Master in Artificial Intelligence (UPC-URV-UB)

Master of Science Thesis

MULTI-MODAL HUMAN GESTURE RECOGNITION COMBINING DYNAMIC PROGRAMMING AND PROBABILISTIC METHODS

Víctor Ponce López

Advisor/s:
Sergio Escalera Guerrero
Xavier Baró Solé

Barcelona

June, 22^{nd} 2012

Acknowledgements

I want to thank the people from different institutions that supported me in the correct development of this work:

People from:

Department of Applied Mathematics and Analysis of the University of Barcelona and the Computer Vision Center, taking part of the Barcelona Perceptual Computing Lab (BCNPCL) virtual group,

with special gratitude to the Human Pose Recovery and Behavior Analysis (HuPBA) group for their support and part of the team work done,

Department of Justice of the Generalitat de Catalunya, and

Fundació Obra Social La Caixa.

Abstract

In this M. Sc. Thesis, we deal with the problem of Human Gesture Recognition using Human Behavior Analysis technologies. In particular, we apply the proposed methodologies in both health care and social applications. In these contexts, gestures are usually performed in a natural way, producing a high variability between the Human Poses that belong to them. This fact makes Human Gesture Recognition a very challenging task, as well as their generalization on developing technologies for Human Behavior Analysis. In order to tackle with the complete framework for Human Gesture Recognition, we split the process in three main goals: Computing multi-modal feature spaces, probabilistic modelling of gestures, and clustering of Human Poses for Sub-Gesture representation. Each of these goals implicitly includes different challenging problems, which are interconnected and faced by three presented approaches: Bag-of-Visual-and-Depth-Words, Probabilistic-Based Dynamic Time Warping, and Sub-Gesture Representation. The methodologies of each of these approaches are explained in detail in the next sections. We have validated the presented approaches on different public and designed data sets, showing high performance and the viability of using our methods for real Human Behavior Analysis systems and applications. Finally, we show a summary of different related applications currently in development, as well as both conclusions and future trends of research.

INDEX

1. Introduction

Nowadays, Human Behavior Analysis (HBA) is a widely considered technology in both health and social applications. As an example, it can be useful for supported diagnosis of mental diseases (e.g. Attention Deficit and Hyperactivity Disorder, which affects about 10% of children and adolescents of the world population), or for the tracking, control, and analysis of gestures on elderly people in order to improve their autonomy (e.g. people with physical incapacities or dementia).

Human Gesture Recognition (HGR) is one of the main challenging areas of Computer Vision and the main core of HBA. Current methodologies have shown robust preliminary results on very simple scenarios. However, these results are still far from human capabilities. On the other hand, the large number of potential applications behind HGR in fields like surveillance [Hampapur et al., 2005; Ivanov et al., 1999; Brown et al., 2005; Howe and Dawson, 1996], Sign Language Recognition [Fang et al., 2007; Yang et al., 2009; Athitsos, 2010], or clinical applications [Pentland, 2005; Wren et al., 1997] among others, is pushing many scientists to devote their efforts in this field of research. The main basis of HBA and HGR comes from psychological analyses, which define the behavioral patterns that a subject presents [Winsor et al., 1997].

Going into more detail, detecting Human Poses (HP) is a main step in the study of HGR. However, HP detection is a challenging task because of the huge inter/intra-limb feature variability in both still images and image sequences. From the point of view of data acquisition, many methodologies treat images captured by visible-light cameras. Computer Vision is then used to detect, describe, and learn visual features. The problem of limb detection and pose recovery becomes even more difficult because of the difficulties of uncontrolled environments: illumination changes, different points of view, or occlusions, just to mention a few.

1.1. Thesis goals

Using Computer Vision feature extraction and Machine Learning approaches our goal in this M. Sc. Thesis is to improve current state-of-the-art strategies for HGR in the field of HBA. In particular, we define the following goals which are addressed in next sections:

1. Use of depth maps for Computer Vision feature extraction to improve the discriminability power of multi-modal descriptors.
2. Probabilistic modelling of temporal multi-modal descriptors for improved recognition of human actions/gestures.
3. Apply the proposed methodology in real and challenging applications.

Given the previous goals, our main objective is to recognize a large number of gestures. The developed system that addresses these goals is composed by the modules shown in Figure 1: a) compute the multi-modal feature-spaces from a large set of videos involving several gestures, b) probabilistic modelling of gestures using the described sequences in a dynamic programming framework, and c) analyzing the effect of HP clustering for gesture recognition.

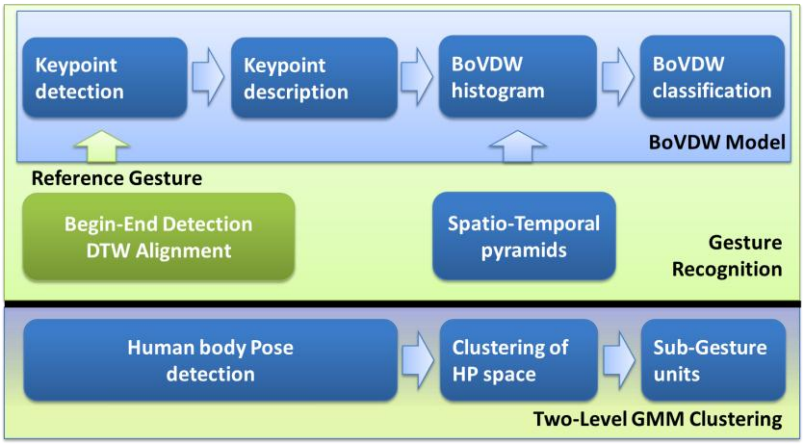

Figure 1: Pipeline of the full framework. BoVDW model at the top (blue), Probabilistic DTW at the middle (green), and clustering of HPs at the bottom (degraded blue and green).

1.1.1. Computing multi-modal feature spaces

First, data is acquired from different sensors that capture the information of people performing a large amount of gestures. This information is stored as multi-modal RGB-D (Red, Blue, Green, plus Depth) data along the time, and it is obtained by the Microsoft Kinect™ device shown in

Figure 2. Then, the post process consists of creating a feature space from the data, where different state-of-the-art multi-modal features are extracted and new ones are proposed, combining them using a fusion strategy. The computed feature space is then codified in a Bag-of-Visual-and-Depth-Words (BoVDW). Moreover, we compare the HGR performance obtained with our BoVDW descriptor respect to other common used descriptors.

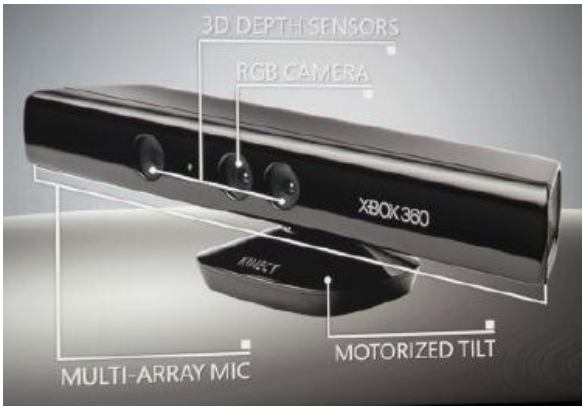

Figure 2: Microsoft Kinect multi-modal sensor device.

1.1.2. Probabilistic modelling of gestures

This module consists of integrating the presented BoVDW approach in a fully-automatic system for HGR, which uses a novel Probabilistic-Based Dynamic Time Warping (DTW) for the prior segmentation of gestures in a sequence. We compare the usual DTW algorithm with our Probability-Based DTW approach using a new proposed distance measure, showing both overlapping and accuracy as the evaluation measurements for the HGR.

1.1.3. Clustering HPs for Sub-Gesture representation

On the other hand, we apply existing methods from the state-of-the-art for constructing a new feature space of HPs through the structural model of human body joints from depth. Then, we propose a linear combination of means using a two-level Gaussian Mixture Model (GMM) process to perform clustering the HPs space. This second technique allows obtaining both more intuitive modelling of features and the first Sub-Gesture units, following our proposal in [Ponce et al., 2011a]. These Sub-Gesture units are coherent HP sets which can be useful to train posterior general purpose HBA systems.

Table 1 summarizes the nomenclature used in the rest of the work.

Table 1: Nomenclature Table

Variable or Constant	Description
V	Cardinality of the visual vocabulary representing the visual words selected.
η	Second-moment 3x3 matrix of first order spatial and temporal derivatives.
$\|\cdot\|$	Determinant.
$\lambda_1, \lambda_2, \lambda_3$	Regions in the image with significant eigen-values.
H	Determinant and the trace of μ.
R	Relative importance constant factor.
$T_r(\cdot)$	Trace computation.
S_{RGB}	Set of interest points for RGB data.
S_D	Set of interest points for Depth data.
$\mathcal{P}$	Number of points in a cloud.
r	Number of stable regions found in a cloud of points.
$\tau^{(i)} of \rho^{(i)}$	Normal of the i-th point $\rho^{(i)}$.
x	Horizontal vector of the camera plane.
y	Vertical vector of the camera plane.
z	Vector between the centroid of the cloud and the camera center.
P_{xy}	x, y plane, orthogonal to the viewing axis z.
ϕ	Angle between the normal $n^{(i)}$ and the viewing axis.
ψ	Frequencies of the projected angle.
h	Histogram that counts the occurrences of each word.
u, v, p	Dimensions of the volume.
b_u, b_v, b_p	Bins along the u, v, and p dimensions of the volume.
h^{RGB}	Histogram corresponding to the vocabulariy for the RGB-based descriptor.
h^D	Histogram corresponding to the vocabulariy for the Depth-based descriptor.
F	Value of the RGB or Depth histogram model.
d^F	Distance of the complementary histogram intersection.
d^{RGB}	Distance of RGB histogram.
d^D	Distance of Depth histogram.
d_{hist}	Weighted sum of distances between histograms.
β	Constant relative important factor.
Q	Long time-serie.
C	Short time serie.
q_j	Vector element of time-serie Q.
c_i	Vector element of time-serie C.

Variable or Constant	Description
q_j	Vector element of time-serie Q.
c_i	Vector element of time-serie C.
n	Length of time-serie Q.
m	Length of time-serie C.
M	Alignment matrix with the costs between c_i and q_j at each (i, j) position.
t	Certain time in a sequence.
f	Feature vector of the image.
T	Length of the warping path.
W	Cost matrix with the Warping path.
w_i	Contiguous matrix elements that defines a mapping between C and Q.
d	Euclidean distance between the feature vectors of the sequences c_i and q_j.
N	Number of training sequences.
S_i	Sequences for a certain gesture category.
L_i	Length in frames of sequence S_i.
$s^i_{L_i}$	Frames of sequence S_i.
$\bar{S}$	Median length sequence.
$\bar{S}_i$	Warped sequences.
$\bar{s}^i_j$	Frames of warped sequence.
G	Component Gaussian Mixture Model.
α	Mixing value.
π, μ, Σ	Parameters of each Gaussian Model in the mixture.
θ	Parameterization of the Gaussian Mixture Model.
λ^i_j	Means of a G-Component Gaussian Mixture Model .
k	Current component Gaussian Mixture Model, or number of clusters.
$p(\cdot)$	Probability distribution function.
$P(t, \lambda)$	Posterior probability of t belonging to the whole Gaussian Mixture Model.
D	Soft-distance based measure of the probability.
D_M	Cummulative cost matrix distance.
HP	Feature vector defining the human pose.
hp	Element in the feature vector HP.
K	Number of components of the Gaussian Mixture Model.
Θ	Discrete Alphabet defining a gesture vocabulary.

The rest of the Thesis is organized as follows: Section 2 reviews state-of-the-art in the fields related to the proposal of the Thesis. Section 3 presents the Bag-of-Visual-Depth-Words approach for multi-modal feature extraction and spatio-temporal representation of people in video sequences. Section 4 describes the probabilistic-based DTW for gestures in a dynamic programming framework. Section 5 describes the Sub-Gesture representation using clustering of HPs. Section 6 validates the different proposed methodologies and shows the results obtained. Section 7 shows real applications of the proposed methodology, and Section 8 a list of publications related to the proposed approaches and applications. Finally, Section 9 includes discussion and conclusion, as well as future lines of research.

2. State-of-the-art

In this section we explain some background on methods that describe HPs, as well as the probabilistic scope for segmentation and recognition of Human Gestures.

2.1. Description of HPs

Nowadays, Bag-of-Visual-Words BoVW is one of the most used approaches in Computer Vision, commonly applied in image retrieval or image classification scenarios. This methodology is an evolution of *Bag-of-Words* (BoW) [Lewis, 1998], a method used in document analysis, where each document is represented using the apparition frequency of each word of a predefined dictionary. In the image domain, these words become visual elements of a certain visual vocabulary. First, each image is decomposed into a large set of patches, either using some type of spatial sampling (grids, sliding window, etc.) or detecting points with relevant properties (corners, salient regions, etc.). Each patch is then described, obtaining a numeric descriptor. A set of V representative visual words are selected by means of a clustering process over the descriptors, where V is the cardinality of the visual vocabulary. Once the visual vocabulary is defined, each image can be represented by a global histogram containing the frequencies of visual words. Finally, this histogram can be used as input for any classification technique (i.e. k−Nearest Neighbor or SVM) [Csurka et al., 2004; Mirza-Mohammadi et al., 2009]. Moreover, extensions of BoW from still images to image sequences have been recently proposed, defining Spatio-Temporal-Visual-Words (STVW) (i.e. in the context of human action recognition) [Niebles et al., 2008].

In addition to visual features computed from RGB data, a new form of representing images has recently emerged, by including the depth as a new source of information. In this sense, the Microsoft Kinect™ sensor released in late 2010 caused a frenetic expansion in the Computer Vision field. Kinect™ is a low cost sensor which is able to capture depth information of the scene, in addition to the RGB image acquired by its camera, providing what is named RGB-D images -RGB plus Depth-. This depth information has been particularly exploited for human body segmentation and tracking. Shotton [Shotton et al., 2011] presented one of the greatest advances in the extraction of the human body pose using RGB-D, which is provided as part of

the Kinect™ human recognition framework. Moreover, motivated by the information provided by depth maps, several 3-D descriptors have been recently developed [Bogdan et al., 2009], which are based on codifying the distribution of normal vectors among regions in the 3D space.

In [Hernández et al., 2011], we have presented the Bag-of-Visual-and-Depth-Words (BoVDW) approach, which is an extension of the BoVW approach that takes profit of multi-modal RGB-D images by combining information of both RGB images and depth maps. We also have proposed a new depth descriptor which takes into account the distribution of normal vectors respect the camera position, as well as the rotation respect the roll axis of the camera. In order to evaluate the presented approach, we have compared the performances achieved with state-of-the-art RGB and depth features separately, and combining them in a *late fusion* fashion. All experiments are run in the proposed framework using the public data set provided by the ChaLearn Gesture Challenge[1] in the context of HGR. Finally, the presented BoVDW approach has been integrated in a fully automatic system for HGR, which uses DTW for the prior segmentation of gestures in a sequence. Results of the proposed BoVDW method have shown better performance using late fusion in comparison to early fusion and standard BoVW model.

2.2. Probabilistic and Dynamic Programming scopes for segmentation and recognition of Human Gestures

Since HRG is considered as the main core of the HBA, there exists the need of handling with problems such as segmentation and recognition. In this way, Human Gestures can be understood as continuous sequences of data points -or temporal series- following certain trajectories, whose information is relevant for the posterior HBA.

There exist a wide number of methods based on dynamic programming algorithms for both alignment and clustering of temporal series [Zhou et al., 2010]. Moreover, one can consider techniques that come from Probabilistic Graphical Models such as Hidden Markov Models (HMM) or Conditional Random Fields (CRF), which has been commonly used [Yang et al., 2009; Rabiner, 1989; Stefan et al., 2008; Fang et al., 2007], especially for classification purposes. Nevertheless, one of the most common methods used for HGR is Dynamic Time Warping (DTW) [Reyes et al., 2011; Stefan et al., 2008], since it offers a temporal alignment between sequences of different lengths.

[1] http://gesture.chalearn.org/

However, the application of such methods to HGR in complex scenarios becomes a hard task due to the high variability of the environmental conditions of each domain. Some common problems are: the wide range of human pose configurations, influence of background, continuity of human movements, spontaneity of humans actions, speed, appearance of unexpected objects, illumination changes, partial occlusions, or different points of view, just to mention a few. These effects can cause dramatic changes in the description of a certain gesture, generating great intra-class variability. In this sense, since usual DTW is applied between a sequence and a single pattern, it fails to take into account such variability. Thus, our proposal in [Bautista et al., 2011] has been focused on the definition of an extension of DTW method to a probability-based framework in order to perform an alignment between a sequence and a set of N pattern samples from the same gesture category using a Gaussian Mixture Model (GMM) [Svensén and Bishop, 2005] to model the variance caused by environmental factors. Consequently, the distance metric used in the DTW framework has been redefined in order to provide a probability-based measure. Results on public and challenging computer vision data have shown better performance of the proposed probabilistic-based DTW in comparison to the classical approach.

At this point, we consider that gestures are usually described as sequence of features per sample. Few works have decomposed either actions/gestures into a set of units [Zhou et al., 2010; Alon et al., 2009], or their features in a multiple eigen-spaces [Tam G. Huynh, 2008]. However, the fact of adding temporality for performing a temporal clustering of gestures of different length based on a clustering cost by means of DTW, allows categorizing gestures in a more robust way. As we proposed in [Ponce et al., 2011a], extending the same idea for HGR is an important improvement for the definition of more precise Sub-Gesture units, which are deformed and aligned in time. Sequence of Sub-Gesture units on the time can be modelled using a temporal model, such as HMM, often used in the literature. From the definition of a gesture vocabulary -or Bag-of-Gestures (BoG) as we call in our proposal-, a discrete alphabet Θ is obtained. If we consider each symbol of the alphabet -a gesture unit- as a possible state of a gesture, one can train state-transition probabilities of a HMM with standard Baum-Welch or Viterbi algorithm. Then, the final inference consists of testing the proposed methodology on a large scale data set of gestures. The process is performed by quantifying a gesture vocabulary and doing inference on trained temporal models of gestures.

Although HMM are the classical used temporal models, more powerful models that represent the hidden (and observed) state in terms of state variables which can have complex

interdependencies can be introduced. Such models regard to the Dynamic Bayesian Networks (DBNs) [Rabiner, 1989], Data Driven Markov Chain Monte Carlo (DD-MCMC) for Switching Linear Dynamical Systems (SLDS) [Oh et. Al., 2008], as well as Parametric and/or Segmental SLDS, where the lasts have been very useful to model natural trajectories (e.g. for the application tracking honey bees explained in such works).

2.3. Human Postures and Human Gestures

In [Elmezain et. al., 2009], authors define Human Gestures and Human Postures on Human Computer Interaction (HCI) problems, introducing an important differentiation between these two concepts. They refer specifically to hand posture and gesture recognition problems, assuring that static morphs of the hand are called postures and hand movements are called gestures. However, this definition can also cover the definition of Human Poses and Human Gestures, considering either separated body parts or all body parts jointly. In this way, a gesture is spatio-temporal pattern which may be static, dynamic or both. On the other hand, several challenging problems are presented for hand posture and gesture recognition in the context of HCI, which also cover the HP and HGR purposes. One of such problems, which arise in real-time hand gesture recognition, is to extract (spot) meaningful gestures from the continuous sequence of hand motions. Another problem is caused by the fact that the same gesture varies in shape, trajectory and duration, even for the same person. The goal of gesture interpretation is to push the advanced human-computer communication to bring the performance of HCI close to human-human interaction. They refer to Sign Language Recognition as an application area for HCI to communicate with computers and for sign language symbols detection. Sign language is categorized into three main groups namely finger spelling, word level sign and non-manual features [Bowden et al., 2003]. Finger spelling is used to convey the words letter by letter. The major communication is done through word level sign vocabulary and non-manual features include the facial expressions, mouth and body position.

The techniques for posture recognition with sign languages are reviewed for finger spelling to understand the research issues. The motivation behind this review is to develop a recognition system which works with high recognition rates. Practically, hand segmentation and computations of good features are important for the recognition. In the recognition of sign languages, different models are used to classify the alphabets and numbers. For example, in [Hussain, 1999], Adaptive Neuro-Fuzzy Inference Systems (ANFIS) model is used for the

recognition of Arabic Sign Language (ASL). In this proposed technique, colored gloves are used to avoid the segmentation problem and it helps the system to obtain good features. In [Handouyahia et al., 1999], the authors presented a recognition system for the International Sign Language (ISL). They used Neural Network (NN) to train the alphabets. NN is used for the recognition purposes because it can easily learn and train from the features computed for the sign languages. Other approach includes the Elliptic Fourier Descriptor (EFD) used in [Malassiotis and Strintzis, 2008] for 3-D hand posture recognition. In their system, they have used orientation and silhouettes from the hand to recognize 3-D hand postures. Similarly, Licsar and Sziranyi [Licsar and Sziranyi, 2002] used Fourier coefficients to represent hand shape in their system which enables them to analyze hand gestures for the recognition. Freeman and Roth [Freeman and Roth, 1994] used orientation histogram for the classification of gesture symbols, but huge training data is used to solve the orientation problem and to avoid the misclassification between symbols.

In the last decade, several methods for advanced hand gesture interfaces have been proposed [Deyou, 2006; Elmezain et al., 2008a; Kim et al., 2007; Mitra and Acharya, 2007, Yang et al., 2007], but they differ from one another in their models. Some of these models are Neural Network [Deyou, 2006], Hidden Markov Models (HMMs) [Elmezain et al., 2008a; Elmezain et al., 2008b] and DTW [Takahashi et al., 1992]. In 1999, Lee et al. [Lee and Kim, 1999] proposed an ergodic model based on adaptive threshold to spot the start and the end points of input patterns, and also classify the meaningful gestures by combining all states from all trained gesture models using HMMs. Kang et al. [Kang et al., 2004] developed a method to spot and recognize the meaningful movements where this method concurrently separates unintentional movements from a given image sequences. Alon et al., [Alon et al., 2005] proposed a new gesture spotting and recognition algorithm using a pruning model that allows he system to evaluate a relatively small number of hypotheses compared to Continuous Dynamic Programming (CDP). Yang et al. [Yang et al., 2007] presented a method for recognition of whole-body key gestures in Human-Robot Interaction (HRI) by HMMs and garbage model for non-gesture patterns.

Mostly, previous approaches use the backward spotting technique that first detects the end point of gesture by comparing the probability of maximal gesture models and non-gesture model. Secondly, they track back to discover the start point of the gesture through its optimal path and then the segmented gesture is sent to the recognizer for recognition. So, there is an inevitable time delay between the meaningful gesture spotting and recognition where this time

delay is not well for on-line applications. Above of all, few researchers have addressed the problems on non-sign patterns -which include out-of-vocabulary signs, epenthesis, and other movements that do not correspond to signs- for sign language spotting because it is difficult to model non-sign patterns [Lee and Kim, 1999].

The different techniques presented for hand gesture recognition are immediately applicable to HBA systems because hands give a huge amount of information at different levels. In addition, although HRG is focused on this extremity of the body, it is proved that modeling gestures at different levels is a good choice even for the rest parts of the body. In the following sections we show different approaches for grouping different feature representations, as well as probabilistic and dynamic programming algorithms related to the approaches presented in this state-of-the-art section.

3. BoVDW Representation

In this section, we present the BoVDW approach for HGR, whose pipeline is shown in the blue part of Figure 1. Figure 3 contains a conceptual scheme of the approach. The steps of the procedure are described below. At the end of this section, we present the application of the BoVDW for HGR (green pipeline in Figure 1).

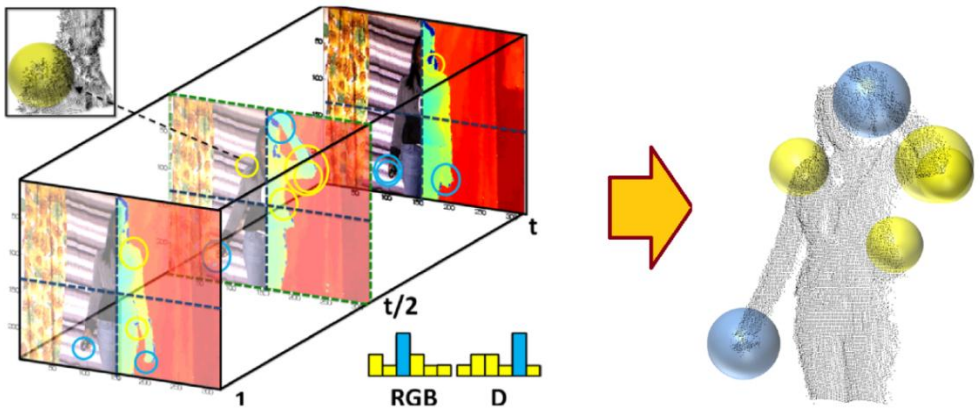

Figure 3: BoVDW approach in a HGR scenario. Interest points in RGB and depth images are depicted as circles. Color of the circles indicates the assignment to a visual word in the shown histogram –computed over one spatio-temporal bin–. Limits of the bins from the spatio-temporal pyramids decomposition are represented by dashed lines in blue and green, respectively. A detailed view of the normals of the depth image is shown in the upper-left corner.

3.1. Keypoint detection

The first step of BoW-based models consists of selecting a set of points in the image/video with relevant properties. In order to reduce the amount of points in a dense spatio-temporal sampling, we use the Spatio-Temporal Interest Point (STIP) detector [Laptev, 2005], which is an extension of the well-known Harris detector in the temporal dimension. The STIP detector first computes the second-moment matrix $\eta \in \mathbb{R}^{3\times3}$ of first order spatial and temporal derivatives. Then, the detector searches regions in the image with the three significant eigenvalues of η, combining the determinant and the trace of η:

$$H = |\eta| - R \cdot T_r(\eta)^3,$$

where $|\cdot|$ corresponds to the determinant, $T_r(\cdot)$ computes the trace, and R stands for a relative importance constant factor. As we have multi-modal RGB-D data, we apply the STIP detector separately on the RGB and Depth volumes, so we get two sets of interest points S_{RGB} and S_D.

3.2. Keypoint description

At this step, we want to describe the interest points detected in the previous step. On one hand, for S_{RGB} we compute state-of-the-art RGB descriptors, including HOG, HOF, and their concatenation HOG/HOF [Laptev et. al., 2008]. On the other hand, for S_D we test the VFH descriptor and propose the VFHCRH, detailed below. We do not use neither the whole image nor the whole object to describe the keypoints. Regions of interest are limited around each keypoint, depending on the scale where keypoints are detected.

3.2.1. VFHCRH

The recently proposed PFH and FPFH descriptors [Bogdan et al., 2009] represent each point in the 3-D cloud with a histogram, where each histogram encodes the distribution of the mean curvature around the described point. Therefore, both PFH and FPFH provide $\mathcal{P}$ histograms, invariants to $6DOF$ (Degrees of Freedom), being $\mathcal{P}$ the number of points in the cloud. Following their principles, VFH describes each cloud of points with only one descriptor of 308 bins, variant to object rotation around pitch and yaw axis. However, VFH is invariant to rotation about the roll axis of the camera. In contrast, CVFH describes each cloud of points using a different number of descriptors r, where r is the number of stable regions found on the cloud. Each stable region is described using a non-normalized VFH histogram and a Camera's Roll Histogram (CRH), and the final object description includes all region descriptors. CRH is computed by projecting the normal τ^i of the i-th point ρ^i onto a plane P_{xy} that is orthogonal to the viewing axis z (i.e. the vector between the centroid of the cloud and the camera center), under orthographic projection:

$$\tau_{xy}^i = \left|\left|\tau^i\right|\right| \cdot sin(\phi) \,,$$

where ϕ is the angle between the normal τ^i and the viewing axis. Finally, the histogram encodes the frequencies of the projected angle ψ between τ_{xy}^i and y-axis, the vertical vector of the camera plane.

In order to avoid descriptors of arbitrary lengths for different point clouds, we describe the whole cloud using VFH. In addition, a 92 bins CRH is computed for encoding $6DOF$ information. The concatenation of both histograms results in the proposed VFHCRH descriptor of 400 bins shown in Figure 4.

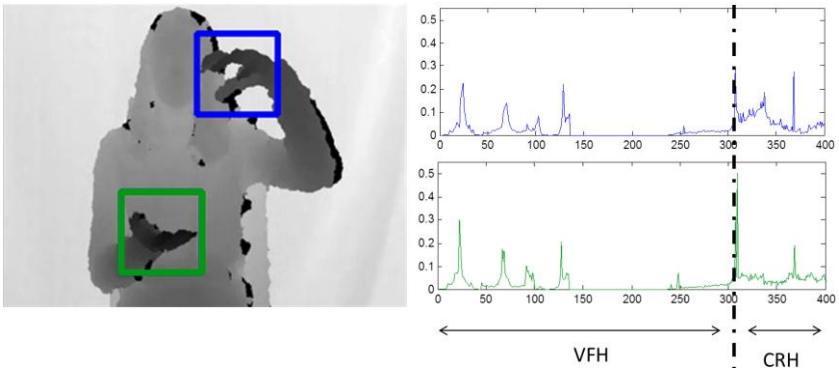

Figure 4: VFHCRH descriptor: Concatenation of VFH and CRH histograms resulting in 400 total bins.

3.3. BoVDW histogram

Once we have described all the detected points, we build our vocabulary of V visual/depth words by applying a clustering method over all the descriptors. Hence, the clustering method -- k-means in our case-- defines the words from which a query video will be represented, shaped like a histogram h that counts the occurrences of each word. Additionally, in order to introduce geometrical and temporal information, we apply spatio-temporal pyramids. Basically, spatio-temporal pyramids consist of dividing the video volume in b_u, b_v, and b_p bins along the u, v, and p dimensions of the volume, respectively. Then, $b_u \times b_v \times b\_p$ separate histograms are computed with the points lying in each one of these bins, and are concatenated with the global histogram computed using all points.

These histograms define the model for a certain class of the problem -in our case, a specific gesture-. Since we deal with multi-modal data, we build different vocabularies for the RGB-based descriptors and the depth-based ones, obtaining the corresponding histograms, h^{RGB} and h^{D}. Finally, the information given by the different modalities is merged in the next and final classification step, hence using late fusion.

3.4. BoVDW-based classification

The final step of the BoVDW approach consists of predicting the class of the query video. For that, any kind of multi-class supervised learning technique could be used. In our case, we use a simple k-Nearest Neighbour classification, computing the complementary of the histogram intersection as distance:

$$d^F = 1 - \sum_i min(h^F_{model}(i), h^F_{query}(i)),$$

where $F \in \{RGB, D\}$. Finally, in order to merge the histograms h^{RGB} and h^D, we compute the distances d^{RGB} and d^D separately, and compute a weighted sum:

$$d_{hist} = (1 - \beta)d^{RGB} + \beta d^D,$$

to perform late fusion, where β is a constant relative importance factor.

The validation of this approach is performed in the evaluation part (Section 6.2.1) together with the rest of methods.

4. Probabilistic-based DTW

In this section we first describe the original DTW and its common extension to detect start-end of sequences given a possible infinite sequence. Then, we extend the DTW in order to align patterns handling with the variance using a posterior probabilistic training with GMM. Finally, we explain the redefinition of the distance measure used in our approach.

4.1. Dynamic Time Warping

The original DTW algorithm was defined to match temporal distortions between two signals, finding a warping path between the two time series $Q = \{q_1, \dots, q_n\}$ and $C = \{c_1, \dots, c_m\}$. In order to align these two sequences, a $M_{m \times n}$ matrix is designed, where the entry (i, j) of the matrix contains the cost between c_i and q_j. Then, a warping path of length T –$W = \{w_1, .., w_T\}$ where w_i indexes a position in the cost matrix– is defined as a set of "contiguous" matrix elements that defines a mapping between C and Q. This warping path is typically subjected to several constraints:

- *Boundary conditions*: $w_1 = (1,1)$ and $w_T = (m, n)$.
- *Continuity and monotonicity*: Given $w_{t-1} = (a', b')$, then $w_t = (a, b)$, $a - a' \leq 1$ and $b - b' \leq 1$, this forces the points in W to be monotonically spaced in time.

We are generally interested in the final warping path that satisfying these conditions minimizes the warping cost:

$$DTW(Q, C) = \min \left\{ \frac{1}{T} \sqrt{\sum_{t=1}^{T} M(w_t)} \right\},$$

where T compensates the different lengths of the warping paths. This path can be found very efficiently using dynamic programming. The cumulative cost at a certain position $M(i, j)$ can be found as the composition of the distance $d(i, j)$ between the feature vectors of the sequences c_i and q_j with the minimum of the cumulative cost of the adjacent elements of the cost matrix up to that point. That is, $M(i, j) = d(i, j) + \min\{M(i-1, j-1), M(i-1, j), M(i, j-1)\}$.

Given the streaming nature of our problem, the input feature vector Q, has no defined length and may contain several instances of the gesture pattern C. In order to detect the beginning and

ending position of the candidate gesture, the current ending cost is checked -the cost of the element in the last row-. If that value is below a certain learned threshold μ, the warping path down to that matrix position is considered to define a matching warping candidate gesture. An example of a begin-end HGR together with the working path estimation is shown in Figure 5.

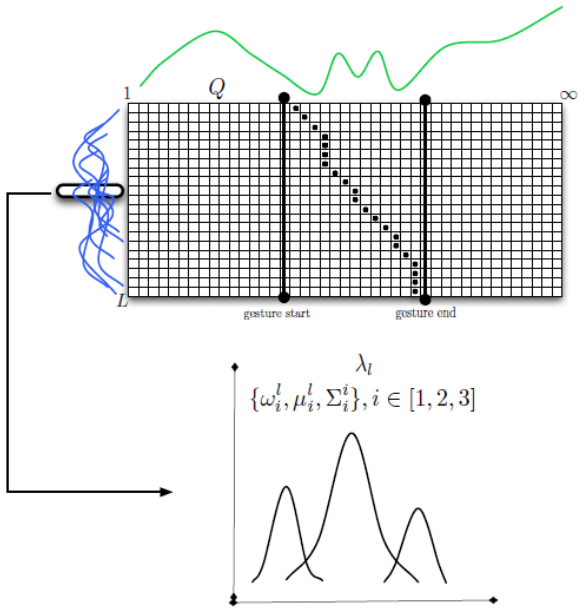

Figure 5: Begin-end of HGR of a gesture pattern in an infinite sequence Q using the probabilistic-based DTW. Note, how different samples of the same gesture category are modelled with a GMM and this model is used to provide a probability-based distance. In this sense, each cell of M will contain the accumulative distance.

4.2. Handling variance with Probability-based DTW

Consider a training set of N sequences $\{S_1, S_2, \dots, S_N\}$ with $S_i = \{s_1^i, \dots, s_{L_i}^i\}$ for a certain gesture category, where L_i is the length in frames of sequence S_i,. Let us assume that sequences are ordered according to their length, so that $L_{i-1} \leq L_i \leq L_{i+1}$ $\forall i \in [2, \dots, N-1]$, the median length sequence is $\bar{S} = S_{\lfloor \frac{N}{2} \rfloor}$. This sequence is used as a reference, and the rest of sequences are aligned with it using DTW in order to avoid the temporal deformations of different samples from a same gesture category. Therefore, after the alignment process, all sequences have length $L_{\lfloor \frac{N}{2} \rfloor}$. The set of warped sequences is $\{\tilde{S}_1, \tilde{S}_2, \dots, \tilde{S}_N\}$. Once all samples are aligned, the set of the j-th elements among all warped sequences, $\tilde{s}_j$ is modelled by means of an G–component Gaussian Mixture Model (GMM) $\lambda_j = \{\alpha_k, \mu_k, \Sigma_k\}$ where $k = 1, \dots, G$, α is the mixing value and μ and Σ the parameters of each Gaussian model in the mixture.

4.3. Distance measure

In classical DTW, a pattern and a sequence are aligned using a distance metric, such as the Euclidean distance. Since our gesture pattern is modelled by means of probabilistic models, if we want to use the principles of DTW, the distance needs to be redefined. In this work, we consider a soft-distance based on the probability of a point belonging to each one of the G components in the GMM.

Let each one of the GMMs that model the gesture patterns be defined as follows:

$$p(\tilde{s}_j) = \sum_{k=1}^{G} \alpha_k \cdot e^{-\frac{1}{2}(f-\mu_k)^T \cdot \Sigma_k^{-1} \cdot (f-\mu_k)}$$

Then, note that for a given feature vector f of the image, the posterior probability of f belonging to each one of the G components of the GMM can be obtained. In addition, since $\sum_1^k \alpha_k = 1$, we can compute the probability of f belonging to the whole GMM as the following:

$$P(f, \lambda) = \sum_{k=1}^{G} \alpha_k \cdot P(f)_k$$

which is the sum of the weighted probability of each component. Nevertheless, an additional step is required since the standard DTW algorithm is conceived for distances instead of similarity measures. In this sense we use a soft-distance based measure of the probability, which is defined as:

$$D = exp^{-P(f,\lambda)}$$

In conclusion, by aligning the set of N gesture sample sequences and modelling each of the elements composing the resulting warped sequences with a GMM, the possible temporal deformations of the gesture category are taken into account, and thus, we obtain a methodology for gesture detection that is able to deal with multiple deformations in data. The algorithm that summarizes the use of the probabilistic-based DTW to detect start-end of gesture categories is shown in Table 1.

Table 1: Probabilistic-based DTW begin-end of HGR algorithm

Input: A gesture model $C = \{c_1, .., c_m\}$ with corresponding GMM models $\lambda = \{\lambda_1, .., \lambda_m\}$, its similarity threshold value μ, and the testing sequence $Q = \{q_1, .., q_\infty\}$. Cost matrix $M_{m \times \infty}$ is defined, where $N(w), w = (i, t)$ is the set of three upper-left neighbor locations of w in M.
Output: Working path W of the detected gesture, if any

// Initialization
for $i = 1 : m$ **do**
  for $j = 1 : \infty$ **do**
    $M(i, j) = \infty$
  end
end
for $j = 1 : \infty$ **do**
  $M(0, j) = 0$
end
for $t = 0 : \infty$ **do**
  for $i = 1 : m$ **do**
    $w = (i, t)$
    $M(w) = D_M(w, \lambda) + \min_{w' \in N(w)} M(w')$
  end
  if $M(m, t) < \mu$ **then**
    $W = \{\text{argmin}_{w' \in N(w)} M(w')\}$
    return
  end
end

The validation of this approach is performed in the evaluation part (Section 6.2.2) together with the rest of methods.

5. Sub-Gesture Representation

This section explains the description of HPs and their clustering process for representing them as Sub-Gestures, [Ponce et. al., 2011c].

5.1. HP description

This section describes the processing of depth data in order to perform the segmentation of the human body by obtaining the skeletal model, and then computing its feature vector. For the acquisition of depth maps we use the public API OpenNI software [OpenNI, 2010]. This middleware is able to provide sequences of images at rate of 30 frames per second. The depth images obtained are $340x280$ pixels resolution. These features are able to detect and track people to a maximum distance of six meters from multi-sensor device.

We use the method of [Shotton et al., 2011] to detect the human body and its skeletal model. This approach uses a huge set of human samples to infer pixel labels through Random Forest estimation, and skeletal model is defined as the centroid of mass of the different dense regions using Mean-Shift algorithm. Experimental results demonstrated that it is efficient and effective for reconstructing 3-D human body poses, even against partial occlusions, different points of view or no light conditions. The articulated human model is defined by the set of 15 reference points shown in Figure 6. This model has the advantage of being highly deformable, and thus, able to fit to complex human poses.

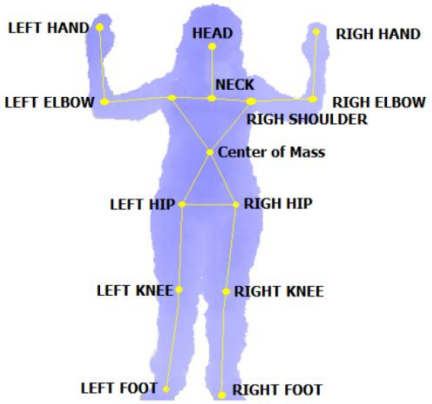

Figure 6: The 3-D articulated human model consisting of 15 distinctive points.

In order to subsequently make comparisons and analyze the different extracted skeletal models, we need to normalize them. In this sense, we use the neck joint of the skeletal model as the origin of coordinates (OC). Then, the neck is not used in the frame descriptor, and the remaining 14 joints are used in the frame descriptor computing their 3-D coordinates with respect to the OC. This transformation allows us to relate pose models that are at different depths, being invariant to translation, scale, and tolerant to corporal differences of subjects.

Therefore, the final feature vector HP_j at frame *j* that defines the human pose is described by 42 elements -*14* joints x 3 spatial coordinates-:

$$HP_j = \left\{ \left\{ hp_{\{j,x\}}^{\{1\}}, hp_{\{j,y\}}^{\{1\}}, hp_{\{j,z\}}^{\{1\}} \right\}, \ldots, \left\{ hp_{\{j,x\}}^{\{14\}}, hp_{\{j,y\}}^{\{14\}}, hp_{\{j,z\}}^{\{14\}} \right\} \right\}$$

5.2. HP clustering

In order to group the previous pose descriptions in pose clusters, we use standard Gaussian Mixture Model. Our goal is to group the set of frame pose descriptions in clusters so that posterior learning algorithms can improve generalization in HBA. We use a full covariance GMM of K components parameterized as follows,

$$\theta = \{\pi(k), \mu(k), \Sigma(k)\}, where\ k \in \{1, \ldots, K\}$$

Then, a likelihood value based on the probability distributions $p(\cdot)$ of the GMM is obtained as follows,

$$\mathrm{GMM}(HP, \mathrm{k}, \theta) = \sum_{\mathrm{i}} -\log \mathrm{p}(HP\ |\mathrm{k_i}, \theta) - \log \pi(\mathrm{k_i})$$

Based on this standard probabilistic GMM model, our two-level clustering procedure is defined as follows,

1) **First level**: Use three spatial components of descriptor HP for each joint $i, i \in [1,..,14]$ and perform GMM of k^1 clusters, namely GMM_1^i.
2) **Second level**:

2.1) Define for each pose a new feature vector,

$$HP^* = \left\{ hp_1^1, \ldots, hp_{k^1}^1, .., hp_1^{14}, .., hp_{k^1}^{14} \right\}$$

of size $14 \cdot k^1$, where hp_i^j is the probability result of applying GMM model GMM_1^i at features from HP corresponding to spatial coordinates of *j*-th joint.

2.2) Use the components of k^2 clusters, namely GMM_2.

Given a new frame, then, human skeleton is obtained as described before, and feature vector HP is computed and tested using two-level GMM description, obtaining a final probability for the most likely cluster from the set of k^2 possible pose clusters.

The validation of this approach is performed in the evaluation part (Section 6.2.3) together with the rest of methods.

6. Experiments and results

In this section, we discuss the data, methods and evaluation measurements before showing the results for the different presented approaches, as well as we discuss about them comparing among other state-of-the-art methods. We show performance of HGR for the BoVDW approach. Then, we show the performance for the probabilistic-based DTW. Finally, we show qualitative results for the HP representation.

6.1. Data

The first data source used is the ChaLearn data set [ChaLearn, 2011][2] provided from the CVPR2011 Workshop's challenge on HGR. The data set consists of 50,000 gestures each one portraying a single user in front of a fixed camera. The images are captured by the Kinect device providing both RGB and depth images. The data used -a subset of the whole- are 20 development batches with a manually tagged gesture segmentation. Each batch includes 100 recorded gestures grouped in sequences of 1 to 5 gestures performed by the same user. The gestures from each batch are drawn from a different vocabulary of 8 to 15 unique gestures and just one training sample per gesture is provided. For each sequence the actor performs a resting gesture between each gesture to classify. For this data set, we performed background subtraction based on depth maps, and defined a 10×10 grid approach to extract HOG+HOF feature descriptors per cell, which are finally concatenated in a full image (posture) descriptor. In this data set we will test for the recognition of the resting gesture pattern, using 100 samples of the pattern in a ten-fold validation procedure. An example of the ChaLearn dataset is shown in Figure 7 (a)-(b) and Figure 8.

[2] http://gesture.chalearn.org/data/data-examples

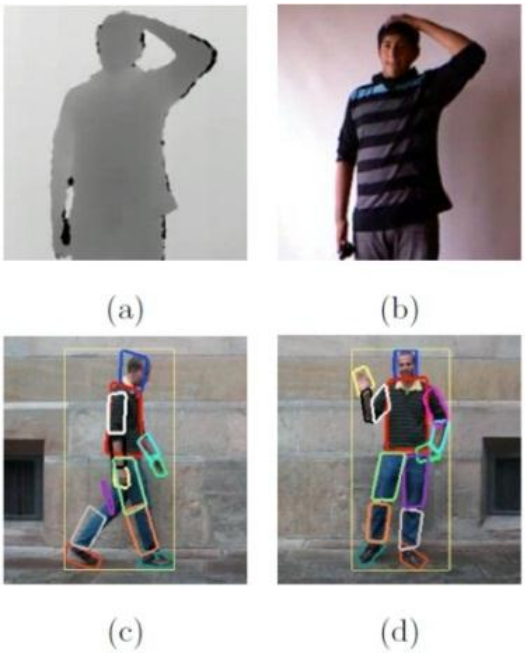

Figure 7: (a,b) Sample depth and RGB image for the ChaLearn database. (c,d) Sample frame of for different activities in the HUPBA dataset.

Figure 8: Gesture samples from the ChaLearn data.

Second, we use the HUPBA dataset, which is composed of 2 multi-actor sequences of RGB video recorded at 24 fps. For each one of those sequences 14 limbs per actor were manually tagged at each frame. In every sequence each actor performs a set of 12 actions, 8 individual actions and 4 actions which involve an additional actor. The features extracted from these sequences are the relative position of the 13 limbs to the centroid of the head. To effectively discriminate actions, the actors perform a resting gesture for a certain amount of time before performing a concrete action. Some images of this data set are shown in Figure 7 (c)-(d). In this data set we also aim to detect the resting gesture performed before each activity/gesture in order to provide a robust gesture segmentation procedure and compare performance with standard DTW approach.

Finally, we designed and used another new data set of gestures using the Microsoft Kinect™ device consisting of seven different categories. It has been considered 10 different actors and different environments, having a total of 130 data sequences with 32 frame gestures. Therefore, the data set contains the high variability from uncontrolled environments. The resolution of the video depth sequences is $340x280$. Some examples are shown in Figure 9.

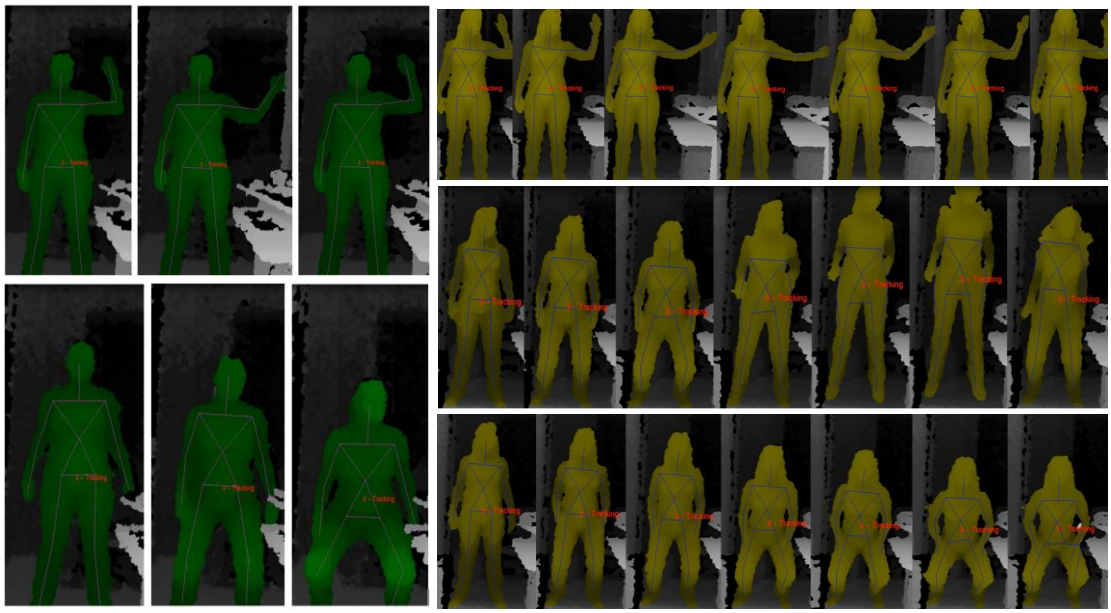

Figure 9: Examples of sequences provided by the people detection system showing the skeleton detection and tracking.

6.2. Methods and Evaluation measurements

Before the presentation of the results, first, we discuss the methods, parameters, and validation protocol of the experiments performed for the different presented approaches.

6.2.1. BoVDW

For the experiments, the vocabulary size was set to $V = 200$ words for both RGB and depth cases. For the spatio-temporal pyramids, the volume was divided in $2 \times 2 \times 2$ bins -resulting in a final histogram of 1800 bins-. In the classification step we use a simple Nearest Neighbor classifier, since we only have one training example for each gesture. Finally, for the late fusion, the weight $\beta = 0.8$ was empirically set. As a pre-processing step, DTW was applied to all sequences in order to segment the gestures.

For the evaluation of the methods, in the context of HGR, we have used the Levenshtein distance or edit distance. This edit distance between two strings is defined as the minimum number of operations -insertions, substitutions or deletions- needed to transform one string into the other. In our case, the strings contain the gesture labels detected in a video sequence. For all the comparison, we compute the Mean Levenshtein Distance (MLD) over all sequences and batches.

In order to test the BoVDW model representation, we designed a continuous HGR system. First, Probabilistic Dynamic Time Warping is used to detect a gesture of reference which splits the multiple gestures to be recognized. Then, each segmented gesture is classified using the BoVDW pipeline described above. These steps are also illustrated in the green pipeline shown in Figure 1.

6.2.2. Probabilistic-Based DTW

We compare the usual DTW algorithm with our probability-based DTW approach using the proposed distance D. The evaluation measurements are overlapping and accuracy of the recognition for the resting gesture, we consider that a gesture is correctly detected if overlapping in the resting gesture sub-sequence is greater than 60% -the standard overlapping value-. The cost-threshold for all experiments was obtained by cross-validation on training data. Each GMM in the probability-based DTW was fit with 4 components.

6.2.3. Sub-Gesture representation

For the HP representation, the people detection system used is provided by the public library OpenNI. This library has a high accuracy in people detection, allowing multiple detections even in cases of partial occlusions. The detection is accurate as people remain at a minimum of 60cm from the camera and up to 4m, but can reach up to 6m but with less robust and reliable detection. We perform classical one-level GMM and the proposed two-level GMM in Matlab programming, using $k^1 = 5$ and $k^2 = [10; 20; 40]$. k^2 is the number of clusters used for standard one-level GMM.

6.3. Results

Considering previous methods and evaluation measurements, in this section we present the results for the different approaches.

Table 2 shows the results for the usual DTW algorithm and our proposal on the ChaLearn and HUPBA datasets. We can see how the proposed probabilistic-based DTW approach outperforms the usual DTW algorithm in both experiments. In addition, this improvement is even higher when measuring the accuracy (up to a 7%).

Table 2: Small Overlapping and Accuracy results.

	ChaLearn		HUPBA	
Dataset	Overlap.	Acc.	Overlap.	Acc.
Probability-based DTW	**0.3908**	**0.6781**	**0.2534**	**0.4434**
Euclidean DTW	0.3003	0.6043	0.2314	0.4032

Table 3 shows a comparison between different state-of-the-art RGB and depth descriptors - including our proposed VFHCRH-, using our BoVDW approach. In the case of RGB descriptors, HOF alone performs the worst. In contrast, the early concatenation of HOF to HOG descriptor outperforms the simple HOG. Thus, HOF contributes adding discriminative information to HOG. In a similar way, looking at the depth descriptors, one can see how the concatenation of the CRH to the VFH descriptor clearly improves the performance compared to the simpler VFH.

Table 3: Mean Levenshtein distance for RGB and depth descriptors

RGB desc.	MLD	Depth desc.	MLD
HOG	0.3452	VFH	0.4021
HOF	0.4144	VFHCRH	**0.3064**
HOGHOF	**0.3314**		

Figure 10 shows the performance in all the 20 development batches separately. When using late fusion in order to merge information from the best RGB and depth descriptors (HOGHOF and VFHCRH, respectively), a MLD of 0.2714 is achieved. Furthermore, we also applied late fusion in a 3-fold way, merging HOG, HOF, and VFHCRH descriptors separately. In this case we assigned the weight α to HOG and VFHCRH descriptors (and $1-\alpha$ to HOF), improving the MLD to 0.2662. From this result we observe that HOGHOF late fusion performs better than HOGHOF early fusion.

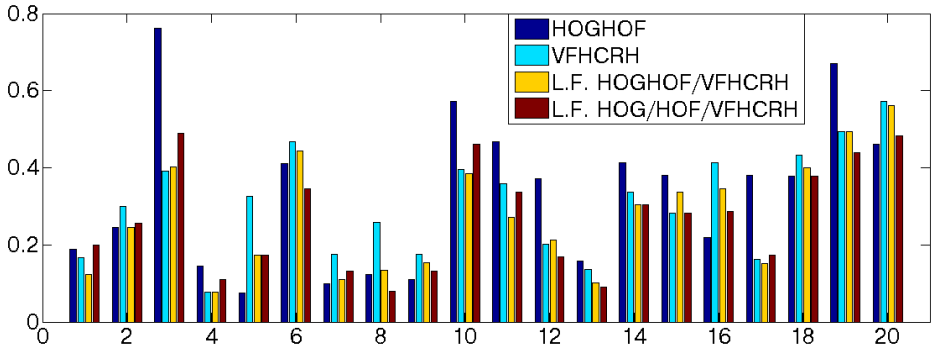

Figure 10: Performance of the best RGB and depth descriptors separately, as well as the 2-fold and 3-fold late fusion of them. X axis represent different batches and Y axis represents the MLD of each batch.

Finally, we tested the one and two-level GMM procedures on the designed data set. We show two qualitative results. First, in Figure 11, we show some examples of samples that fall in a particular cluster using one-level GMM and some samples that fall in a two-level GMM cluster. Up and down results are poses from different subjects. From these qualitative results, we can observe that in the case of one level cluster, samples have more visual variability, grouping different pose from different subjects in the same cluster. This is mainly because all joints and spatial coordinates are considered independent in the one-level GMM procedure, and large

movement in a particular joint affects global clustering for a particular pose. On the other hand, in the proposed two-level GMM clustering, all joints are first independently clustered, and grouped with equal probability in the second GMM level. As a result, we can observe that samples from the proposed clustering have more visual similarity, offering more discriminative information for better generalization of Human Behavior recognition techniques.

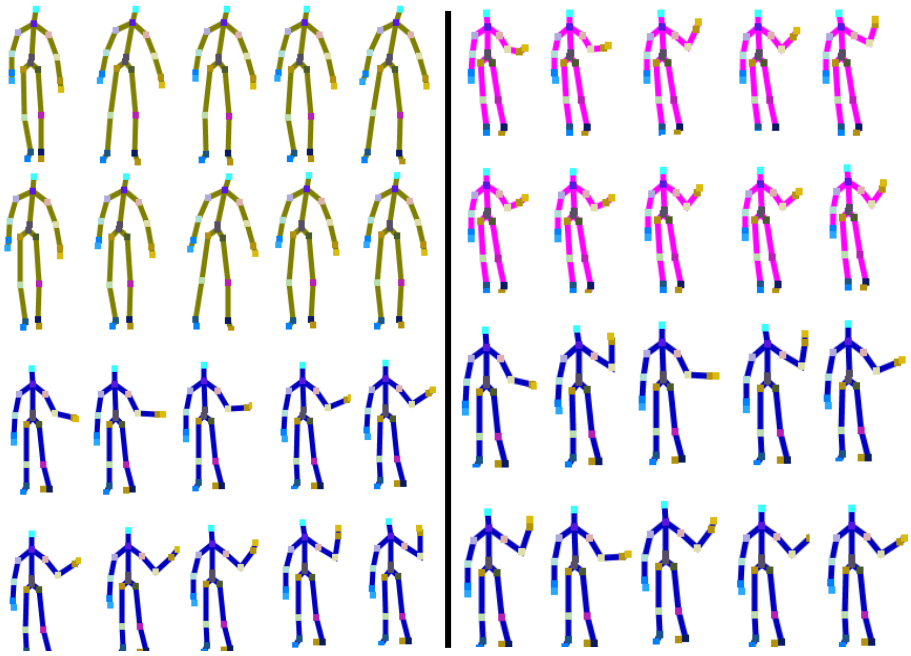

Figure 11: Sample examples from clusters defined using one-level GMM (left) and two-level GMM (right), respectively.

As an example of applicability, in the second qualitative result shown in Figure 12, we can see consecutive visual descriptions of some data set gestures. At the bottom of the sequences, we show a first row that represents the cluster number assigned by a one-level GMM, and a second row with the assigned cluster using two-level GMM. One can see that in most cases both grouping techniques assigns consecutive poses to same clusters, but as shown earlier, the clusters assigned by the one-level GMM have more visual variability, being inefficient for human behavior generalization purposes.

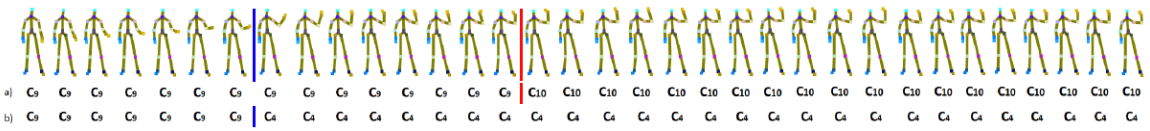

Figure 12: Cluster assignments for some gesture pose sequences. Red line shows the clustering of the first level, while blue one shows the clustering of the second level with less visual pose variability.

7. Applications

This section describes a set of real applications based on the proposed methodologies. In particular, we focus on monitoring and health care purposes, as well as social applications.

First, we explain the project funded by the Department of Applied Mathematics and Analysis (MAiA) in the University of Barcelona, as part of our presented work in the Doctoral Consortium of the International Joint Conference on Artificial Intelligence (IJCAI) 2011 [Ponce et al., 2011a], where we reviewed some applications coming from a B.S. Thesis and their immediately extension to other areas. Then, we briefly comment our current project funded by the Department of Justice of the Generalitat of Catalonia. Finally, we summarize another current project funded by the "Obra Social La Caixa".

7.1. Project funded by Dept. of MAiA in UB

As part of a previous B.S. Thesis funded by the Department of Applied Mathematics and Analysis (MAiA) of the University of Barcelona, we tackled the problem of HBA, where one can consider the non-verbal communication as a specific case. This analysis was performed in order to see the overlapping between the annotations provided by the feedback from the evaluation committee of the student's competences when presenting final bachelor thesis, and the impressions provided by our system. Then, we evaluate how the system discriminates the best grades from the worst ones.

As we explained in [Ponce et al., 2010, Ponce et al., 2011b], oral expression and non-verbal communication are one of the most relevant competences, and it is considered a critical factor to the personal, academic, professional and civic life of the graduates [Allen, 2002]. In this direction, Curtis and Winsor proved that oral communication was the second most important factor for the American Society of Personnel Administrators [Curtis et al., 1989], and subsequently conducted a survey of over 1000 human resources managers, concluding that good oral communication skills are important either for obtaining a job or to give a good performance at the job [Winsor et al., 1997].

In the particular case of Computer Science for oral expression and non-verbal communication, the development of this competence has usually been relegated to the defense of final bachelor projects. The list and methods for evaluating both the specific and transverse competences of a

final bachelor project has been analysed and widely discussed in the field of engineering, where such kind of activities are being developed for many years [Valderrama et al., 2009a; Valderrama et al., 2009b]. In many cases, the defense of the final bachelor project is the first opportunity for the student to be with the need of communicating their results orally, without prior training.

S. Indra Dexi and F. Shahnaz Feroz [Indra and Shahnaz, 2008] made a study of the fear effect to the grade earned by the students in oral presentations. What derives from their work is that fear leads to worse outcomes and that the more convinced the students are on their communication skills, the more comfortable they feel, and hence their grades become better. To improve the perception of students on their communication skills is necessary to generate activities that require communicating concepts or results, providing them a good feedback so they can keep improving their skills.

As we reviewed in previous sections, most automatic methods for HBA have a first stage of feature extraction and a second phase of analysis of the extracted data. In many works, the first phase is based on data mining by using special clothes or specific colors with sensors that allow to easily determine the position and acceleration of certain regions (hands, arms, head, etc. [Triesch et al., 1998]). In order to work in uncontrolled environments, other studies have focused on the skin color detection, movement, shapes, or background subtraction, which automates and give more independence to both the recognition system and the subject who performs the actions [Chen et al., 2003; Martin et al., 1998].

For the feature extraction stage of a HGR system, we defined a set of simple visual RGB features for our first works. These features were based on face detection, skin modelling, and feature tracking processes. We used the Face Detector of Viola & Jones [Viola and Jones, 2004] in order to detect the region of the face and define our origin or coordinates, which is similar to the approach presented in [Stefan et al., 2008]. Inner pixels of the detected region are used to train a skill color model [Jones and Rehg, 2002], which is used to look for hand/arm candidate regions. Finally, those blobs connected with the highest density correspond to our regions of interest, which are tracked using mean shift [Fukunaga and Hostetler, 1975]. All the spatial coordinates of the detected regions are computed in reference to the face coordinates and normalized using the face area. Examples of detected and tracked regions are shown in Figure 13.

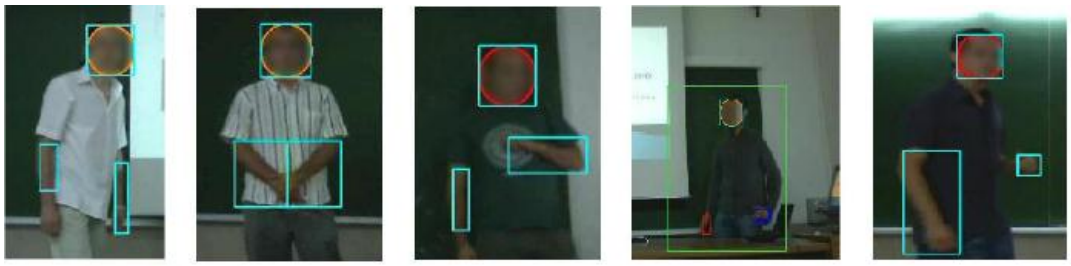

Figure 13: Examples of detected regions for different student presentations.

With the computed feature space, we performed an initial experiment where 15 bachelor thesis videos were recorded some examples are shown in Figure 13). Using the social signal indicators defined in [Pentland, 2005], we computed a set of activity, stress, and engagement indicators from the extracted feature space [Ponce et al., 2010]. Using the score obtained by the teachers at the presentations, we categorized the videos in two levels: those with the lowest score, and those with the highest score, and trained a Discrete Adaboost binary classifier [Friedman et al., 1998]. Applying stratified ten-fold cross-validation, we obtained interesting results, showing high prediction performance of student score based on his/her non-verbal communication using the extracted features. Then, we used Adaboost margin in order to rank features by relevance.

Moreover, we have also tested our system on different applications, such as Sign Language recognition using a novel multi-target dynamic gesture alignment, Attention Deficit Hyperactivity Disorder (ADHD), corporal physiotherapy analysis, and inpatient monitoring, with high success. See some examples in Figure 14.

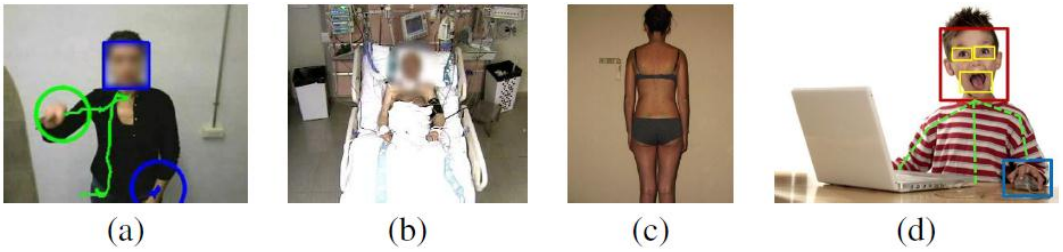

Figure 14: (a) Sign Language Recognition, (b) Impatient monitoring, (c) Physiotherapy analysis, and (d) ADHD analysis.

7.2. Project funded by the Department of Justice

At the present time, we are enrolled in a project funded by the Center of Juridical Studies and Specialized Formalization (CEJFE), as part of the Department of Justice of the *Generalitat de Catalunya*. In this project, we are interested in the analysis of the communication process in criminal mediation situations. In these situations, several people participate in the same session trying to reach an agreement of an event involving legal problems, helped by the figure of a mediator as part of the session participants. First, we are acquiring multi-modal data from these situations in different Cities of Justice from different localities: Barcelona, Marnesa, and Vilanova i la Geltrú. This is a long procedure that has to take in account all the required conditions for this purpose: etic, environmental conditions, distribution of people, required recording artifacts, and the invasiveness factor. Some examples of the acquiring process and the environment conditions are shown in Figure 15. Then, the second step is to handle with the data obtained for extracting human behavioral patterns of each of these situations applying the HRG techniques defined in this Thesis for HBA. Finally, using these patterns and the expert's feedback of the full mediation process, the main goal is to detect and analyze the situations when people opinions begin to match ("click" moment), as well as the reasons of why an agreement is being produced from a psychological point of view.

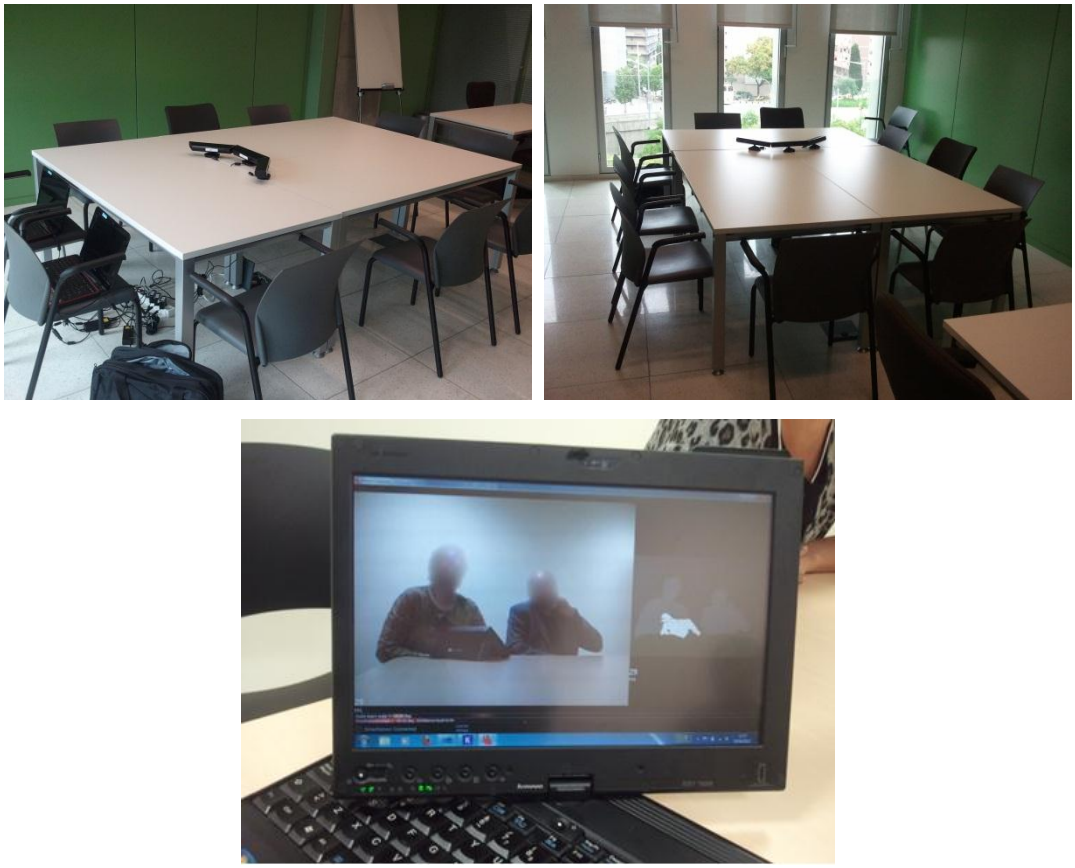

Figure 15: Top: Examples of the acquiring process and environment conditions from the City of Justice of Barcelona. Down: RGB (left) and depth maps (right) data acquisition from one of the three devices that records in the environment.

7.3. Project funded by the “Obra Social La Caixa”.

We are taking part of another present time project funded by the “Obra Social La Caixa” foundation. In this project, we are interested in monitoring and control the actions of ancients having several kinds of dementias like Alzheimer. As a particular case, we want to detect if a certain ancient has taken her pills in order to avoid performing the action twice. For that, we use a developed application which contains different methodologies of this Thesis. For instance, depth descriptors are used for detecting and tracking the person, as well as the interaction between that person and prior trained objects. Also, a qualitative DTW can be employed to correctly identify simple gestures like taking an object. Although the objects for our case are usually pillboxes, the application is able to learn a several kind of objects. Moreover, since the appearance of the person is also learned, the application is able to discriminate the correct interaction of the ancient with his/her pillbox. If a pillbox does not belong to a certain ancient, then an alarm can be triggered to alert that there could be a problem with the medication of the ancient. Moreover we have also included an extra functionality to the implemented system in order to perform ‘remembers’ of the location of different kind of objects in the environment, such as keys, glasses, or mobile phones. Figure 16 shows some examples of the Human-Object Interaction application and examples of the pillboxes used. Our plan is to use this technology either for geriatric centers or automatic home assistance.

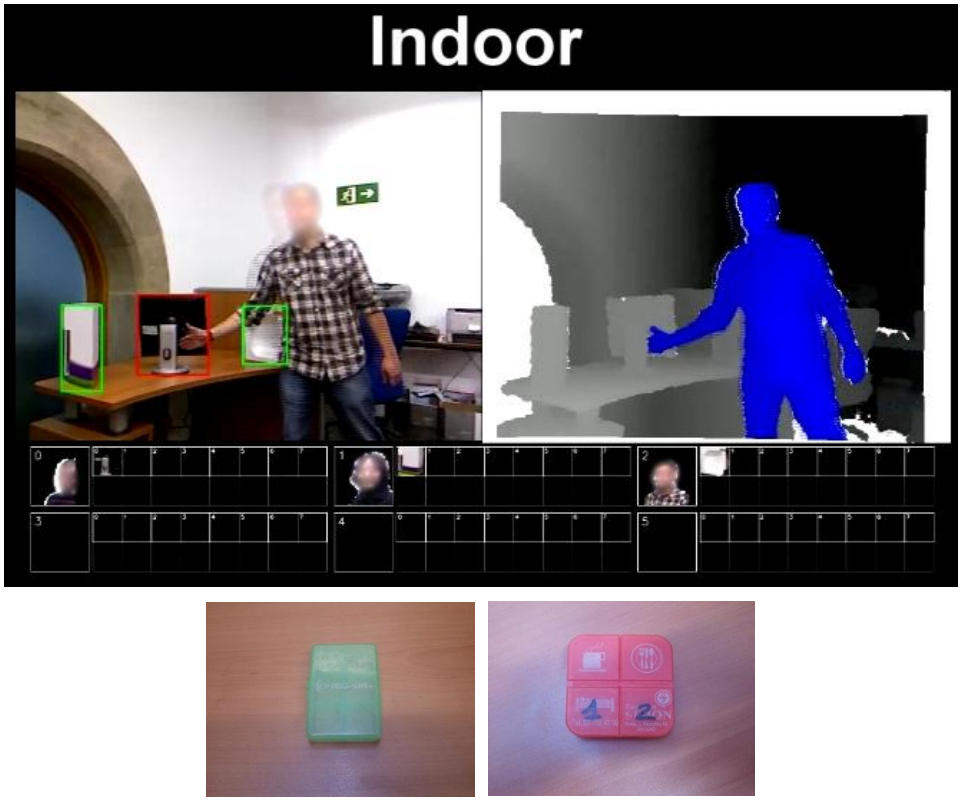

Figure 16: Examples of the Human-Object Interaction application and examples of the pillboxes used.

8. Publications

This section shows a list of publications of the author categorized according to the relation with the content of this M. Sc. Thesis. Some of the publications listed below can be also found in section 9 with the references.

[1] [Bautista et al., 2011] M. A. Bautista, A. Hernández, V. Ponce, X. Pérez, X. Baró, O. Pujol, C. Angulo, and S. Escalera. Probability-based Dynamic Time Warping for Gesture Recognition. Under revision for the International Workshop on Depth Image Analysis on ICPR, 2011.

[2] [Hernández et al., 2011] A. Hernández, M. A. Bautista, X. Pérez, V. Ponce, X. Baró, O. Pujol, C. Angulo, and S. Escalera. BoVDW: Bag-of-Visual-and-Depth-Words for Gesture Recognition. Accepted for Publication at ICPR, 2011.

[3] T. Hernández, Miguel Reyes, Víctor Ponce y Sergio Escalera. GrabCut-based Human Segmentation in Video Sequences. Under revision at the International Journal on Pattern Recognition and Artificial Intelligence (IJPRAI) 2011.

[4] [Ponce et al., 2010] V. Ponce, S. Escalera, X. Baro, and P. Radeva. Automatic analysis of non-verbal communication. CVCRD10 Achievements and New Opportunities in Computer Vision, pages 105–108, 2010.

[5] [Ponce et al., 2011a] V. Ponce, M. Gorga, X. Baró, S. Escalera, Human Behavior Analysis from Video Data Using Bag-of-Gestures. International Joint Conference on Artificial Intelligence, Doctoral Consortium, pp. 2836-2837, 2011.

[6] [Ponce et al., 2011b] V. Ponce, M. Gorga, X. Baró, P. Radeva, and S. Escalera. Análisis de la Expresión Oral y Gestual en Proyectos Fin de Carrera vía Un Sistema de Visión Artificial. ReVisión, Vol 4(1), 2011.

[7] [Ponce et. al., 2011c] V. Ponce, M. Reyes, X. Baró, M. Gorga and S. Escalera. Two-level GMM Clustering of Human Poses for Automatic Human Behavior Analysis. Proceedings of the Sixth CVC WorkShop CVCR&D State of the Art of Research and Development in Computer Vision, ISBN 978-84-938351-5-6, pp. 47-50, 2011.

Summarizing, in [2,3,4,5,6,7] we use different methodologies for describing features related with sections 3 and 5, as well as some new approaches for those purposes. In addition, [4,5,6] present similar applications related with the section 7.1 of this work. Finally, in [1,5] we discuss about different methodologies of the state of the art for HGR, and propose a new approach related to the one presented in section 4.

9. Conclusion and Future Work

In this M. Sc. Thesis we have presented both feature level and sequence level approaches for improving HGR using large data sets, as well as a first representation of clustered HPs in order to perform accurate HBAs considering gesture units.

First, we have presented the BoVDW approach for HGR using multi-modal RGB-D images. We have proposed a new depth descriptor VFHCRH, which outperforms VFH. Moreover, we have analyzed the effect of the late fusion for the combination of RGB and depth descriptors in the BoVDW, obtaining better performance in comparison to early fusion. Finally, we have presented a fully-automatic HGR system, using DTW for a prior segmentation of the video sequences, and the BoVDW approach for the classification of each segmented gesture.

On the other hand, we have proposed a probabilistic-based DTW for HGR, where different samples of the same gesture category are used to build a Gaussian-based probabilistic model of the gesture in which possible deformations are implicitly encoded. In addition, to embed these models into the DTW framework, soft-distance based on the posterior probability of the GMM was defined. In conclusion, a novel methodology for gesture detection that is able to deal with multiple deformations in data was presented.

Moreover, in this work we designed a data set of human actions and described individual frames using pose skeleton models from depth map information. We proposed a two-level GMM clustering algorithm in order to group similar poses so that posterior HBA techniques can improve generalization. We showed some preliminary qualitative results comparing our approach with the classical one-level GMM clustering strategy, showing a more visual coherent grouping of poses.

Finally, in this work we have presented several real and challenging applications using the proposed methodologies.

The future work consists of advancing in the different interconnected lines of research presented in this M. Sc. Thesis belonging to Machine Learning and Computer Vision fields, as part of the Artificial Intelligence scope. For the feature level, it could be interesting to compare with more descriptors and methods for emphasize the benefits of using the depth feature. Therefore, we can combine them on a new late fusion fashion, as we performed for our RGB-D descriptors. For the sequence level, we are interested in analyzing other alternatives for handing with the variance among gestures, as well as perform more theoretical analyses for the use of other distance measures. In addition, we think that it would be interesting to perform more experimentation by comparing with other methods from the Probabilistic Graphical Models. Since our approach for segmenting sequences is a priori used for posterior recognizing gestures, we believe on combining these approaches for improving HGR. Furthermore, we are interested in obtaining more precise Sub-Gesture units in order to obtain more accurate HGR systems. For that, we plan to propose, train, and validate different Probabilistic Graphical Model structures within a HBA framework. Then, the final inference consists of the testing of the proposed methodology on a large scale data set of gestures. The process will be performed by quantifying a gesture vocabulary in our discrete alphabet Θ (recalled from Section 2.2) and doing inference on trained temporal models of gestures.

10. References

[Allen, 2002] T. Allen, Charting a communicative pathway: Using assessment to guide curriculum development in a re-vitalized general education plan. Communicative Education, 51(1) 26-39. 2002.

[Alon et al., 2005] J. Alon, V. Athitsos, and S. Sclaroff. Accurate and Efficient Gesture Spotting via Pruning and Subgesture Reasoning. In Lecture Notes in Computer Sciences, Springer Berlin / Heidelberg, Vol. 3766, pp. 189-198, ISBN 978-3-540-29620-1, 2005.

[Alon et al., 2009] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff. A unified framework for gesture recognition and spatiotemporal gesture segmentation. PAMI, 31(9):1685–1699, 2009.

[Athitsos, 2010] V. Athitsos. Large lexicon project: Asl video corpus and sl indexing/retrieval algorithms. RPSL: Exploitation of Sign Language Corpora, 2010.

[Bautista et al., 2011] M. A. Bautista, A. Hernández, V. Ponce, X. Pérez, X. Baró, O. Pujol, C. Angulo, and S. Escalera. Probability-based Dynamic Time Warping for Gesture Recognition. Under revision for the International Workshop on Depth Image Analysis on ICPR, 2011.

[Bogdan et al., 2009] R. Bogdan, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *ICRA*, 2009.

[Bowden et al., 2003] R. Bowden, A. Zisserman, T. Kadir, and M. Brady. Vision Based Interpretation of Natural Sign Languages. In International Conference on Computer Vision Systems, 2003.

[Brown et al., 2005] L. M. Brown, A. W. Senior, Y. Tian, J. Connell, A. Hampapur, C. Shu, H. Merkl and Max Lu. Performance evaluation of surveillance systems under varying conditions. IEEE PETS Workshop, pp. 1-8, 2005.

[ChaLearn, 2011] Chalearn gesture dataset, California, 2011.

[Chen et al., 2003] Chen, F., Fu, C. y Huang, C.: Hand gesture recognition using a real-time tracking method and Hidden Markov Models. Image and Video Computing, vol. 21, No. 8, pp. 745—758, 2003.

[Csurka et al., 2004] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV*, pages 1–22, 2004.

[Curtis et al., 1989] D.B. Curtis, J. L. Winsor, and R.D. Stephens. National preferences in business and communication education. Communication Education, Vol. 38 (1), pp. 6-14. 1989.

[Deyou, 2006] X. Deyou. A Network Approach for Hand Gesture Recognition in Virtual Reality Driving Training System of SPG. In ICPR, pp. 519-522, 2006.

[Elmezain et. al., 2009] M. Elmezain, A. Al-Hamadi, O. Rashid, and B. Michaelis. Posture and Gesture Recognition for Human-Computer Interaction. Advanced Technologies, ISBN: 978-953-307-009-4, 2008.

[Elmezain et al., 2008a] M. Elmezain, A. Al-Hamadi, and B. Michaelis. Real-Time Capable System for Hand Gesture Recognition Using Hidden Markov Models in Stereo Color Image Sequences. In Journal of WSCG'08, Vol. 16, No. 1, pp. 65-72, ISSN 1213-6972, 2008.

[Elmezain et al., 2008b] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis. A Hidden Markov model-Based Continuous Gesture Recognition System for Hand Motion Trajectory. In International Conference on Pattern Recognition (ICPR), pp. 519-522, 2008.

[Fang et al., 2007] G. Fang, W. Gao, and D. Zhao. Large-vocabulary continuous sign language recognition based on transition-movement models. SMC-A, 37(1), 2007.

[Freeman and Roth, 1994] W. Freeman and M. Roth. Orientation histograms for hand gesture recognition. In International Workshop on Automatic Face and Gesture Recognition, pp. 296-301, 1994.

[Friedman et al., 1998] J. Friedman, T. Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. Annals of Statistics, 28:2000–2030, 1998.

[Fukunaga and Hostetler, 1975] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory, 21(1):32–40, 1975.

[Hampapur et al., 2005] A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, and S. Pankanti. Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking. SPM, IEEE, 22(2):38–51, 2005.

[Handouyahia et al., 1999] M. Handouyahia, D. Ziou, and S. Wang. Sign Language Recognition Using Moment-Based Size Functions. In International Conference of Vision Interface, pp. 210-216, 1999.

[Hernández et al., 2011] A. Hernández, M. A. Bautista, X. Pérez, V. Ponce, X. Baró, O. Pujol, C. Angulo, and S. Escalera. BoVDW: Bag-of-Visual-and-Depth-Words for Gesture Recognition. Accepted for Publication at ICPR, 2011.

[Howe and Dawson, 1996] Howe and Dawson, K. Active Surveillance. Using Dynamic Background Subtraction. Trinity College Dublin. Technical Report No. TCD-CS-96-06, 1996

[Hussain, 1999] M. Hussain. Automatic Recognition of Sign Language Gestures. Master Thesis, Jordan University of Science and Technology, 1999.

[Indra and Shahnaz, 2008] Indra Devi, S. y Shahnaz Feroz, F.: Oral Communication Apprehension and Communicative competence among Electrical Engineering undergraduates in UTeM. Journal of Human Development and Technology, Vol. 1, No. 1. 2008.

[Ivanov et al., 1999] Y. Ivanov, A. Bobick, Y. A. Ivanov, and A. F. Bobick. Recognition of Multi-agent Interaction in Video Surveillance. ICCV, pp. 169-176, 1999.

[Jones and Rehg, 2002] M. Jones and J. Rehg. Statistical color models with application to skin detection. IJCV, 46:81–96, 2002.

[Kang et al., 2004] H. Kang, C. Lee, and K. Jung. Recognition-based Gesture Spotting in Video Games. In Journal of Pattern Recognition Letters, Vol. 25, No. 15, pp. 1701-1714, ISSN 0167-8655, 2004.

[Kim et al., 2007] D. Kim, J. Song, and D. Kim. Simultaneous Gesture Segmentation and Recognition Based on Forward Spotting Accumulative HMMs. In Journal of the Pattern Recognition Society, Vol. 40, No. 11, pp. 3012-3026, ISSN 0031-3203, 2007.

[Laptev, 2005] I. Laptev. On space-time interest points. *Int. J. Comput.Vision*, 64(2-3):107–123, Sept. 2005. 107-123. Doi: 10.1007/s11263-005-1838-7. URL: http://dx.doi.org/10.1007/s11263-005-1838-7.

[Laptev et. al., 2008] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld.Learning realistic human actions from movies. *CVPR*, pages 1–8, 2008.

[Lee and Kim, 1999] H. Lee and J. Kim. An HMM-Based Threshold Model Approach for Gesture Recognition. IEEE Transaction on PAMI, Vol. 21, No. 10, pp. 961-973, ISSN 0162-8828, 1999.

[Lewis, 1998] D. Lewis. Naive (Bayes): The independence assumption in information retrieval. *ECML*, pages 4–15, 1998.

[Licsar and Sziranyi, 2002] A. Licsar and T. Sziranyi. Supervised Training Based Hand Gesture Recognition System. In International Conference on Pattern Recognition, 2002.

[Malassiotis and Strintzis, 2008] S. Malassiotis and M. Strintzis. Real-time Hand Posture Recognition using Range Data. In Image and Vision Computing, Vol. 26, No. 7, pp. 1027-1037, ISSN 0262-8856, 2008.

[Martin et al., 1998] Martin, J., Devin, V. y Crowley, J.: Active hand tracking, Proceedings of the III conference on Automatic Face and Gesture Recognition, pp. 573—578. 1998.

[Mirza-Mohammadi et al., 2009] M. Mirza-Mohammadi, S. Escalera, and P. Radeva. Contextual-guided bag-of-visual-words model for multiclass object categorization. *CAIP*, pages 748–756, 2009.

[Mitra and Acharya, 2007] S. Mitra and T. Acharya. Gesture Recognition: A Survey. In IEEE Transaction on Systems, MAN, and Cybernetics, Vol. 37, No. 3, pp. 311-324, ISSN 1094-6977, 2007.

[Niebles et al., 2008] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3):299–318, 2008.

[Oh et. Al., 2008] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert. Learning and Inferring Motion Patterns using Parametric Segmental Switching Linear Dynamic Systems. GVU Technical Report; Georgia Institute of Technology-GVU-06-02, 2006.

[OpenNI, 2010] Open natural interface. November 2010. Last viewed 14-07-2011 13:00.

[Pentland, 2005] A. Pentland. Socially aware computation and communication. Computer, 38:33–40, 2005.

[Ponce et al., 2010] V. Ponce, S. Escalera, X. Baro, and P. Radeva. Automatic analysis of non-verbal communication. CVCRD10 Achievements and New Opportunities in Computer Vision, pages 105–108, 2010.

[Ponce et al., 2011a] V. Ponce, M. Gorga, X. Baró, S. Escalera, Human Behavior Analysis from Video Data Using Bag-of-Gestures. International Joint Conference on Artificial Intelligence, Doctoral Consortium, pp. 2836-2837, 2011.

[Ponce et al., 2011b] V. Ponce, M. Gorga, X. Baró, P. Radeva, and S. Escalera. Análisis de la Expresión Oral y Gestual en Proyectos Fin de Carrera vía Un Sistema de Visión Artificial. ReVisión, Vol 4(1), 2011.

[Ponce et. al., 2011c] V. Ponce, M. Reyes, X. Baró, M. Go rga and S. Escalera. Two-level GMM Clustering of Human Poses for Automatic Human Behavior Analysis. Proceedings of the Sixth CVC WorkShop CVCR&D State of the Art of Research and Development in Computer Vision, ISBN 978-84-938351-5-6, pp. 47-50, 2011.

[Rabiner, 1989] L. R. Rabiner. A tutorial in Hidden Markov Models and selected applications in speech recognition. Proc. of the IEEE, 77(2):257–286, 1989.

[Reyes et al., 2011] M. Reyes, G. Dominguez, and S. Escalera. Feature weighting in dynamic time warping for gesture recognition in depth data. ICCV, 2011.

[Shotton et al., 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. *CVPR*, 2011.

[Stefan et al., 2008] A. Stefan, V. Athitsos, J. Alon, and S. Sclaroff. Translation and scale invariant gesture recognition in complex scenes. PETRA, 2008.

[Svensén and Bishop, 2005] M. Svensén and C. M. Bishop. Robust Bayesian mixture modelling. ESANN, 64:235–252, 2005.

[Takahashi et al., 1992] K. Takahashi, S. Sexi, and R. Oka. Spotting Recognition of Human Gestures From Motion Images. In Technical Report IE92-134, pp. 9-16, 1992.

[Tam G. Huynh, 2008] D. Tam G. Huynh. Human Activity Recognition with Wearable Sensors. Dissertation submitted to Technische Universität Darmstadt. Chapter 4, 2008.

[Triesch et al., 1998] Triesch, J. y von der Malsburg, C.: Robotic gesture recognition, Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction, pp. 233—244. Springer-Verlag, Londres, 1998.

[Valderrama et al., 2009a] Valderrama, E., Rullán, M., Sánchez, F., Pons, J., Cores, F. y Bisbal, J.: La evaluación de competencias en los Trabajos Fin de Estudios, Actas de las XV Jornadas de Enseñanza Universitaria de la Informática, Jenui ’09, pp. 405—412, Barcelona, 2009.

[Valderrama et al., 2009b] Valderrama, E., Rodríguez, S. y Prades, A.: Guía para la evaluación de competencias en los trabajos de fin de grado y de máster en las Ingenierías, AQU Catalunya. 2009.

[Viola and Jones, 2004] P. Viola and M. J. Jones. Robust real-time face detection. IJCV, 57(2):137–154, 2004.

[Winsor et al., 1997] J. L. Winsor, D.B. Curtis, and R.D. Stephens. National preferences in business and communication education. JACA, 3:170–179, 1997.

[Wren et al., 1997] C. Wren, A. Azarbayejani, T. Darrel and A. Pentland. Pfinder: Real-time tracking of the human body. IEEE TPAMI 19(7):780-785, 1997.

[Yang et al., 2007] H. Yang, A. Park, and S. Lee. Spotting and Recognition for Human-Robot Interaction. In IEEE Transaction on Robotics, Vol. 23, No. 2, pp. 256-270, ISSN 1552-3098, 2007.

[Yang et al., 2009] H.-D. Yang, S. Sclaroff, and S.-W. Lee. Sign language spotting with a threshold model based on conditional random fields. IEEE TPAMI, 31(7):1264–1277, 2009.

[Zhou et al., 2010] F. Zhou, F. D. la Torre, and J. K. Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. IEEE TPAMI, 2010.