Introduction to pose estimation

Lesson given by **Sébastien Piérard** in the course
"Vision 3D" (ULg, Pr. M. Van Droogenbroeck)

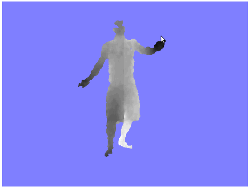
INTELSIG, Montefiore Institute, University of Liège, Belgium

November 30, 2011

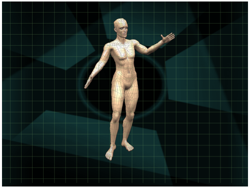
What is pose recovery ?

Example with a *Kinect*:

input data



input segmented



result

[image source: J-F Hansen & D Leroy, "Réalisation d'une plateforme d'immersion pour jeux 3D interactifs", 2011]

Outline

1. Introduction
   - A preview of the technique used in the film industry
   - Definition of the human pose
   - Image-based pose recovery
   - Which input data can be used ?
2. Human models
   - 2D human models
   - 3D human models
3. Pose recovery methods
   - Model-based methods (generative methods)
   - Example-based methods
   - Learning-based methods (discriminative methods)
4. State-of-the-art and conclusions
   - An example of state-of-the-art method
   - Conclusions

Outline

1. Introduction
   - A preview of the technique used in the film industry
   - Definition of the human pose
   - Image-based pose recovery
   - Which input data can be used ?
2. Human models
   - 2D human models
   - 3D human models
3. Pose recovery methods
   - Model-based methods (generative methods)
   - Example-based methods
   - Learning-based methods (discriminative methods)
4. State-of-the-art and conclusions
   - An example of state-of-the-art method
   - Conclusions

Outline

1. Introduction
   - A preview of the technique used in the film industry
   - Definition of the human pose
   - Image-based pose recovery
   - Which input data can be used ?
2. Human models
   - 2D human models
   - 3D human models
3. Pose recovery methods
   - Model-based methods (generative methods)
   - Example-based methods
   - Learning-based methods (discriminative methods)
4. State-of-the-art and conclusions
   - An example of state-of-the-art method
   - Conclusions

Motion capture for character animation

[image source: http://franciszgx.wordpress.com]

Motion capture for character animation

Several types:

- passive markers / active markers
- anonymous markers / markers with IDs.

Drawbacks and advantages:

- ☹ Intrusive ($\rightarrow$ field of applications very limited).
- ☹ Often more than $n = 20$ cameras are needed ($\rightarrow$ costly).
- ☹ Manually controlling the matching of markers is done to improve the reliability. This is laborious.
- ☺ Very accurate : the 3D location of a marker is computed by intersecting n lines in the least squares sense.

Too many drawbacks ! It is possible to do something simpler ?

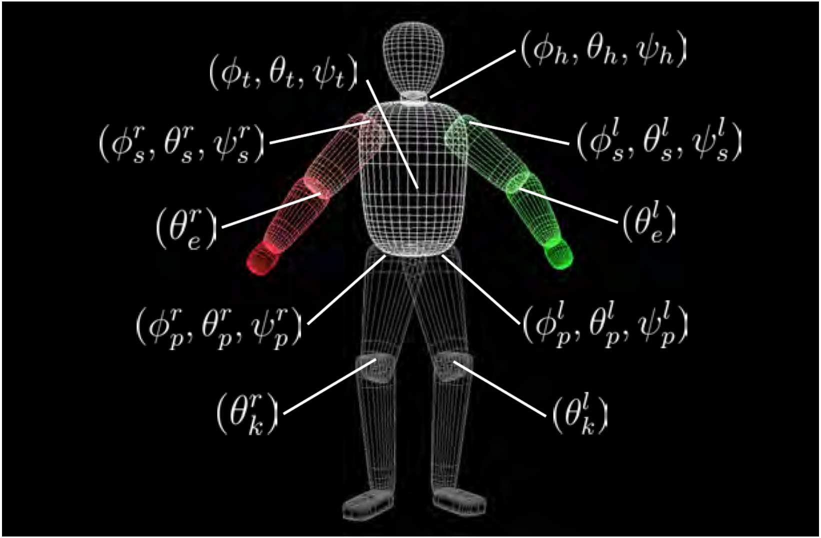
Outline

How can be define the pose ?

- Is it 2D or 3D ?
- Is it related to the position of the person in the 3D scene ?
- Is it related to the orientation of the person in the 3D scene ?

neck :	neck :	neck :
$(u, v) = (100, 200)$	$(x, y, z) = (0.0, 1.6, 0.0)$	$(\theta, \phi, \psi) = (0.0, 0.0, 0.0)$
left shoulder :	left shoulder :	left shoulder :
$(u, v) = (50, 175)$	$(x, y, z) = (0.3, 1.5, 0.0)$	$(\theta, \phi, \psi) = (0.0, 0.2, 0.1)$
left elbow :	left elbow :	left elbow :
$(u, v) = (50, 125)$	$(x, y, z) = (0.3, 1.2, 0.0)$	$(\theta) = (0.35)$
...	...	...

At least, 22 kinematic parameters are needed.



[image source: M. Hofmann, "Multi-view 3D Human Pose Recovery in Complex Environment", 2011]

Outline

1. Introduction
   - A preview of the technique used in the film industry
   - Definition of the human pose
   - Image-based pose recovery
   - Which input data can be used ?
2. Human models
   - 2D human models
   - 3D human models
3. Pose recovery methods
   - Model-based methods (generative methods)
   - Example-based methods
   - Learning-based methods (discriminative methods)
4. State-of-the-art and conclusions
   - An example of state-of-the-art method
   - Conclusions

Is the pose recovery from an image possible ?

Can you estimate their poses ?

[image source: http://images.allmoviephoto.com]

If a human expert is able to estimate the pose from an image, why a computer wouldn't be able to do it too ?

Pose recovery is difficult

The pose recovery is the ability to learn the "function"

range or color image(s) $\rightarrow$ kinematic parameters

Pose recovery from images is a difficult problem:

- the human visual appearance is highly variable (morphology, clothing, lighting, ...)
- occlusions (self-occlusions, occlusions by scene elements)
- high dimension of the input (images $640 \times 480 \Rightarrow \mathbb{R}^{921600}$)
- high dimension of the output (typically $\mathbb{R}^{20} \rightarrow \mathbb{R}^{100}$)
- the function that has to be learned is multivalued
- the kinematic parameters are highly dependent

Preliminary remarks

In engineering (or computer science), it is very easy to solve problems that are linear or that can be approximated as linear.

Examples: camera calibration with a pinhole model, linear filtering.

The relationship between the visual perception of a complex 3D scene and its state variables is *not* linear.

Example: deformations, self-occlusions.

Solving a problem begins by understanding it and choosing an appropriate model for it. Machine learning methods do not eliminate the need for a good understanding.

This is the menu of this introductory course.

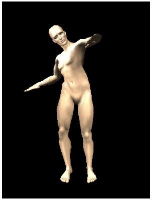
Outline

Is "silhouette $\to$ pose" possible ?

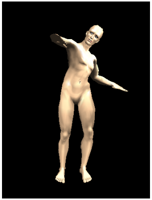
Let us consider:

1. That we observe a person from the side view (i.e. the camera looks horizontally)
2. That the perspective effects are negligible.[1]

To answer the question, we need to consider two mirror poses p_1 and p_2 like these ones:



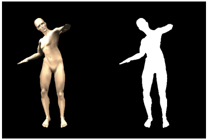
pose p_1



pose p_2

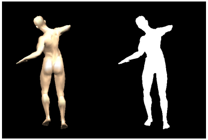
[1] With an orthographic camera, there is no perspective effect. With a pinhole camera that is not too close to the observed person, perspective effects are small.

Silhouettes ambiguities : $(p_1, \theta) \equiv (p_2, 180^\circ - \theta)$

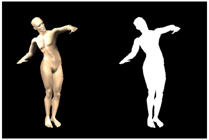


$(p_1, 0^\circ)$

$\Updownarrow$

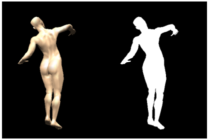


$(p_2, 180^\circ)$

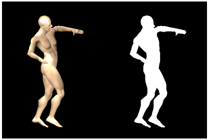


$(p_1, 30^\circ)$

$\Updownarrow$

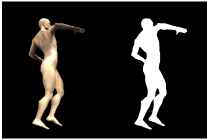


$(p_2, 150^\circ)$



$(p_1, 90^\circ)$

$\Updownarrow$



$(p_2, 90^\circ)$

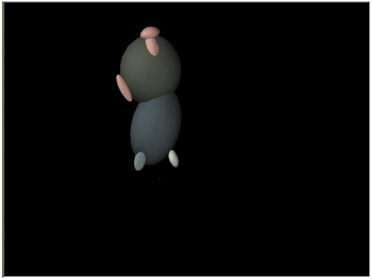
There are always two poses corresponding to a side-view silhouette.
$\hookrightarrow$ Learning correctly "silhouette $\rightarrow$ pose" is impossible.
$\hookrightarrow$ Which supplementary information can be taken into account ?

Outline

1. Introduction
   - A preview of the technique used in the film industry
   - Definition of the human pose
   - Image-based pose recovery
   - Which input data can be used ?
2. Human models
   - 2D human models
   - 3D human models
3. Pose recovery methods
   - Model-based methods (generative methods)
   - Example-based methods
   - Learning-based methods (discriminative methods)
4. State-of-the-art and conclusions
   - An example of state-of-the-art method
   - Conclusions

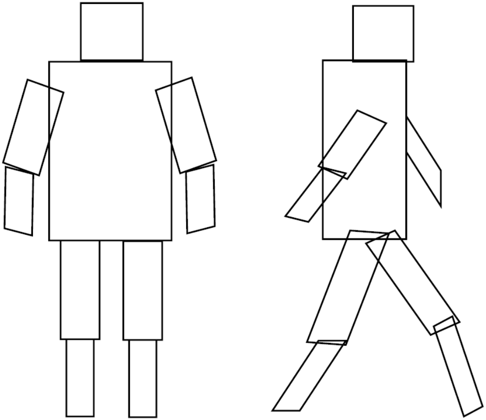
Outline

{2D blobs} (x,y,color gaussians)



[image source: C Wren et al., "Real-time tracking of the human body", 1997]

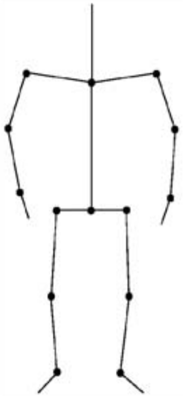
2D “cardboard”

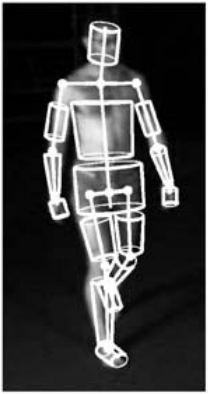


[image source: S Ju et al., "Cardboard people: a parametrized model of articulated image motion", 1996]

Outline

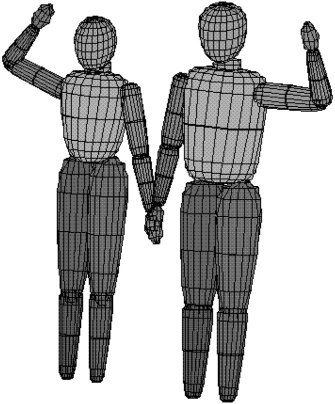
3D cones with elliptical cross-sections





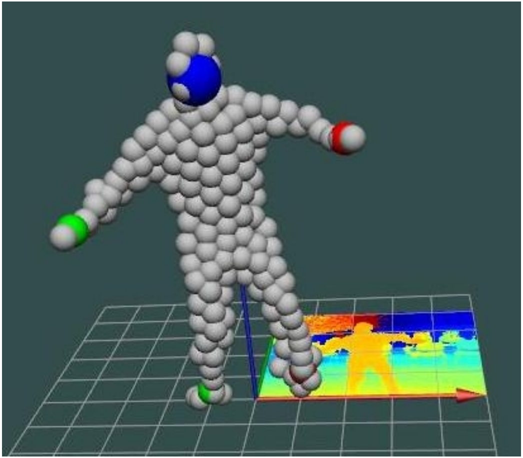
[image source: J. Deutscher & I. Reid, "Articulated body motion capture by stochastic search", 2005]

3D tapered super-quadrics



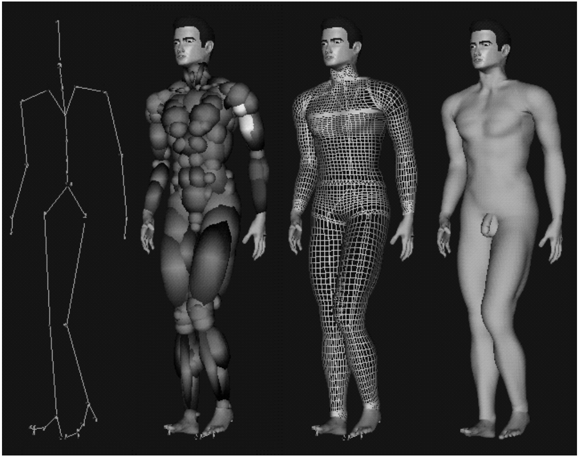
[image source: D. Gavrila & L. Davis, "3-D model-based tracking of humans in action", 1996]

3D volume estimation based on a depth map without explicitly recovering pose parameters:



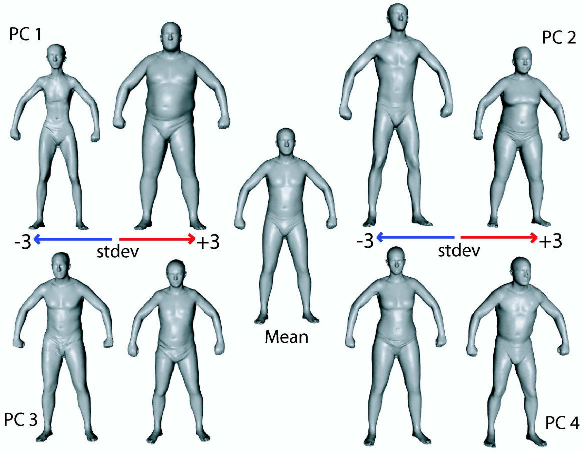
[image source: softkinetic]

Skeleton + {3D gaussians} + polygonal mesh



[image source: R. Plankers & P. Fua, "Articulated soft objects for video-based body modeling", 2001]

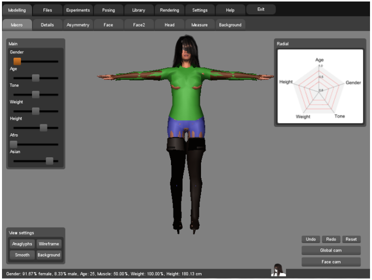
Scape



[image source: D. Anguelov et al., "SCAPE: shape completion and animation of people", 2005]

MakeHuman [1, 5]

version 0.9 (2007)



version 1.0 (201?)

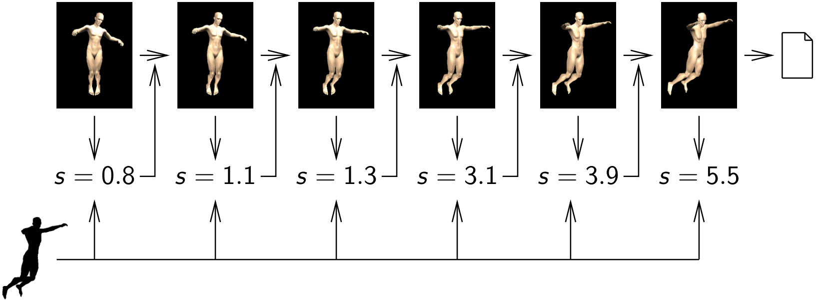
[right image source: http://www.makehuman.org]

Outline

Outline

Model-based methods

Goal: find the pose(s) that maximize the likelihood.



- Can we use a 2D model ? Can we use a 3D model ?
- Does this procedure always converge ?
- How can we define the score s (likelihood).

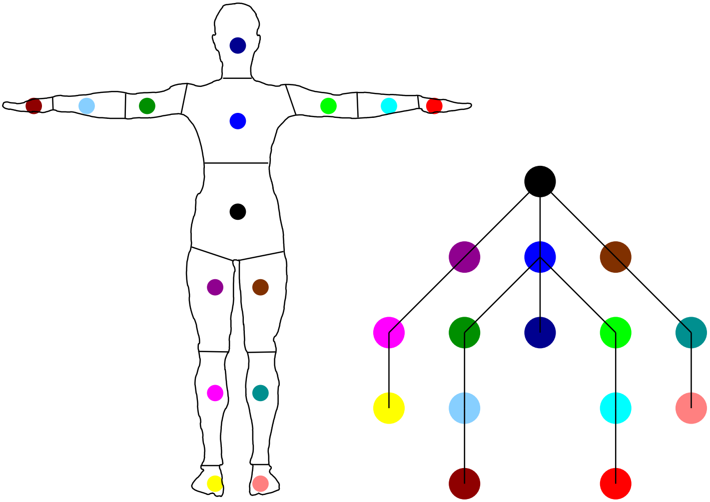
Alternative based on body parts

- Instead of using a model of the full body, one can use a model for each body part (head, hands, feet, legs, arms, torso, etc).
- The goal is then to find all body parts in the image.
- A part is defined by (*location*, *orientation*, *appearance*).
- We can define a score s_{part} for each body part.

$$s = \sum_{part \in parts} s_{part} + s_{coverage} + s_{kinematic} + s_{symmetry} + s_{autointersect} + s_{temporal} + \cdots$$

- $s_{kinematic}$ can be efficiently handled when the human body is considered as a tree, but we are then limited to consider only pairs of connected parts.

The human body can be represented as a tree



The nodes represent the "rigid" body parts, and the links (edges) represent articulations (1, 2, or 3 *dofs*).

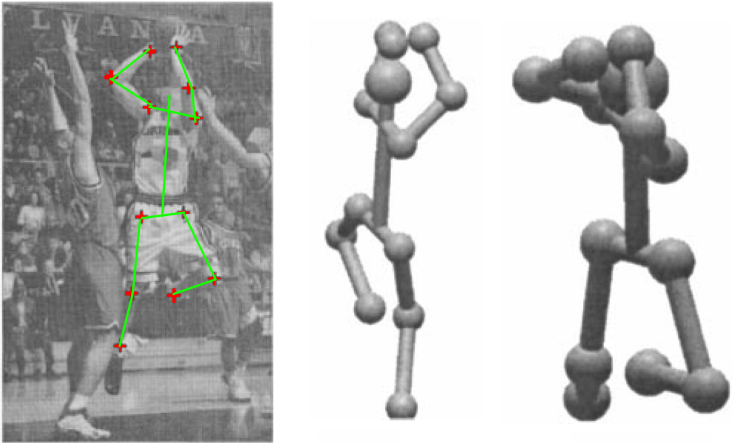
Constraints and cues

- The angles of the joints have limits
- There should not be any self-intersection
- The body is symmetrical
- Clothes are often symmetrical
- Hand and head : skin color
- Gravity center
- Temporal continuity
- Is the activity known ?

Are these constraints are our friends or our enemies?

2D → 3D with Taylor's algorithm [4] I

What can we do if a 2D model has been used instead of a 3D one ?



[image source: C. Taylor, "Reconstruction of Articulated Objects from Point Correspondences ...", 2000]

$2D \rightarrow 3D$ with Taylor's algorithm [4] II

Let $(x_1, y_1, z_1) \longleftrightarrow (x_2, y_2, z_2)$ be a 3D rigid segment of length l, and $(u_1, v_1) \longleftrightarrow (u_2, v_2)$ its 2D projection. We assume an orthographic camera for which 1 *m* corresponds to *k pixels*.

$$\begin{cases} l^2 & = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 \\ u_1 - u_2 & = k\ (x_1 - x_2) \\ v_1 - v_2 & = k\ (y_1 - y_2) \end{cases}$$

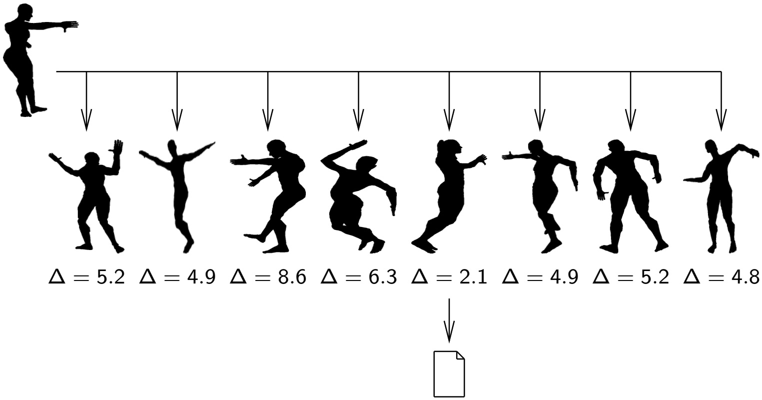
$$\Rightarrow \qquad z_1 - z_2 = \pm\sqrt{l^2 - \left(\frac{u_1 - u_2}{k}\right)^2 - \left(\frac{v_1 - v_2}{k}\right)^2}$$

There are 2^n possible 3D skeletons corresponding to a 2D stick-figure with n links (not all are physically possible).

Outline

Example-based methods

With a database { (pose parameters , visual data) } :

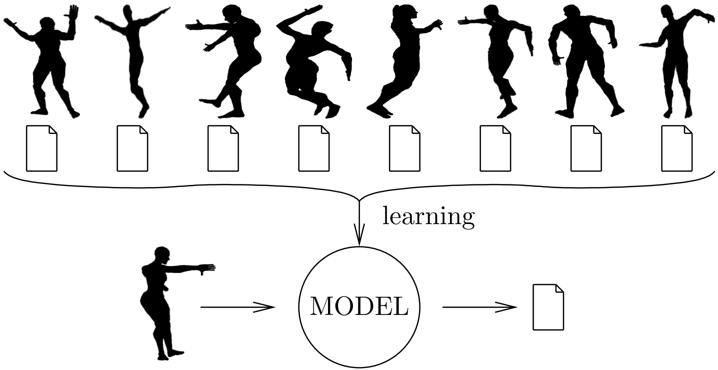


- Distance Δ between visual data is needed: global or by parts ?
- How many samples should we place in the database ?
- How can we obtain a database of samples ?
- Is this method fast enough ?

Outline

Learning-based methods

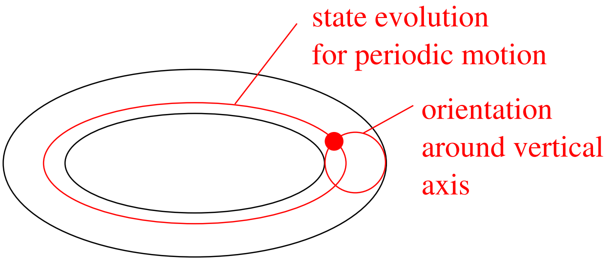
generalization + fast thanks to the pre-computing of the model:



- We need to describe the visual data to obtain attributes.
- What happens if there are ambiguities ?

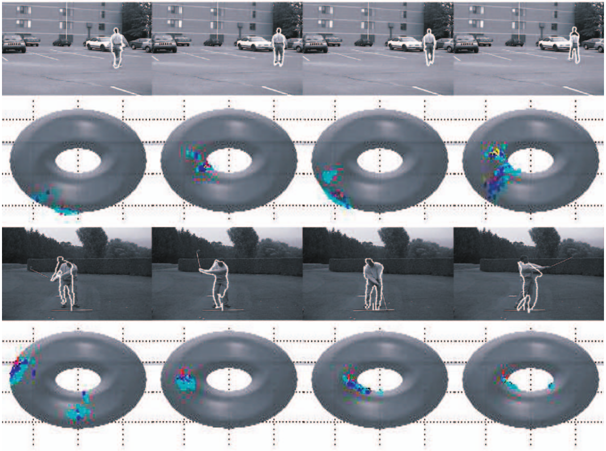
If the movement is known ... manifold learning I

A movement $\equiv$ a state machine with continuous evolution. The observation depends on the state and the orientation (2 d.o.f.).



1. Learning the visual manifold : (*state*, *orientation*) $\rightarrow$ *visual*
2. Learning the kinematic manifold *state* $\rightarrow$ *pose*

If the movement is known ... manifold learning II



[image source: A. Elgammal, "Tracking People on a Torus", 2009]

Outline

Outline

What is behind your friendly *Xbox-Kinect* application ?

	ICVPR Jun. 2011 [3]	ICCV Nov. 2011 [2]
FPS on the Xbox GPU	~ 200	?
FPS on a 8 core desktop CPU	~ 50	~ 200
body parts	31	
images in LS	900.000	$15.000 \to 300.000$
% of joints $< 10\ cm$ error	73.1	$73.6 \to 79.9$

Homework : Read and understand [3] ! You can download it at `http://research.microsoft.com/apps/pubs/default.aspx?id=145347`.

What is behind your friendly *Xbox-Kinect* application ?

(click here to play video)

[video source: webpage of Microsoft Research]

Outline

Conclusions

- Human pose recovery with a single camera has a lot of applications.
- A separation between intrinsic parameters (kinematic parameters) and extrinsic parameters (view-point, color, texture, etc) is often preferred: *cf.* the manifolds and the likelihood scores.
- There is not enough informations in binary silhouettes.
- There are three kinds of methods : learning, example, and model-based.
- Human pose recovery based on color images is a challenge.
- Human pose recovery with a range camera works very well.

Bibliography I

- M. Bastioni, S. Re, and S. Misra.
  Ideas and methods for modeling 3D human figures: the principal algorithms used by MakeHuman and their implementation in a new approach to parametric modeling.
  In *Proceedings of the 1st Bangalore Annual COMPUTE Conference*, pages 10.1–10.6, Bangalore, India, 2008. ACM.

- R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon.
  Efficient regression of general-activity human poses from depth images.
  In *International Conference on Computer Vision (ICCV)*, Barcelona, Spain, November 2011.

J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake.
Real-time human pose recognition in parts from single depth images.
In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, June 2011.

C. Taylor.
Reconstruction of articulated objects from point correspondences in a single uncalibrated image.
Computer Vision and Image Understanding, 80(3):349–363, 2000.

The MakeHuman team.
The MakeHuman website.
`http://www.makehuman.org`, 2007.