Master in Artificial Intelligence (UPC-URV-UB)

HUMAN POSE RECOVERY AND BEHAVIOR ANALYSIS FROM RGB AND DEPTH MAPS

Student: Miguel Reyes Estany

Advisor: Sergio Escalera Guerrero

Instituto Fisioterapia Global Mèzieres

Computer Vision Center

Fundació Bosch i Gimpera

Universitat de Barcelona

Corporació Parc Taulí

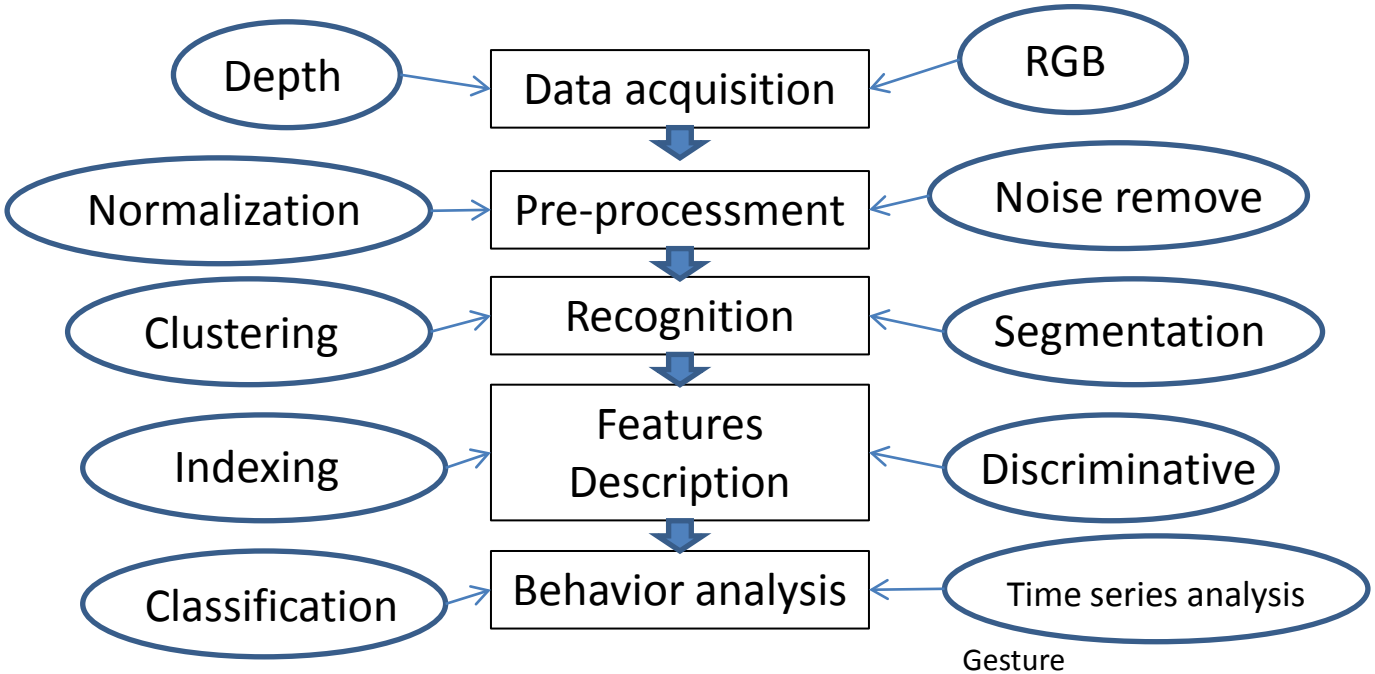
Outline

- Introduction
- RGB-D
- Modeling cameras
- Pose recovery
- Gesture recognition
- Demo
- Conclusions

Introduction

•Automatic Human Pose Recovery and Behavior Analysis from different input sensor data is an open challenging problem.

•This problem has been treated for several years in the fields of Computer Vision, Pattern Recognition, and Machine Learning.

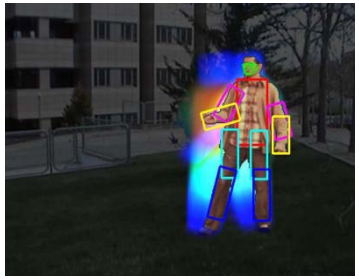


Motivation

Up to now, there no exist a direct solution for the mentioned problem, and achieving these goals will allow the design of new technologies that can open several lines research and application.

Among these applications one can find:

- Entertainment.
- Intelligent searches for content retrieval of video data per content.
- Security, video surveillance.
- Health care, augmented autonomy of people with different physical and mental diseases, advanced assisted living, impatient monitoring in hospitals, behavior supported diagnosis.



Introduction

Modeling cameras

RGB-D

Pose recovery

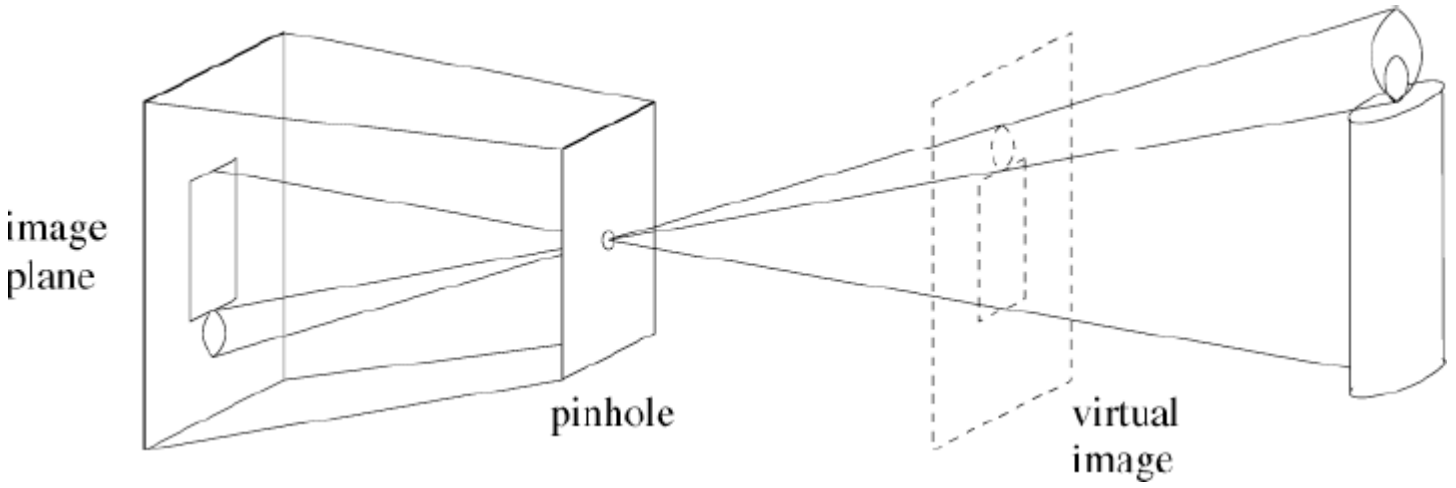
Gesture recognition

Conclusions

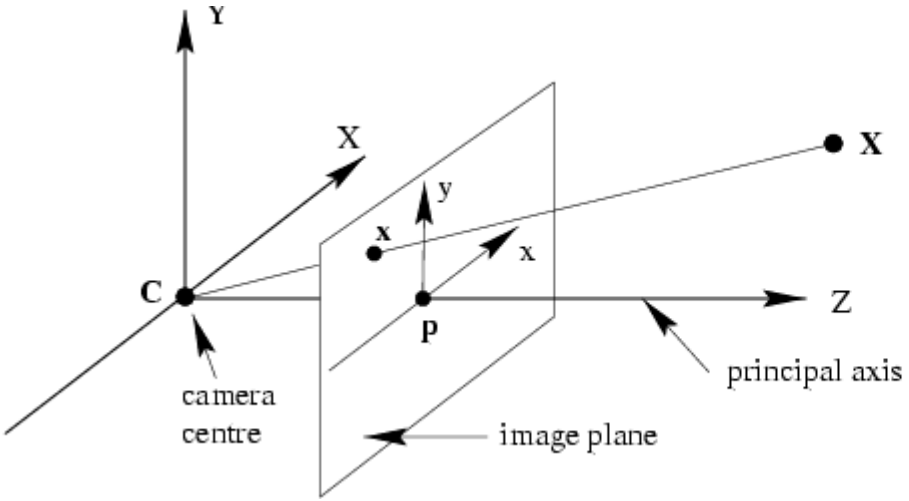
Modeling Cameras

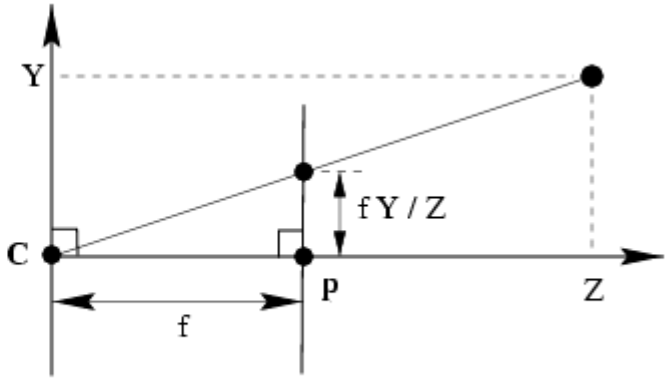
• We need the geometric relationship between a real world (3D) object point and its 2D corresponding projection onto the image plane.

• Controlling each of the parameters that make up the geometric model, we can recover 3D data to treat over 2D image. Cameras are modeled as pinhole.



Pinhole Camera Model

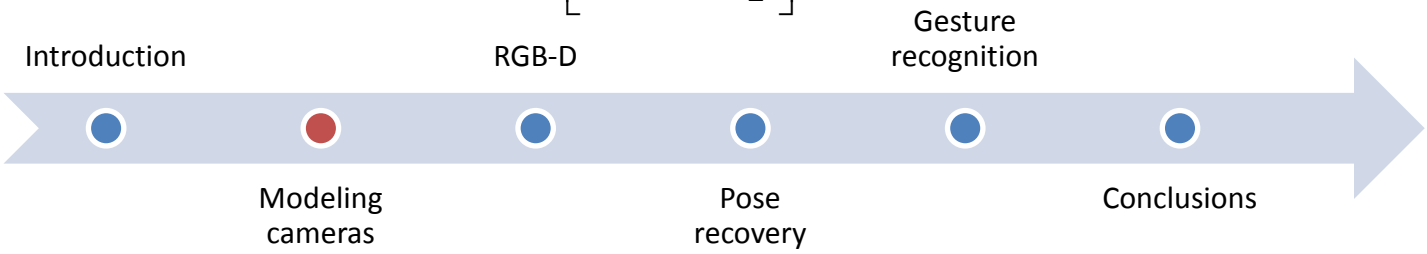


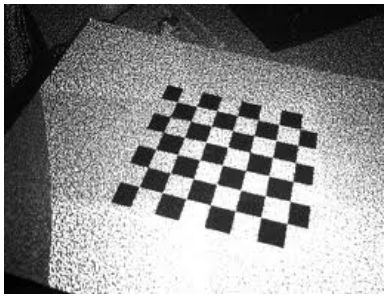


The reality is modeled by the extrinsic and intrinsic parameters

- Extrinsic parameters denotes the position of the camera relative to the scenario.
- Intrinsic parameters denotes the features of the camera lens.

$$K = \begin{bmatrix} f_x & s & p_x \\ & f_y & p_y \\ & & 1 \end{bmatrix}$$





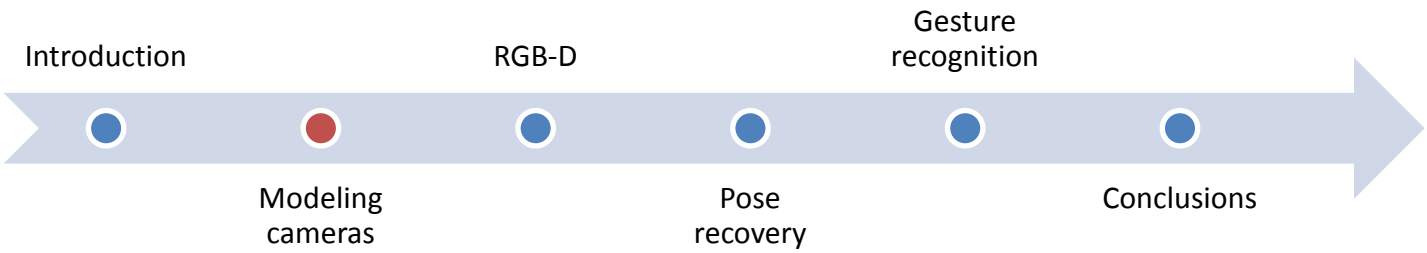
Calibration

•Zhang[1] proposed a calibration technique based on the observation of a flat template from various positions.

• To calibrate the camera with this method is necessary to estimate the homographies of each of the images taken from the template.

• Checkerboards with black-and-white squares are most widely used because the easy sub-pixel detection algorithm for X-corners with high precision.

1. Zhang Z. (2002) Camera calibration with one-dimensional objects. Technical Report MSR-TR-2001-120 Microsoft research.



Lens distortion

- Reduce these imperfections in order to obtain very realistic image.

- **Radial distortion:** The visible effect is that lines that do not go through the centre of the image are bowed inwards, towards the centre of the image, like a pincushion.

- **Tangential distortion:** The apparent effect is that of an image which has been mapped around a sphere.

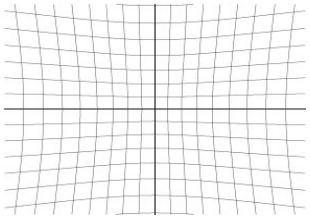
Brown's distortion model:

$$x_u = x_d + (x_d - x_c)(K_1 r^2 + K_2 r^4 + \ldots) + (P_1(r^2 + 2(x_d - x_c)^2) + 2P_2(x_d - x_c)(y_d - y_c))(1 + P_3 r^2 + \ldots)$$

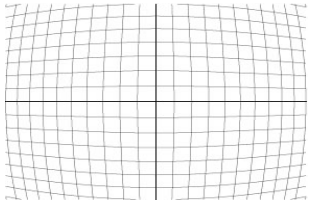
$$y_u = y_d + (y_d - y_c)(K_1 r^2 + K_2 r^4 + \ldots) + (P_2(r^2 + 2(y_d - y_c)^2) + 2P_1(x_d - x_c)(y_d - y_c))(1 + P_3 r^2 + \ldots)$$

(x_u, y_u) undistorted point
(x_d, y_d) distorted point
(x_c, y_c) center of distortion
$K_n = n^{th}$ radial distortion coefficient
$P_n = n^{th}$ tangential distortion coeficient
$r = \sqrt{(x_d - x_c)^2 + (y_d - y_c)^2}$

Radial distortion

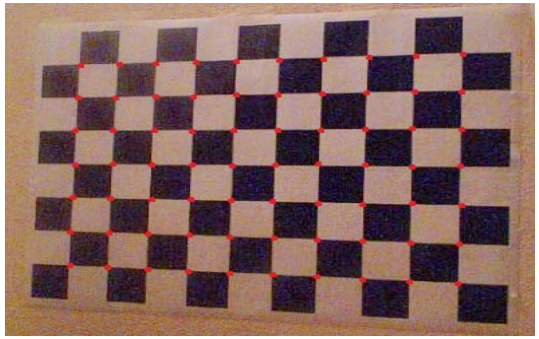


Tangential distortion



Calibration Results: World coordinates

- Two-dimensional template of size: 115,5cmx84cm
- RGB camera resolution 640x480 pixels
- OpenCv library

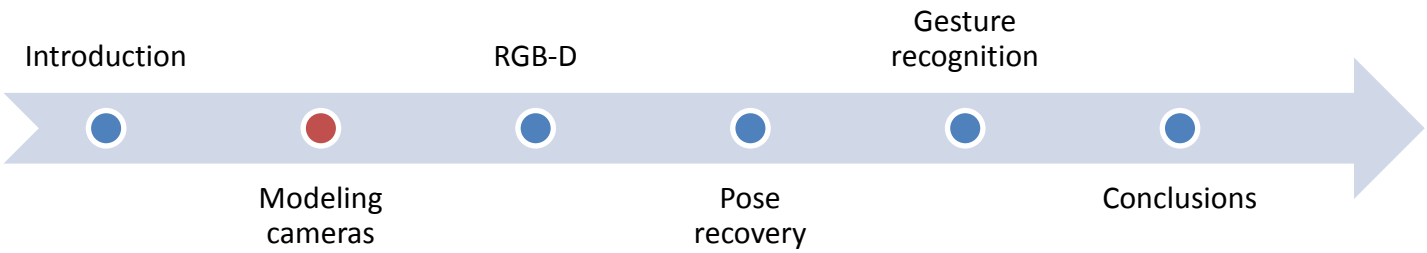


Corners	Pixel postion		Real coord.		Computed coord.		Directional error	
	u (pixels)	v (pixels)	x (cm)	y(cm)	X (cm)	Y (cm)	X (cm)	Y (cm)
1	52	32	18.5	8.5	18.25634	8.785612	0.2436	-0.285
2	108	65	29	19	29.45855	19.32213	-0.4585	0.322
3	166	100	39.5	29.5	39.81254	29.49242	-0.3125	-0.008
4	226	140	50	40	49.95812	40.28745	0.0418	0.287
5	286	182	60.5	50.5	60.65413	50.12544	-0.1541	0.375
6	347	225	71	61	71.63214	61.48521	-0.6432	-0.485
7	408	272	81.5	71.5	81.1256	71.36515	0.3744	0.135
8	471	322	92	82	92.2325	82.36541	-0.2325	-0.365
9	534	377	102.5	92.5	102.5478	92.87125	-0.0478	-0.371
10	597	437	113	103	112.8542	103.3548	0.1458	0.354

Average accuracy:

$$Q_k = \frac{\sum_{i=1}^{m} \sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2}}{m}$$

The average value is 0.3547cm



Applications

Accurate photometry

AR advertisement

RGB-D, Vision advantage

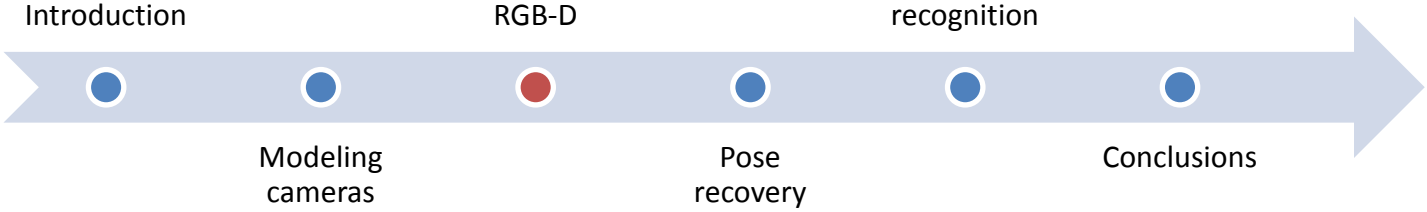
• RGB-D cameras are novel sensing systems that capture RGB images along with per-pixel depth information.

• Currently are they being packaged in form factors that make them attractive for research.

• RGB-D cameras have some important drawbacks with respect to 3D mapping:

- They provide depth only up to a limited distance (typically less than 5m).
- Their depth estimates are very noisy
- Their reduced field of view (≈60°)

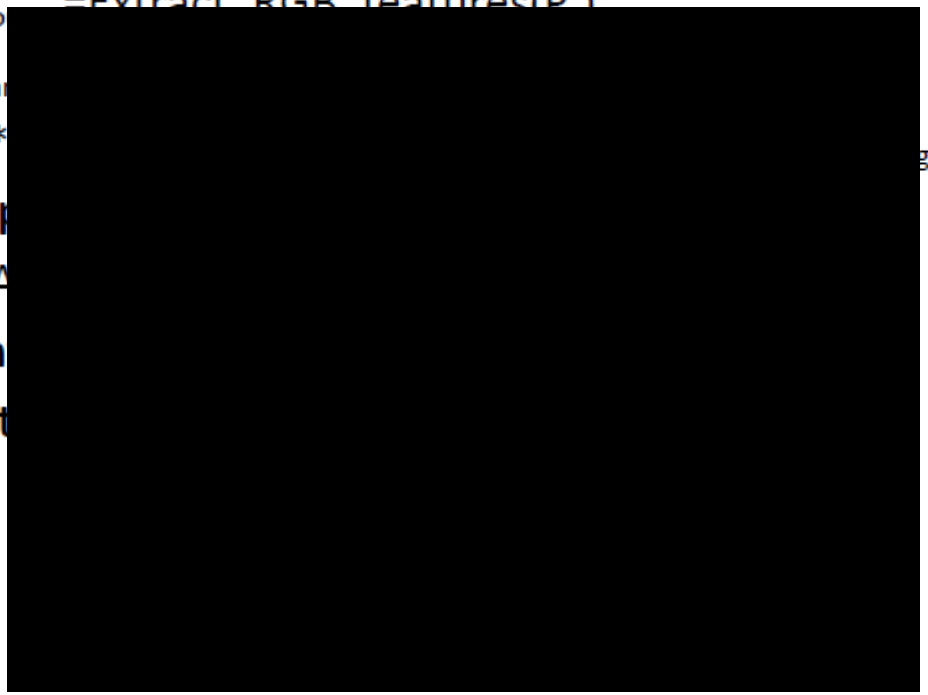
Gesture recognition

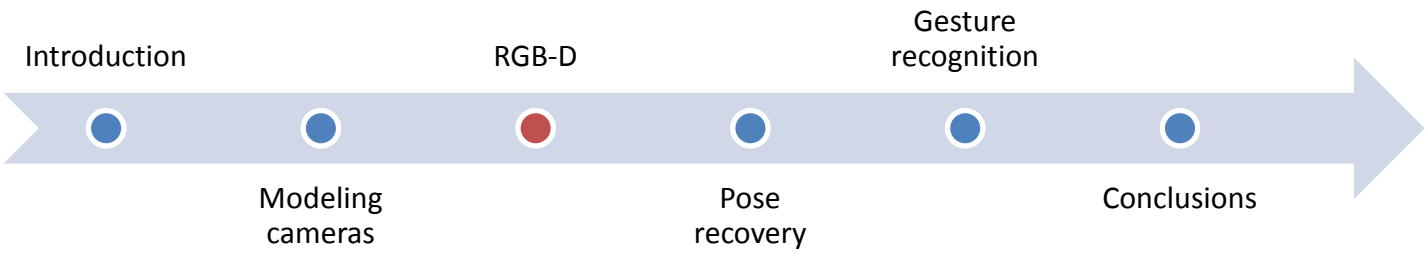


RGB-D Mapping

RGBD-ICP (P_s,P_t):

1. F_{so} =Extract_RGB_features(P_s)
2. F_{ta}
3. (t* ... get)
4. **rep**
5. A
6. **un**
7. **ret**

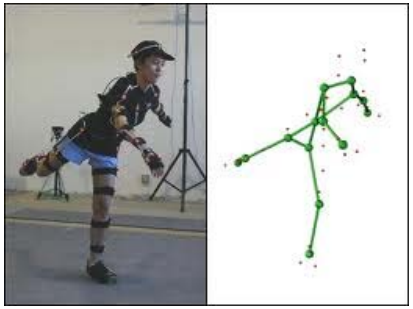




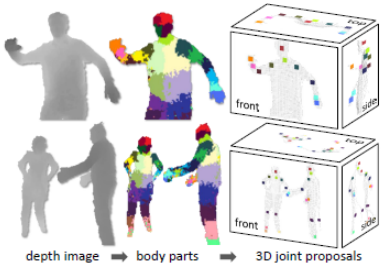
Modeling Human Pose

GOAL : Get an accurate model based on the configuration of a human body posture.

- Two proposals:
  - Get the articulated model based on the manual placement of markers.



  - Creating a skeletal model based on 15 joints examining the RGB-depth map.



Introduction

RGB-D

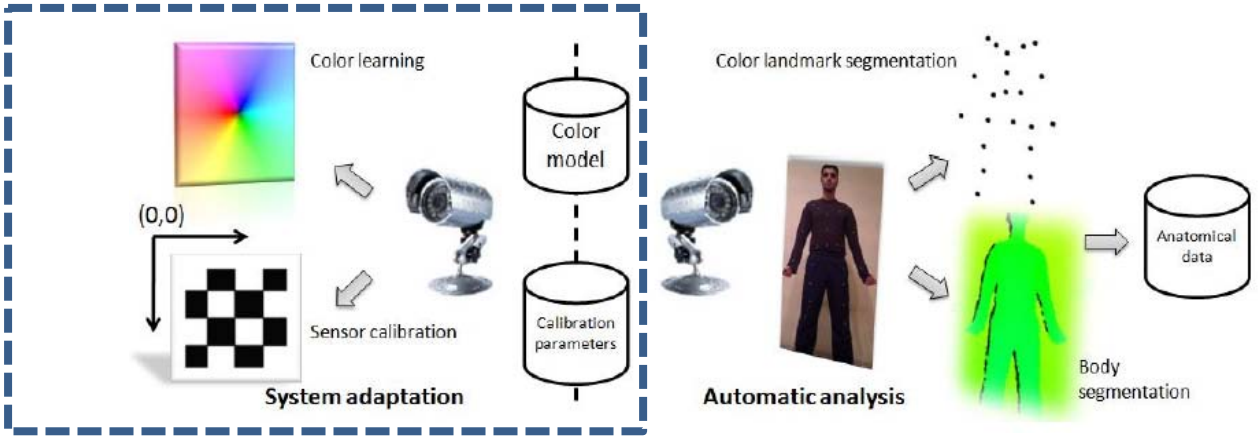
Gesture recognition

Modeling cameras

Pose recovery

Conclusions

Human Pose by landmarks



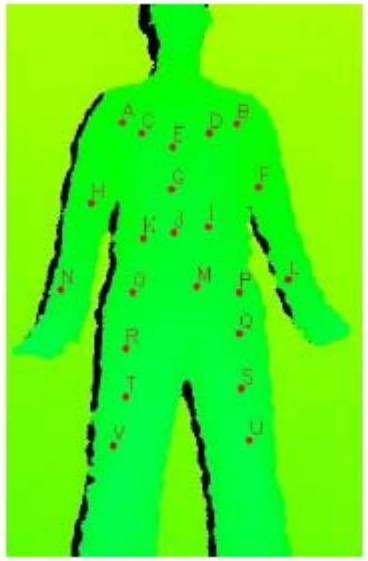
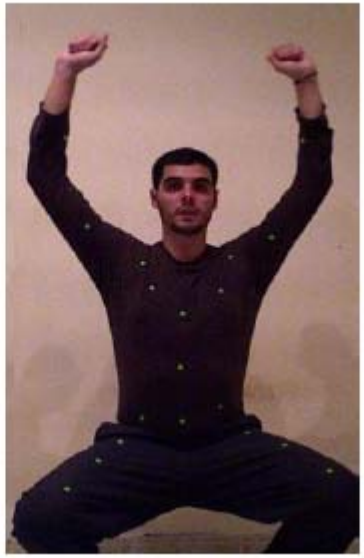
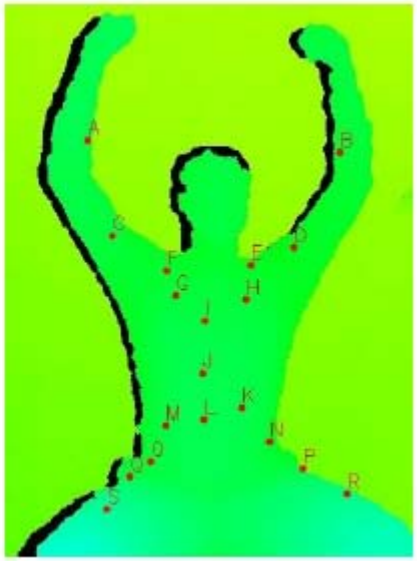
- **Scene Calibration** (pinhole model)
- **Color Learning**: measures the color difference in CIELAB color with a combination with the channel Hue from HSV space color.

$$v = v \cup \forall_j \mid j \in N_v, d(x_j, x_i) < \theta$$

- The **CIELAB-H** color model of pixels is then clustered into ***k*** clusters using *k*-means algorithm.

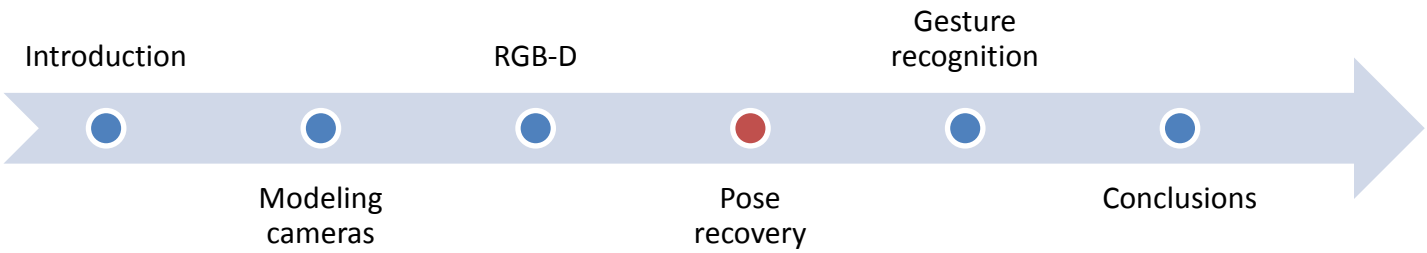
$$S_i^t = \{ j : \left\| x_j - m_i^t \right\| \leq \left\| x_j - m_{i*}^t \right\| \forall i* \in [1,\ldots, j] \}$$

Human Pose by landmarks results

- Number of pictures: 70
- Number of landmarks: 22

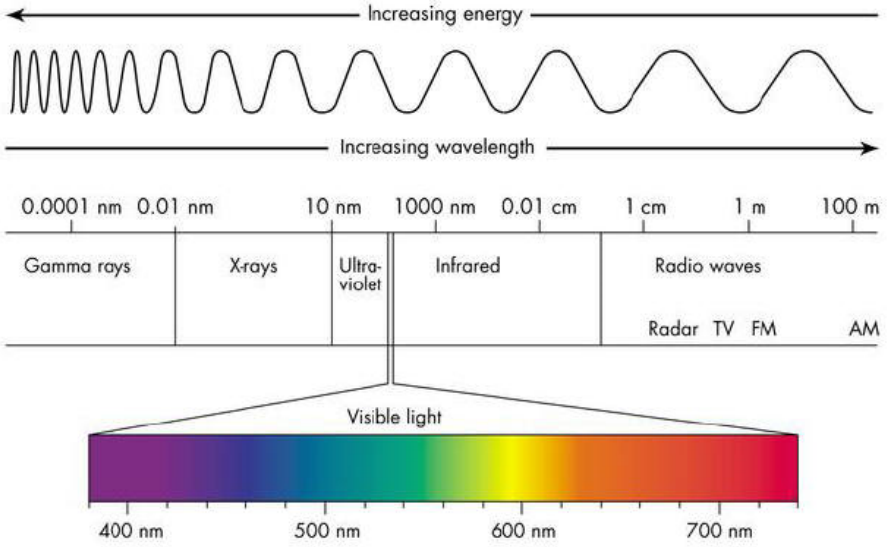
Markers detection	mm	Degree
76,3%	2,3±1,2	3,7±2,4

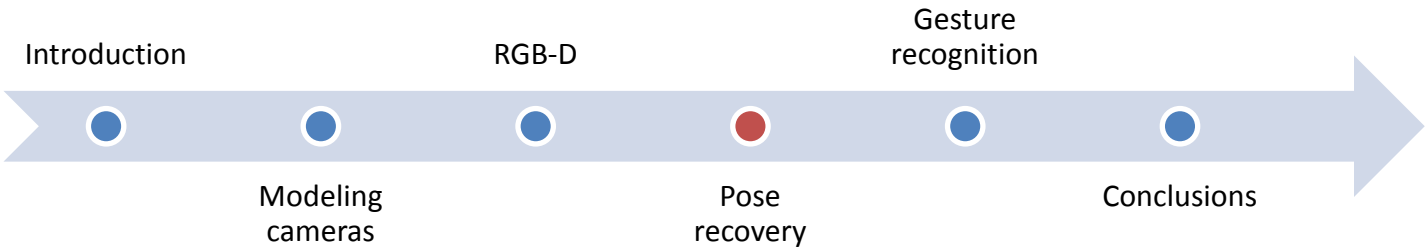


Improving the landmarks recovery system

GOAL: create a robust system against different lighting conditions

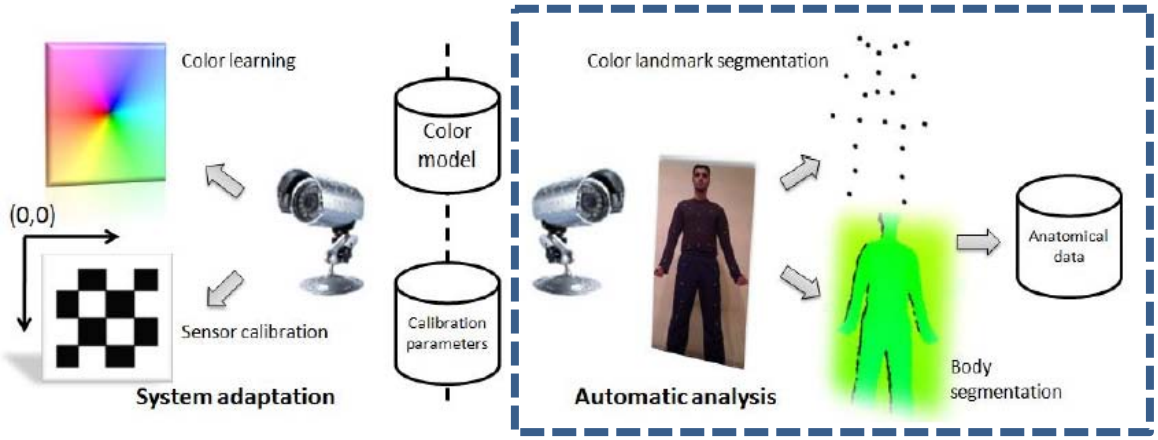
- Colored markers replaced by infrared LED
- Introduced a cut-off filter (820 nm)





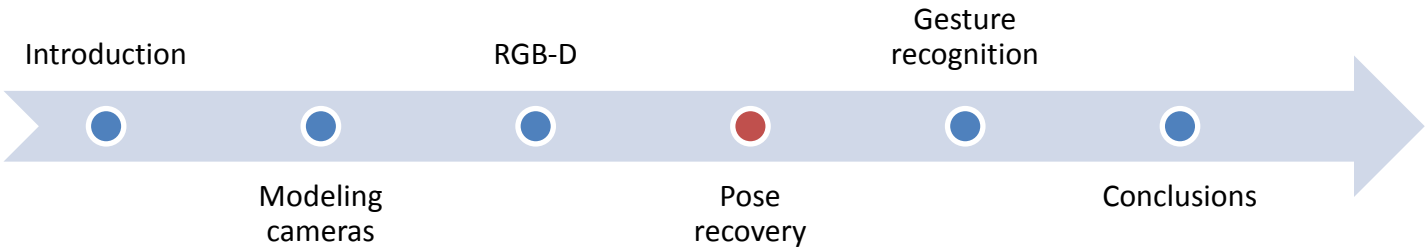
Postural Consistency

GOAL: characterize the output as a consistent feature vector



• **Shape Context**: samples points from generates histograms, and are used to match the most suited point from the first drawing to find a perfect match.

• **Active shape models (ASMs):** are statistical models of the shape of postures which iteratively deform to fit to an example of the shape based in its principal component analysis; $x = \bar{x} + Pb$



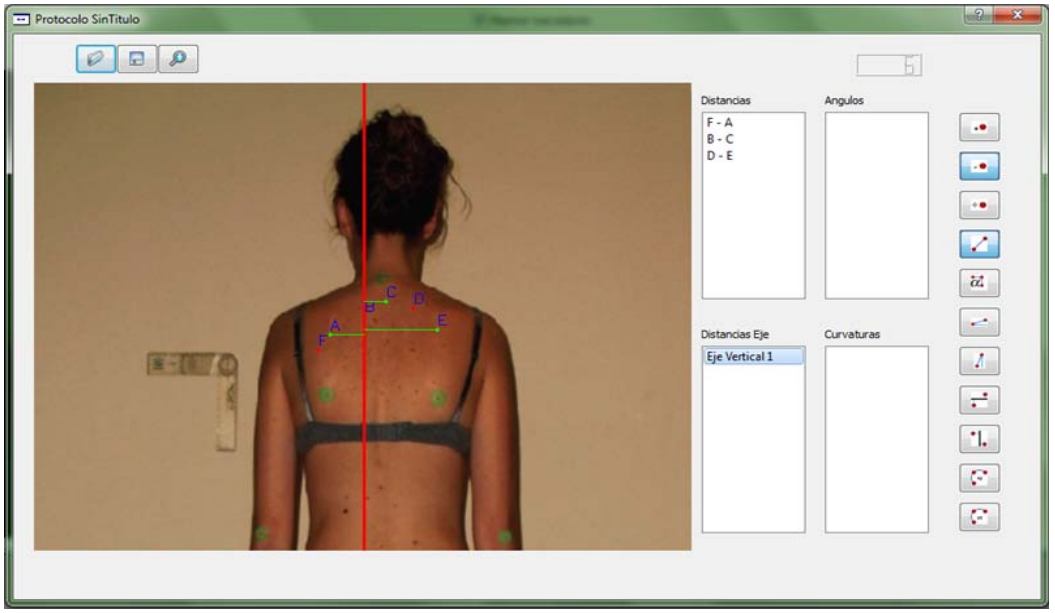
Applications

ADIBAS posture
Automatic Digital Biometry Analysis System

www.adibas.es

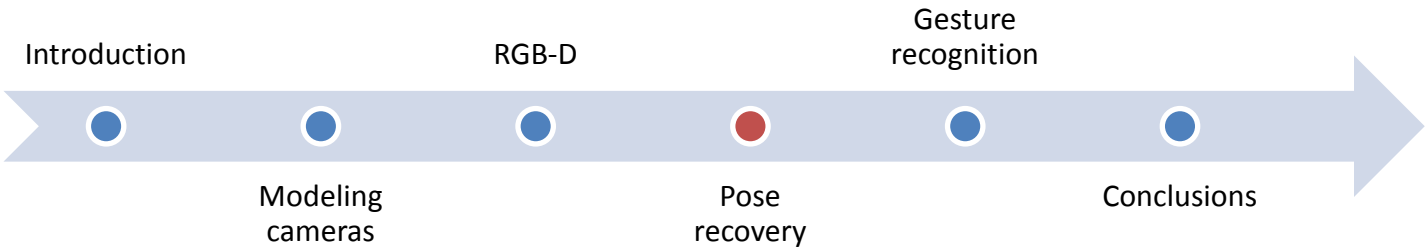
Features:

- Non-intrusive
- Full automatic
- Medical analysis support
- High accuracy
- Easily installable
- Low cost



Software and Hardware:

- Implemented C/C++
- RGB-D technology
- OpenCv library
- OpenNI middleware
- Nokia Qt Interface



Introduction

Modeling cameras

RGB-D

Pose recovery

Gesture recognition

Conclusions

Human Pose using depth maps

Generic framework for human segmentation using depth maps based on Random Forrest and Graphcuts theory, and apply it to the segmentation of human limbs.

Introduction

Modeling cameras

RGB-D

Pose recovery

Gesture recognition

Conclusions

Framework segmentation in depth maps

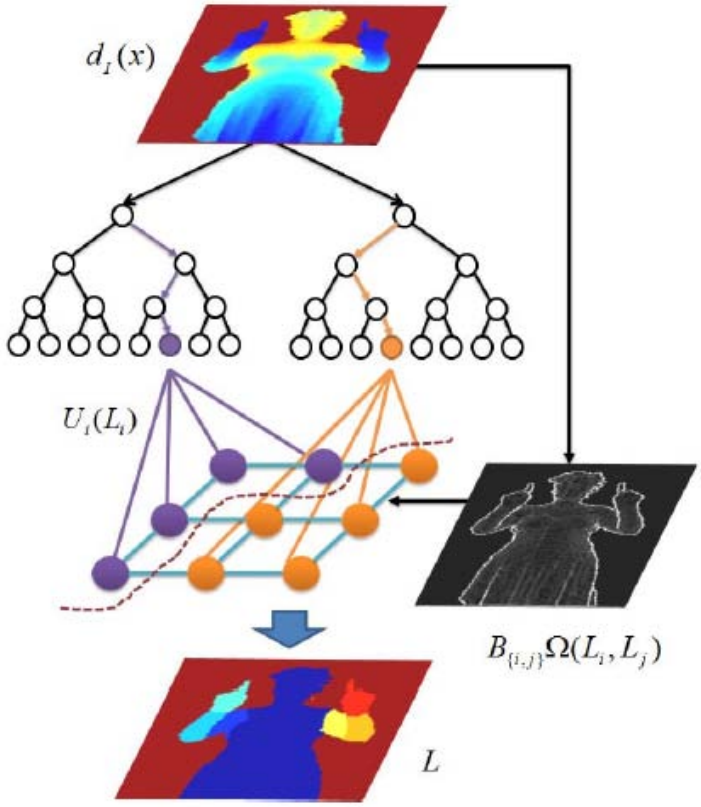
•The pose recognition is addressed by re-projecting the pixel classification results and inferring the 3D positions of several skeletal joints using RF based in the work of Shoton[1].

$$f_\theta(I,x) = \mathrm{d_I}\left(\mathrm{x} + \frac{\mathrm{u}}{\mathrm{d_I(x)}}\right) - \mathrm{d_I}\left(\mathrm{x} + \frac{\mathrm{v}}{\mathrm{d_I(x)}}\right)$$

•The second stage is a globally optimum segmentation of human limbs in depth images with GC.

$$E(L) = U(L) + \lambda B(L)$$

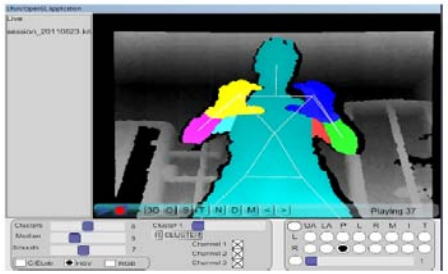
$$U(L) = \sum_{i \in P} U_i(L_i). \qquad B(L) = \sum_{\{i,j\} \in N} B_{\{i,j\}} \Omega(L_i, L_j)$$



1. 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. 2011.

Experiments and results

Data: For the purposes of gathering ground truth data, we defined a new data set of different sessions where the actors are performing different gestures with his/her hands.

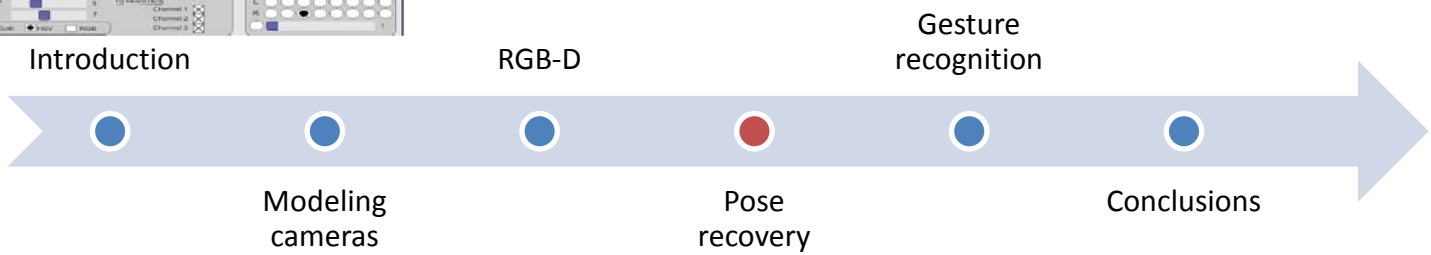


Random forrest results:

	Torso	LU arm	LW arm	L hand	RU arm	RW arm	R hand	Avg. per class
100 f_θ, O_{max} = 30, dt = 20	92.90	73.29	71.42	57.75	74.25	76.26	59.38	72.18
100 f_θ, O_{max} = 60, dt = 20	94.17	79.83	77.69	77.10	81.04	82.65	80.17	81.81
80 f_θ, O_{max} = 60, dt = 20	94.22	79.08	76.46	74.19	81.24	83.26	79.05	81.07
60 f_θ, O_{max} = 60, dt = 20	94.09	78.86	75.86	73.49	79.43	82.60	78.08	80.34
100 f_θ, O_{max} = 60, dt = 15	94.06	79.81	78.69	76.59	81.18	83.10	80.23	81.95
100 f_θ, O_{max} = 60, dt = 10	91.83	81.47	78.98	72.30	83.00	83.74	76.85	81.17
60 f_θ + 20 g_θ, O_{max} = 60, dt = 20	94.04	77.73	74.93	71.97	77.62	81.22	76.64	79.17

Graph-Cut segmentation results:

	Torso	LU arm	LW arm	L hand	RU arm	RW arm	R hand	Avg. per class
Depth, $\Omega_1(L_i, L_j)$	98.86	75.05	82.87	91.45	77.57	87.35	93.96	86.73
Depth, $\Omega_2(L_i, L_j)$	98.86	0.7503	83.36	92.41	77.54	87.67	94.20	87.01
RGB+Depth, $\Omega_1(L_i, L_j)$	99.02	72.02	81.86	90.29	76.56	86.84	92.14	85.53
RGB+Depth, $\Omega_2(L_i, L_j)$	99.02	72.03	81.95	91.19	76.53	87.12	92.12	85.71



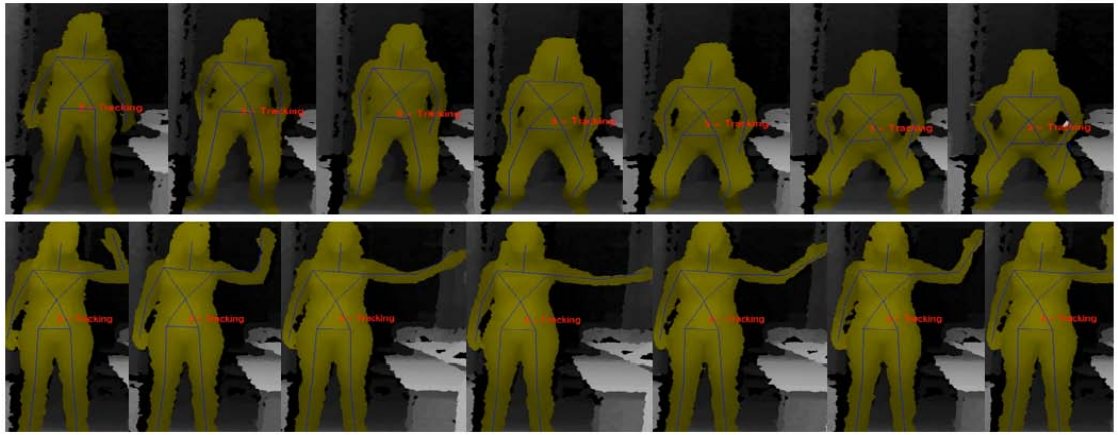
Gesture Recognition

• We attempt to identify simple gestures that a person might perform with his or her full body.

• Is used the vector of features obtained from the skeletal model described in the previous chapter

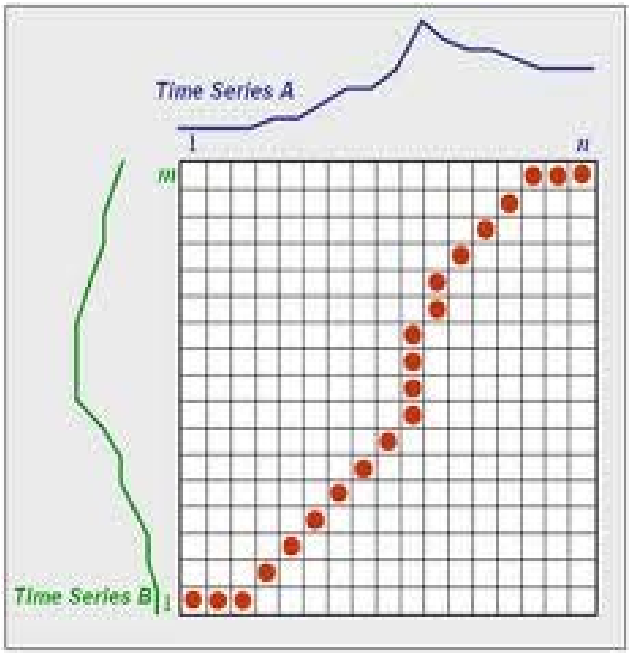
$$V_j = \{\{v_{j,x}^1, v_{j,y}^1, v_{j,z}^1\}, \ldots, \{v_{j,x}^{14}, v_{j,y}^{14}, v_{j,z}^{14}\}\}$$

• Our proposal is focused within the Dynamic Time Warping framework



Dynamic Time Warping Approach

DTW[1] is an algorithm for measuring similarity between two sequences which may vary in time or speed.



Was defined to match temporal distortions between two models, finding an alignment warping path between the two time series:

$$A = \{a_1, \ldots, a_n\}$$
$$B = \{b_1, \ldots, b_n\}$$

In order to align these two sequences M_{nxm} where the position *(i,j)* contains the distance between a_i and b_j.

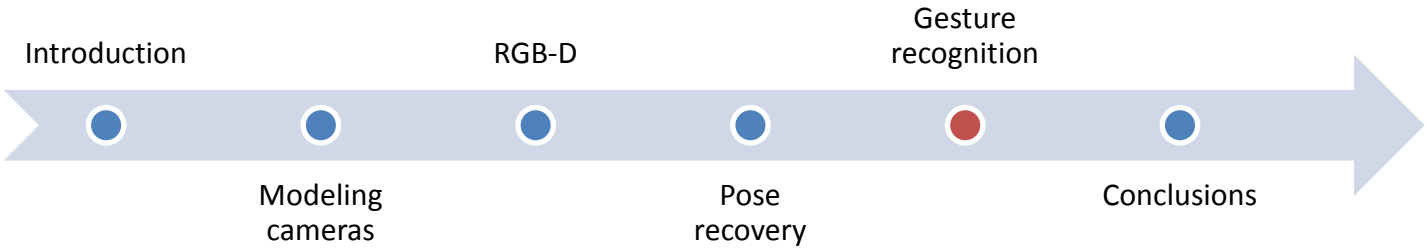
The Euclidean distance is the most frequently applied. Then, a warping path:

$$W = \{w_1, \ldots, w_T\}, \max(m,n) \le T < m+n+1$$

final warping path:

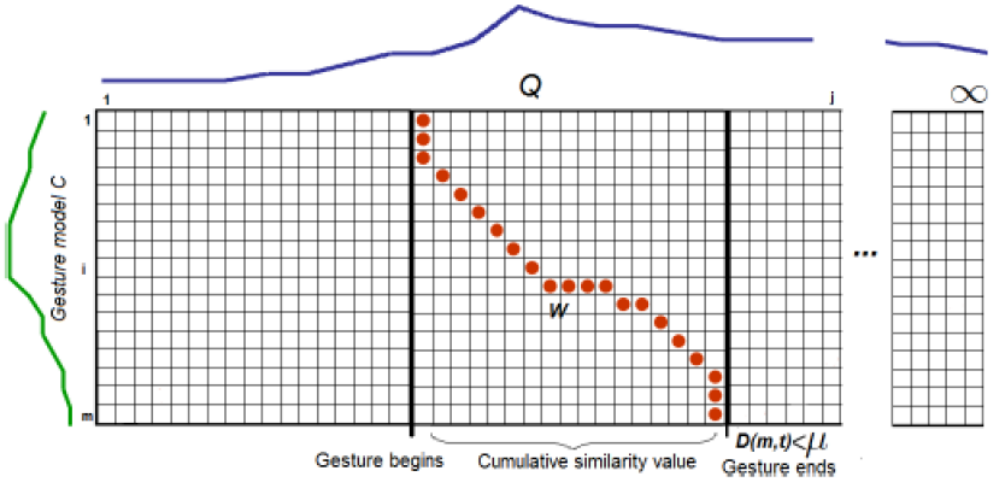
$$DTW = (A,B) = \min\{\frac{1}{T}\sqrt{\sum_{t=1}^{T} w_t}\}$$

1. M. Parizeau and R. Plamondon. A comparative analysis of regional correlation, dynamic time warping, and skeletal tree matching for signature verification.



Dynamic Time Warping Approach Adapted

Dynamic Time Warping allows to align two temporal sequences taking into account that sequences may vary in time based on the subject that performs the gesture. The alignment cost can be then used as a gesture appearance indicator.



$$Q = \{q_1, \ldots, q_\infty\}$$

$$C = \{c_1, \ldots, c_n\}$$

Our aim is focused on finding segments of ***Q*** sufficiently similar to the sequence ***C***.

The system considers that there is correspondence between the current block **k** in ***Q*** and a gesture if satisfying the following condition:

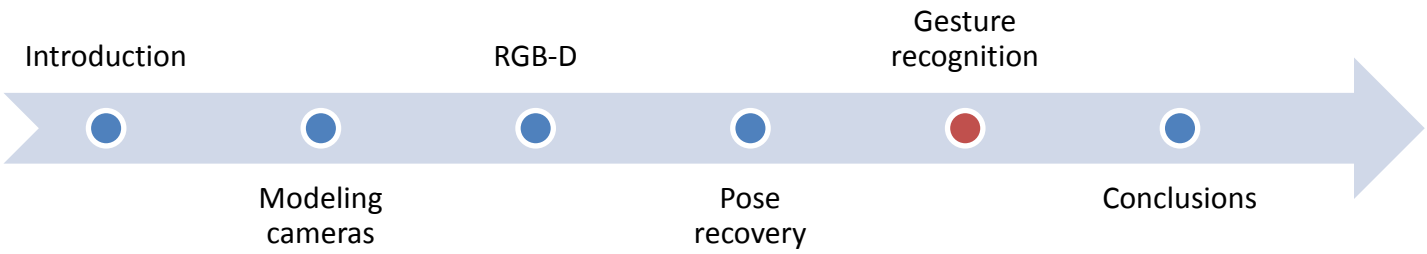
$$M(m,k) < \mu, k \in [1, \ldots, \infty]$$

Feature Weighting DTW

• Is proposed a Feature Weighting approach in order to improve the cost distance computation of the classical DTW framework.

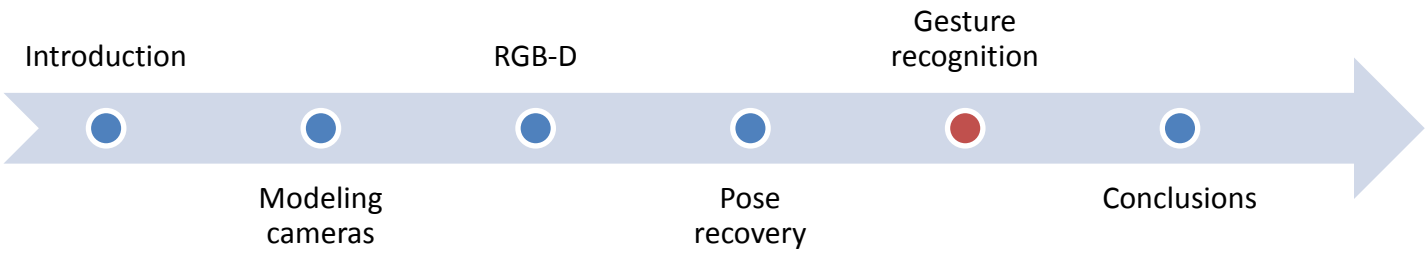
•Associate a discriminatory weight to each joint of the skeletal model depending on an inter-intra gesture similarity measure

• In order to automatically compute this weight per each joint, we propose an inter-intra gesture similarity algorithm.



Feature Weighting DTW Algorithm

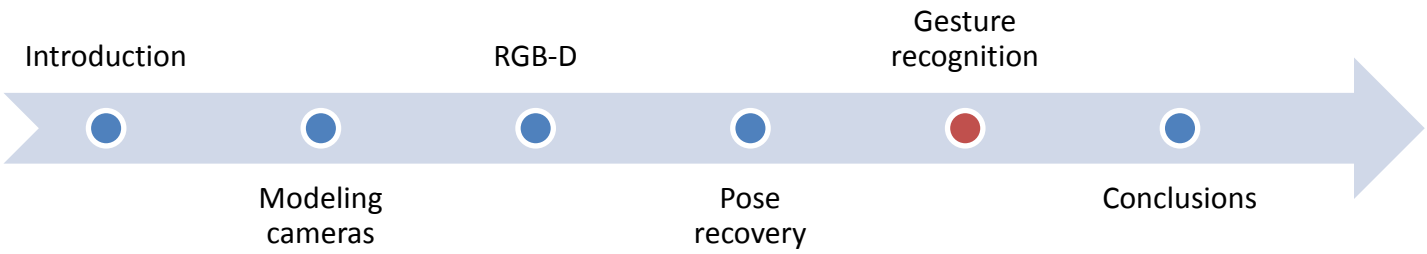
Input: Ground-truth data formed by N sets of gestures $\{n_1, .., n_N\}$.
Output: Weight vector $\nu = \{\nu^1, .., \nu^z\}$ associated with skeletal joints so that $\sum_{i=1}^{z} \nu^i = 1$.
$\nu = \emptyset$
for $p = 1 : z$ **do** // Number of joints
  for $i = 1 : N$ **do**
    for $j = i : N$ **do**
      $D^p(i,j) = \text{mean}(DTW(C_v^i, C_w^j)), \forall v, w$ gesture samples of categories i and j.
    end
  end
  $\nu_{\text{intra}} = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} D^p(i,j)}{\frac{N \times (N-1)}{2}}$ // Computer intra-class variability
  $\nu_{\text{inter}} = \frac{\text{Trace}(D^p)}{m}$ // Computer inter-class variability
  $\nu^p = max(0, \frac{\nu_{\text{intra}} - \nu_{\text{inter}}}{\nu_{\text{intra}}})$ // Compute global weight for joint p
  $\nu = \nu \cup \nu^p$
end
Normalize ν so that $\sum_{i=1}^{z} \nu^i = 1$



Feature Weighting DTW Results

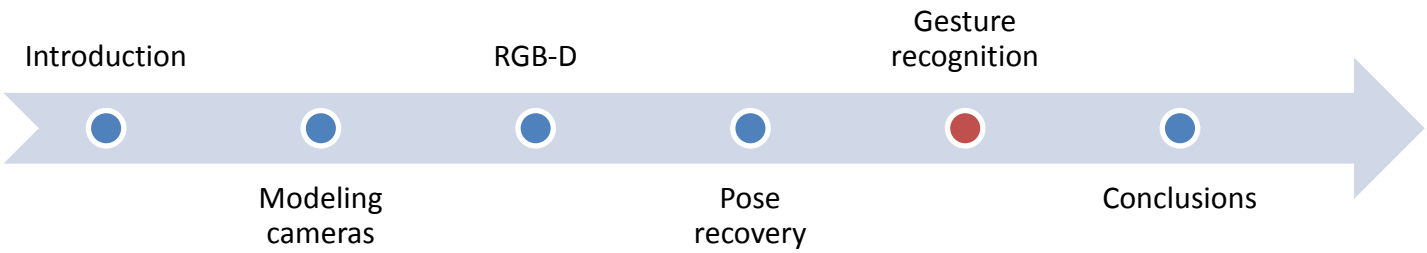
-Data set of five categories: 100 sequences of jumping, bending, clapping, greeting, and noting with the hand
-Performance evaluation: Stratified ten-fold cross-validation

Classification Results Feature Weighting DTW		
Gesture	Begin-end DTW	Feature Weighting
Jump	**68**	**68**
Bend	63.4	**68**
Clap	42	**55**
Greet	64.2	**73**
Note	68	**76**



Feature Weighting DTW DEMO

ADIBAS motion



Conclusions

•In this master thesis we proposed a general methodology for **human pose recovery and behavior** analysis using multi-modal RGB-Depth data.

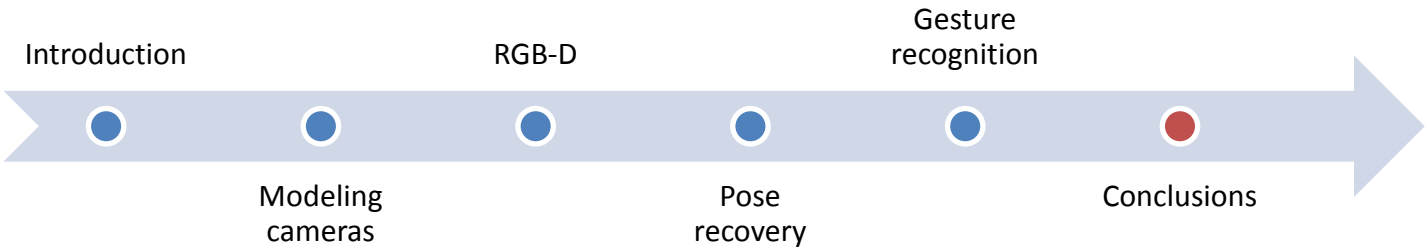
•The camera model is a necessary element if one needs to use the information in images to make measurements of the scene or to make **3D reconstructions**. Is defined a calibration method that is effective and valid **for most calibration problems** that may arise today.

•These methods are based on **image segmentation** using several state-of-the-art technologies, such as **Random Forrest and Graph Cuts theory**.

•The presented results show robust pose segmentation for different configurations and points of view, **improving previous state-of-the-art results in the field**.

•We presented different feature descriptions based on depth map information and also tested **gesture recognition** approaches for time series analysis. Focused on **Dynamic Time Warping**, and we showed that **Feature Weighting improves classical DTW results**.

• The methodolgy explained has been applied in **real scenarios** and using **clinical databases**.



Contributions

1. Miguel Reyes, Gabriel Domínguez, and Sergio Escalera, Feature Weighting in Dynamic TimeWarping for Gesture Recognition in Depth Data, 1st IEEE Workshop on Consumer Depth Cameras for Computer Vision, **International Conference in Computer Vision**, Barcelona, 2011

2. Miguel Reyes, José Ramírez-Moreno, Juan R. Revilla, Petia Radeva, and Sergio Escalera, ADiBAS: Sistema Multisensor de Adquisición Automática de Datos Corporales Objetivos, Robustos y Fiables para el Análisis de la Postura y el Movimiento, **VI Congreso Iberoamericano de Tecnologías de Apoyo a la Discapacidad**, Mallorca, 16/06/2011-17/06/2011.

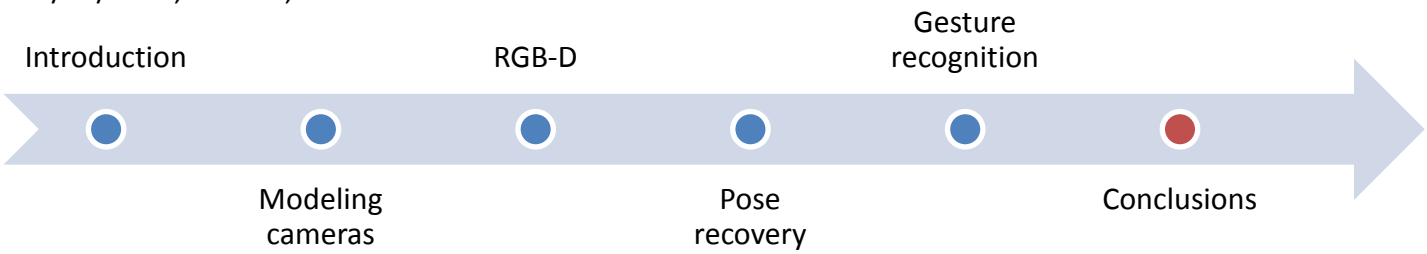
3. M. Reyes, S. Escalera, and Petia Radeva, Real-time head pose classification in uncontrolled environments with Spatio-Temporal Active Appearance Models, **CVCRD'10**, Achievements and New Opportunities in Computer Vision, pp. 101-104, CVC, 29/10/2010, Barcelona, ISBN 978-84-938351-0-1, 2010.

4. Registered software number B3342-11, **ADiBAS Posture**: Automatic Digital Biometry Analysis System, Miguel Reyes, Sergio Escalera, José Ramírez, Juan Ramón Revilla, and Petia Radeva, 2011.

5. Antonio Hernández, Miguel Reyes, Sergio Escalera, and Petia Radeva, Spatio-Temporal GrabCut Human Segmentation for Face and Pose Recovery, IEEE International Workshop on Analysis and Modeling of Faces and Gestures, **Computer Vision and Pattern Recognition**, IEEE Computer Society, 13/06/2010-18/06/2010, San Francisco, ISBN 978-1-4244-7029-7, 2010.

6. Laura Igual, Antonio Hernandez, Sergio Escalera, Miguel Reyes, Josep Moya, Joan Carles Soliva, Jordi Faquet,Oscar Vilarroya, Petia Radeva, Automatic Techniques for Studying Attention-Deficit/Hyperactivity Disorder, Jornada **TIC Salut Girona**, 04/05/2011-05/05/2011, Girona, Spain, 2011.

7. M. Reyes, J. Vitrià, P. Radeva, and S. Escalera, Real-Time Activity Monitoring of Inpatients, **MICCAT**, 28/10/2010, Gerona, 2010.



Thank you very much