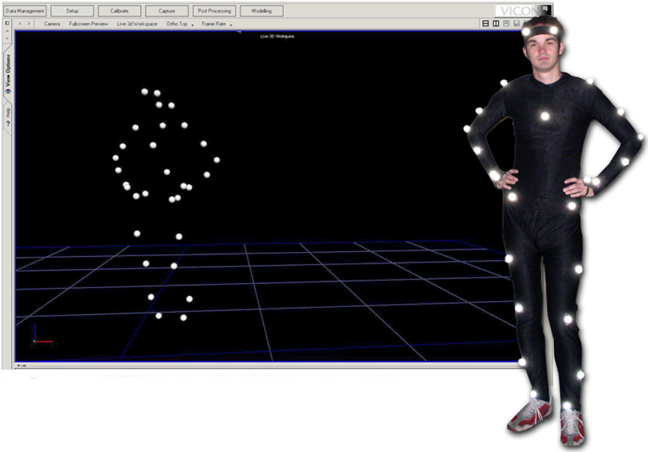
The camera-based pose estimation systems (or motion capture systems) can be marker-based or markerless:

maker-based systems: markers are put on the subject and the pose is recovered by localizing these markers with a multi-camera setup.

markerless systems: the subject has nothing to wear and its pose is recovered using a body model tracking method or a machine learning technique.

State-of-the-art marker-based system with passive markers

- The Vicon system:



How does the Vicon system works ?

- It uses more than 10 calibrated IR cameras with IR LEDs.

- A set of reflective markers are placed on anatomical landmarks of the subject.

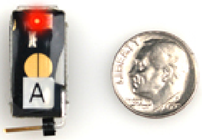
- The images taken by the cameras are filtered to keep only the markers.
- A 3D representation of the markers is constructed based on all the images.
- The body joint locations are recovered based on the markers positions.

State-of-the-art marker-based system with active markers

- The PhaseSpace system:

Advantages of active markers compared to passive ones ?

- *Each marker* is powered to emit its own light and can be *uniquely identified*.



$\Longrightarrow$ The marker swapping problem is eliminated.

$\Longrightarrow$ It provides much cleaner data.

Pros and cons of marker-based pose estimation systems

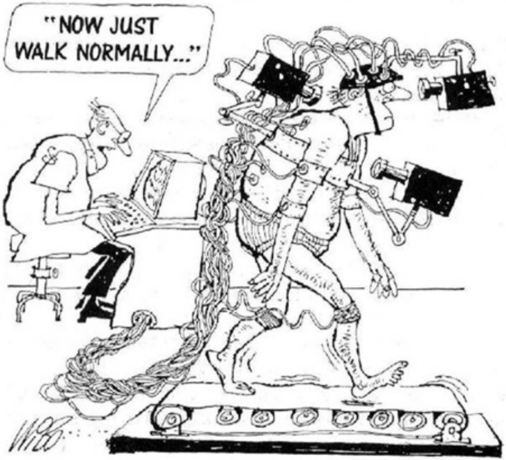
Pro

- Accuracy (error smaller than 1 mm on the markers positions)

Cons

- Long time needed to equip the person
- Errors due to markers misplacement
- Errors due to soft tissue artifact
- Large number of cameras needed wrt the tracking area ($>$10 cameras for a 5 x 5 m area)
- Markers can modify the gait

Pros and cons of marker-based systems



Markerless systems

For what purpose ?

Markerless systems can solve nearly all the mentioned disadvantages of marker-based systems

Depending on the application the objective is to make them either:

- as accurate as possible (medical and sports analysis).

or

- as fast as possible (gaming).

or

- a trade-off between the two (animation movies).

State-of-the-art markerless system for medical and sports analysis (ref: Corazza et al. 2010)

Main characteristics of the method:

- Multiple color cameras (> 8)
- 3D reconstruction of the subject's body
- Subject-specific model
- Accurate and anatomically consistent tracking algorithm
- Not realtime

3D reconstruction of the subject's body

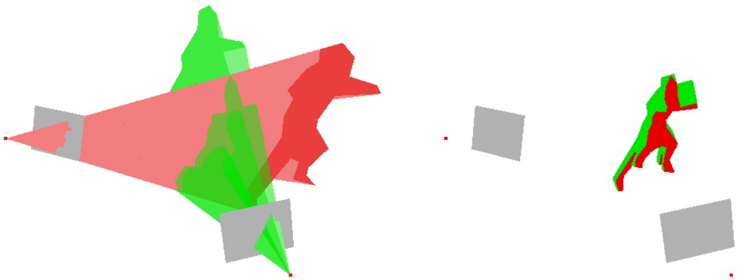
A 3D reconstruction of the subject's body is obtained from the calibrated color cameras:

1. The background is subtracted in each color camera image sequence using an intensity and color threshold.
2. The 3D reconstruction is achieved through *visual hull*.

Visual hull

Definition (Visual hull [Laurentini, 1994])

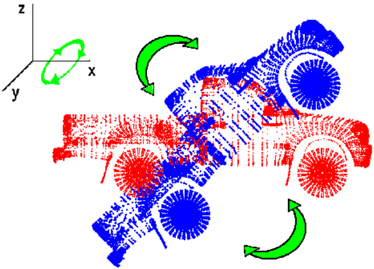
The visual hull is defined as the maximal volume consistent with an object's silhouettes as seen from a set of viewpoints.



Tracking algorithm

The visual hull reconstruction is tracked using an *articulated Iterative Closest Point* (*ICP*) method and the subject-specific body model.

ICP: algorithm that minimizes the difference between two clouds of points by using translation and rotation transformations



Tracking algorithm

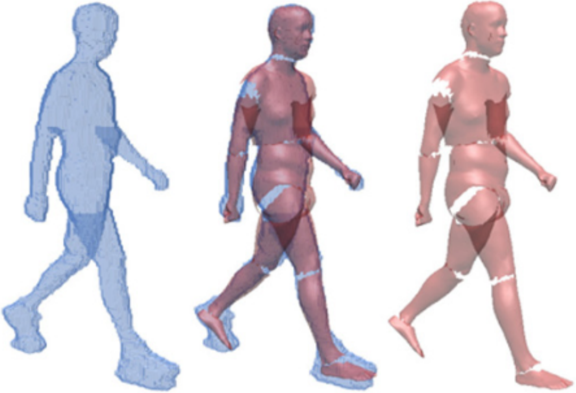
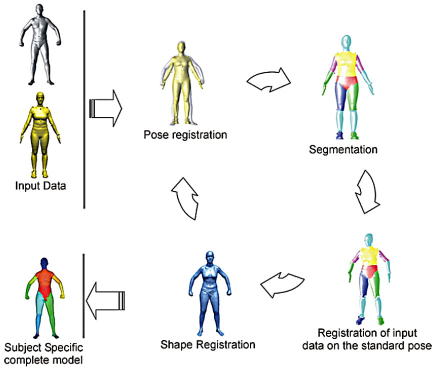


Figure: Visual hull (blue) and body model (red) matched with an articulated ICP algorithm.

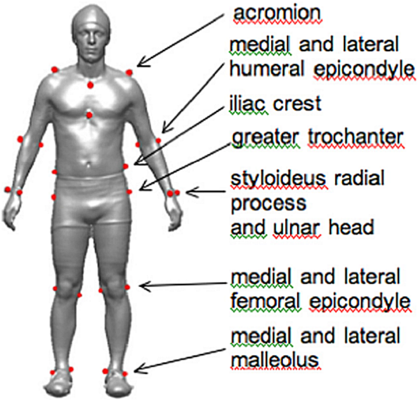
Generation of a subject-specific model

The key of the method to improve accuracy is to generate a subject-specific body model with joint center locations.

- This is done using just one static scan (mesh) of the subject



A training data set of nine subjects was used to learn the optimal joint center locations in a subject-specific model.



Learning joint centers locations

To make the process of model generation fully automatic, the joint center locations are linked to the n nearest vertexes in the mesh.

$$(a_1 a_2 \dots a_n)_j \begin{bmatrix} x_{1i} & y_{1i} & z_{1i} \\ x_{2i} & y_{2i} & z_{2i} \\ \dots & \dots & \dots \\ x_{ni} & y_{ni} & z_{ni} \end{bmatrix} = (\bar{x_{ji}} \bar{y}_{ji} \bar{z}_{ji}) \tag{1}$$

where $(\bar{x_{ji}} \bar{y}_{ji} \bar{z}_{ji})$ are the coordinates of the joint center j.

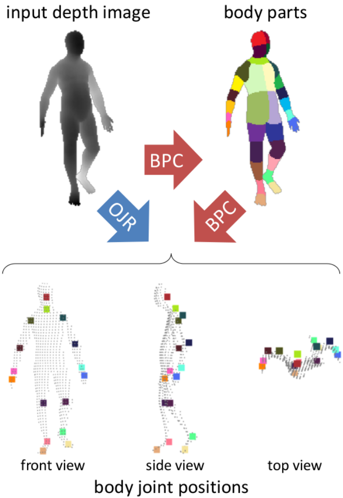
- It was found that $n = 7$ minimizes the generalization error.

State-of-the-art markerless system for entertainment (ref: Shotton et al. 2012)

Main characteristics of the method:

- One depth camera
- Machine learning approach with a large synthetic training set
- Each frame is treated independently (no temporal information)
- Super-realtime (around 200 fps)

Outline of the method



Body part classification approach

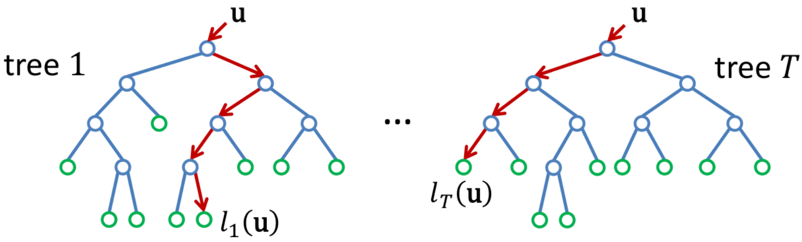
The *body part classification* (BPC) estimates the human pose in 2 steps:

1. It predicts a body part label for each pixel.
2. It uses the inferred body part labels to localize the body joint centers.

The method use a large (1 million images) and highly varied training set of synthetic data.

Body part prediction model

- A forest of decision trees is used to predict a body part label for each pixel $\mathbf{u}$.

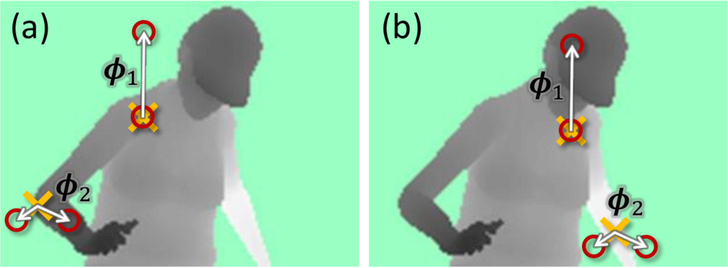


Body parts prediction model

- A feature value is thresholded at each split node and the pixel $\mathbf{u}$ takes a different path depending on the result.
- The leaf where the pixel $\mathbf{u}$ ends determines the probabilities to belong to the different body parts.
- The final prediction is obtained by averaging the predictions over all the trees.

Features

The features used are simple depth comparisons



$$f(\mathbf{u}|\phi) = z\left(\mathbf{u} + \frac{\delta_1}{z(\mathbf{u})}\right) - z\left(u + \frac{\delta_2}{z(\mathbf{u})}\right) \tag{2}$$

with feature parameters $\phi = (\delta_1, \delta_2)$

Problem:

- In the world space coordinates, the pixels lie on the body surface and so they are not aligned with a body joint in the z direction.

Solution:

1. The 3D coordinates of each pixel are computed: $\mathbf{x}(\mathbf{u}) = (x(\mathbf{u}), y(\mathbf{u}), z(\mathbf{u}))^T$
2. An offset along the z direction ζ_j is used to push back the 3D coordinates to better align with the interior body joint j: $\mathbf{x}_j(\mathbf{u}) = \mathbf{x}(\mathbf{u}) + (0, 0, \zeta_j)$

Recovering body joint locations

How do we map the surface body parts to the interior body joint locations?

1. Each pixel $\mathbf{u}$ provides exactly one vote $\mathbf{x}_j(\mathbf{u})$ for each body joint j.
2. Each vote is given a weight

$$w_j(\mathbf{u}) = p(c = c(j)|\mathbf{u}).z^2(\mathbf{u}) \qquad (3)$$

where $c(j)$ is the body part associated with joint j.

3. The body joint locations are then given by the modes of the following density estimators:

$$p_j(\mathbf{x}') \propto \sum_{\mathbf{u}} w_j(\mathbf{u}).exp\left(-\left\|\frac{\mathbf{x}' - \mathbf{x}_j(\mathbf{u})}{b_j}\right\|^2\right) \qquad (4)$$

Conclusion

- There exist a lot of different pose estimation methods.
- The choice of a method strongly depends on the application and should be based on three main aspects:

$\Longrightarrow$ the setup complexity

$\Longrightarrow$ the computing time

$\Longrightarrow$ the precision of the pose estimation

A fun application: combining a markerless pose estimation system with a virtual reality system

Oculus Rift : a virtual reality headset for 3D gaming

Bibliography

S. Corazza, E. Gambaretto, L. Mundermann, and T. Andriacchi.
Automatic generation of a subject-specific model for accurate markerless motion capture and biomechanical applications.
IEEE Transactions on Biomedical Engineering, 2010

S. Corazza, L. Mündermann, E. Gambaretto, G. Ferrigno, and T. Andriacchi.
Markerless motion capture through visual hull, articulated ICP and subject specific model generation. International Journal of Computer Vision, 2010.

J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake.
Efficient human pose estimation from single depth images.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013.