A Comparison of Dynamic Naive Bayesian Classifiers and Hidden Markov Models for Gesture Recognition

H.H. Avilés-Arriaga*[1], L.E. Sucar-Succar[2], C.E. Mendoza-Durán[3], L.A. Pineda-Cortés[4]

[1,4]Department of Computer Science, Instituto de Investigaciones
en Matemáticas Aplicadas y en Sistemas,
Universidad Nacional Autónoma de México Circuito Escolar,
Ciudad Universitaria, 04510 Mexico City, Mexico
*haviles@live.com
[2] Computer Science Department, Instituto Nacional de Astrofísica,
Óptica y Electrónica, Luis Enrique Erro 1, 72840
Tonantzintla, Mexico
[3] Universidad Anáhuac (México Norte),
Av. Universidad Anáhuac, Núm. 46, Col. Lomas Anáhuac, 52786
Huixquilucan, Mexico

ABSTRACT
In this paper we present a study to assess the performance of dynamic naive Bayesian classifiers (DNBCs) versus standard hidden Markov models (HMMs) for gesture recognition. DNBCs incorporate explicit conditional independence among gesture features given states into HMMs. We show that this factorization offers competitive classification rates and error dispersion, it requires fewer parameters and it improves training time considerably in the presence of several attributes. We propose a set of qualitative and natural set of posture and motion attributes to describe gestures. We show that these posture-motion features increase recognition rates significantly in comparison to motion features. Additionally, an adaptive skin detection approach to cope with multiple users and different lighting conditions is proposed. We performed one of the most extensive experimentation presented in the literature to date that considers gestures of a single user, multiple people and with variations on distance and rotation using a gesture database with 9441 examples of 9 different classes performed by 15 people. Results show the effectiveness of the overall approach and the reliability of DNBCs in gesture recognition.

Keywords: Gesture recognition, hidden Markov models, motion analysis, visual tracking.

RESUMEN
En este documento se compara el desempeño de los clasificadores Bayesianos dinámicos simples (CBDSs) y los modelos ocultos de Markov (MOM) en el reconocimiento visual de ademanes. Los CBDSs extienden a los MOM incorporando suposiciones de independencia condicional entre los atributos dado el estado del modelo. Esta factorización ofrece porcentajes de clasificación y dispersión de error competitivos, un menor número de parámetros para el modelo y una mejora considerable del tiempo de entrenamiento. Para describir los gestos se propone un conjunto de atributos simples de postura y movimiento que incrementan el porcentaje de reconocimiento en comparación a modelos que sólo utilizan información de movimiento. Adicionalmente, se propone un esquema de detección de color de piel adaptativo para considerar diferentes usuarios y condiciones de iluminación. Se describe uno de los conjuntos de experimentos más exhaustivos presentados en la literatura de reconocimiento de gestos hasta el momento que incluyen gestos de un usuario, de diferentes personas, con variaciones de distancia y de rotación. Se presenta también una base de datos con 9441 ejemplos de 9 gestos de 15 personas. Los resultados muestran la efectividad de esta aproximación y la confiabilidad de los CBDSs en el reconocimiento de gestos.

1. Introduction

Hidden Markov models are successful and widely used classifiers in gesture recognition [1,2,3,4]. In the presence of several attributes, however, observation probability functions of HMMs imply conditional dependence among attributes given the state. This makes difficult to visualize independence relationships of attributes and their statistical behavior. Clarity in knowledge description is essential to a better understanding of gesture execution and recognition processes. Naive

Bayesian classifiers (NBCs) strongly relax the assumption of conditional dependence. This factorization improves clarity in the description of the attributes and decreases the number of parameters to be estimated. Moreover, NBCs are competitive to other more complex probabilistic and non-probabilistic classifiers, even when conditional independence does not hold [5,6]. However, in contrast with HMMs, NBCs do not naturally cope with sequential data. For these reasons, a model that combines the advantages of HMMs and NBCs in gesture recognition is desirable.

In this paper, we present an extensive empirical comparison between dynamic naive Bayesian classifiers [7] and HMMs for gesture recognition. DNBCs incorporate the concept of conditional independence among attributes given the state into standard HMMs to combine the descriptiveness capabilities of NBCs with the capacity of HMMs to model data streams. These models have received diverse names and used in different problem domains in the past. For example, multidimensional HMMs (MDHMMs) [8] to mix various sources of information in modeling teleoperation tasks for a robot manipulator; hybrid Naive Bayes HMMs (HNBHMMs) [9] to merge individual word information to classify multi-page documents; Output HMMs (OHMMs) [10] to combine expert's opinions in gene classification; and multi-observation HMMs (MOHMMs) [11] for fusioning behavioral patterns in the detection of abnormal actions in scenes. More recently, similar ideas have been applied successfully in activity recognition [12,13]. In all these works, factorization is proposed for mixing various sources of information. However, this is implicitly done in common HMMs applications in which, as stated above, each observation is defined by the conjunction of each feature value. By contrast, in our work we emphasize the importance of DNBCs to decrease the number of parameters of the model and improve training time, clarity in the representation of the attributes and to allow structural learning and feature selection [14,15]. Additionally, no methodical and systematic experimental evidence on the performance of this extension in comparison to standard HMMs in gesture recognition has been presented in the literature.

We propose to describe gestures in terms of a set of qualitative and fairly simple discrete motion and posture features. We show that posture and motion attributes increase recognition rates in comparison to models with motion features only, with gestures: a) taken from a single person, b) from multiple people, and c) with variations on distance and rotations. In addition, we describe a monocular visual system with a simple adaptive skin color strategy to cope with different users and lighting conditions. This visual system was used to construct a gesture database that comprises 9441 gestures samples of 9 gestures classes executed by 15 people used in our experimentation. This is one of the more extensive set of experiments to compare probabilistic graphical models in gesture recognition, with one of the highest number of gesture samples documented in the literature to date. Our results demonstrate the competitiveness of DNBCs in comparison to standard HMMs to learn, represent and classify gestures, and the effectiveness of the overall approach in the gesture recognition problems described above. Early results were presented in [16]. Here, we elaborate new experiments and results along with several improvements and corrections to our methodology. The main contributions of this paper are 1) DNBCs that provide competitive recognition results and efficient learning, 2) a set of simple and natural posture and motion features to effectively describe gestures, 3) improvement of posture-motion features to describe gestures, and 4) a more complete set of experiments than previous work presented in literature to date.

1.1 Outline

This document is organized as follows: Section 2 reviews various extensions to HMMs and different alternatives for the selection of gesture features. In section 3, we describe DNBCs. The adaptive strategy of our visual system is presented in section 4. Section 5 and 6 describe our gesture database and posture and motion features, respectively. Experiments and results for the validation of DNBCs and their comparison to HMMs, and a brief discussion, are described in Section 7. Finally, Section 8 summarizes our conclusions.

2. Related work

2.1 HMMs Classifiers for gesture recognition

HMMs describe statistical properties of dynamic gestures, with well-known probability estimation algorithms for learning and recognition [17] -See Fig. 1a for a Bayesian network description of these models; shaded nodes mean hidden variables. Several extensions to standard HMMs have been proposed to deal with particular issues in gesture recognition. Parametric HMMs (PHMMs) [18] represent gestures that involve spatial variations in their execution -*e.g.*, "This length" or "Go there". In PHMMs, observation variables are conditioned to the state variable and one or more parameters that account for such variations -Fig. 1b. Parameter values are known and constant on training. On testing, values that maximize the likelihood of the PHMM are recovered via a tailored EM algorithm. Coupled HMMs (CHMMs) [19] join HMMs by introducing conditional dependencies between state variables -see Fig. 1c. These models are suitable to represent influences between sub-processes that occur in parallel -e.g., two-hand gestures. Input-Output HMMs (IOHMMs) [20] consider an extra "input" parameter that affects the states of the Markov chain, and optionally, observation variables, -Fig. 1d. The input variable corresponds to the gesture observations. The output signal of IOHMMs is the class of the gesture that is being executed. A single IOHMM can describe a complete set of gesture classes. Parallel HMMs (PaHMMs) [21] require fewer HMMs than CHMMs for composite processes, by assuming mutual independence between HMMs -Fig. 1e. The idea is to construct independent HMMs for the possible motions of each hand and combine them by multiplying their individual likelihoods. PaHMMs with the most probable joint likelihood define the desired class. Hierarchical hidden Markov models (HHMMs) [22] arrange HMMs into layers at different levels of abstraction -Fig. 1f. In a two-layer HHMMs, the lower layer is a set of HMMs that represents sub-gesture sequences. The upper layer is a Markov chain that governs the dynamics of these sub-gestures. Layering allows re-using the basic HMMs simply by changing upper layers. Mixed-state dynamic Bayesian networks (MSDBNs) [23] combine discrete and continuous state spaces into a two-layer structure. MSDBNs are composed by a HMM in the upper layer and a linear dynamic system (LDS) in the lower layer. LDS is used to model transitions between real-valued states. Output values of the HMM drive the linear system -Fig. 1g. In MSDBNs, HMMs can describe discrete high-level concepts, such as a gesture grammar, while the LDS describes the motion of the hand in a continuous-state space. Hidden semi-Markov models (HSMMs) [24] exploit temporal knowledge of the process by defining explicit durations on each state -Fig. 1h. HSMMs are suitable to avoid an exponential decay of the state probabilities when modeling large observation sequences. More recently, derivations of HMMs that incorporate some of the characteristics presented above have been proposed as well [25,26]. Partially observable Markov decision processes (POMDPs) [27] generalize HMMs by including action and reward functions –Fig. 1i. The POMDP framework is usually used to quantify the "convenience" of the states of a system although its real situation is not completely known, and hence, to plan actions to reach a goal state. In [28] POMDPs are focused on actions to infer: i) the reaction to be taken in response to a gestural stimulus, ii) the cause that generates a gesture, or iii) decisions to maximize the return in a cooperative game between two players using gesture communication.

HMMs-based architectures have been successfully applied to challenging problems faced by novel applications of gesture recognition. In general, these approaches incorporate new variables to represent specific concepts into the standard HMMs framework, or factor the state space into various Markov chains to simplify its representation. Despite the usefulness of this framework, little attention has been paid to other important aspects of the problem, such as factorization to reduce the number of parameters of gesture features and to improve its description, and the evaluation of this extension in gesture recognition.

2.2 Gesture features

The selection of accurate and general gesture features is one of the most pursued goals in gesture recognition [29,30,31,32,33,34]. In practice, features are selected according to the characteristics of the gestures, and the application domain. Roughly speaking, alternatives to describe gestures can be divided in a) motion features, b) posture attributes, and c) posture-motion features. In the early 70s, Johannson [35] showed that isolated visible points over the joints of human actors in motion are enough to infer postures and activities. He named this visual phenomenon *biological motion*. After his findings, many authors have focused on features that emphasize motion

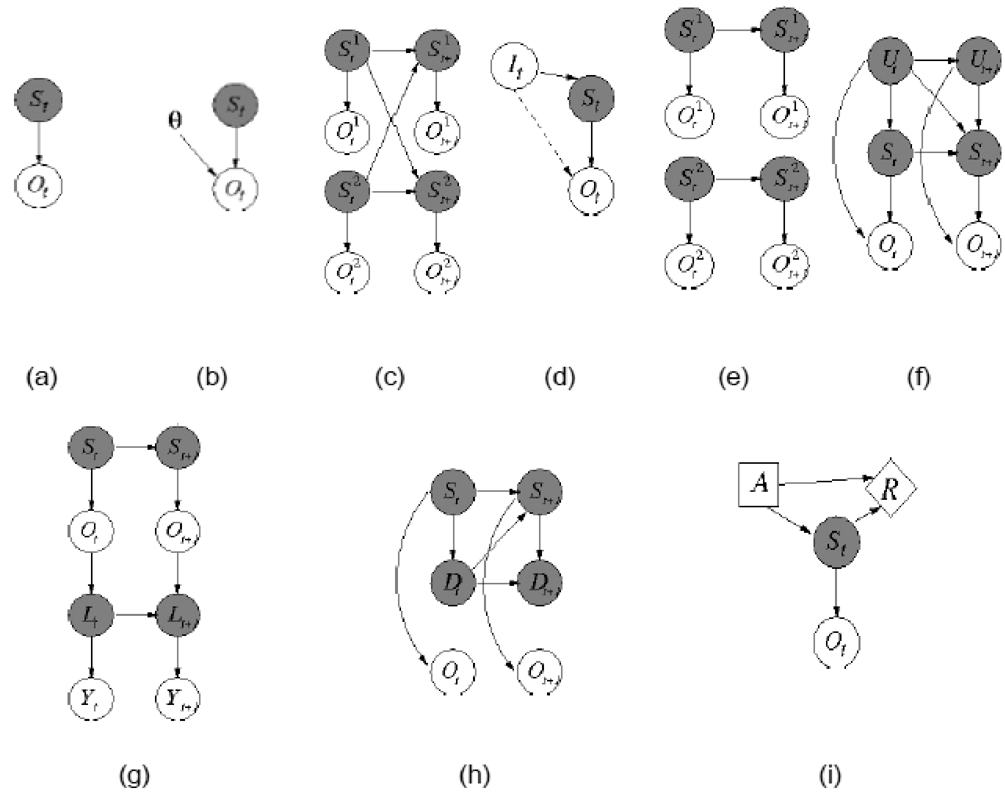


Figure 1. Bayesian networks representation of (a) standard HMMs with state and observation variables S_t and O_t, respectively, (b) PHMMs with a single parameter θ, (c) 2-Coupled HMMs, (d) IOHMMs with the input parameter I_t, (e) PaHMMs with two independent HMMs, (f) HHMMs with the Markov chain transitions in the upper layer denoted by U_t and U_{t+1}, (g) MSDBNs with a HMM in the upper layer, and a LDS in the lower layer indicated by L_t and L_{t+1}; in this case, Y_t and Y_{t+1} correspond to observations obtained from the process, (h) HSMMs with duration variables D_t and D_{t+1}, (i) POMDPs with an action function A and a reward function R. Shaded nodes indicate hidden states. Dashed arrows indicate optional dependencies. Models are unrolled two times only when required.

signals as the core information to describe gestures [36,37,38]. One representative example is temporal templates [39]. This technique -inspired in stroboscopic photography- collapses "motion appearance" into a single image, without regarding posture information to classify activities.

Another alternative is to see gestures as sequences of body postures [40] or global image coordinates -e.g., "raw" *(x,y)* data. In accordance to this scheme, some neurobiological experiments [41] have shown that motion information may be inferred from form *stimuli*, more than from form motion, as it was suggested by Johannson's work. These results have generated a live research field in feature selection from the neurobiological point of view [42,43,44].

Stokoe [45] suggests that gestures are characterized by motion, posture, orientation and position. In this form, some approaches have described gestures with a hybrid set of posture-motion features [46,47,48]. However, most of these proposals focus on the architectural design of the classifiers, without regarding on the discriminatory power of the features. This is usual in gesture recognition, where features are evaluated in conjuction with classifiers as a whole. Only a few tests for comparison of posture and motion features have been reported in literature. Campbell *et al*. [49] conducted experiments to test ten different feature sets based on posture -*e.g*., raw data, or polar coordinates- and motion information -*i.e*., Cartesian, polar velocities, instantaneous speed and local curvature using HMMs. Their results showed that velocity-based features obtained better recognition rates than posture data. Vogler and Metaxas [50] presented a comparison of ten feature sets of 2D and 3D posture and motion attributes for the classification of ASL. The attributes are similar to those used by [49] and include (*x,y,z*) data, polar and spherical coordinates of the hands, and its derivatives. However, in contrast to Campbell's work, their results showed that posture attributes slightly outperformed velocity attributes. Recently, coincidently in time to our evaluation on the combination of posture and motion features [16], Ahmad et al. [51] mixed 2D optical flow with a description of the human body shape based on *principal component analysis* for activity recognition. Their results were similar to our findings. These authors showed that the combination of posture and motion information improves recognition rates in activity recognition in comparison to models that consider posture or motion features only. However, it is difficult to draw strong conclusions from their results due to the small number of gesture classes and examples. Because of this, more extensive and conclusive experiments showing the importance of the combination of posture and motion data on different gesture recognition problems, and how these attributes can be represented, is still required. The approach presented in this document is a contribution to solve these problems.

3. Dynamic naive bayesian classifiers

In order to describe dynamic naive Bayesian classifiers, consider first a sequence $S=\{S_t \mid t=1,\ldots,T\}$ that is a realization of the states of the process, wher $1\leq S_t \leq N$ being N the number of possible states; and, a sequence $A=\{A_t \mid t=1,\ldots,T\}$ wh ere each $A_t=\{A_t^{(m)} \mid 1\leq m\leq M\}$ is a set of M attribute values generated by the process at state S_t. Superscripts m identify a specific attribute, in our case, an individual gesture feature.

Each attribute $A_t^{(m)}$ can be either discrete or continuous, although in this paper we consider the finite discrete case only; let $A_t^{(m)} \in \{k^{(m)} \mid 1\leq k^{(m)} \leq K^{(m)}\}$where $K^{(m)} \in$ be the possible values of each attribute m. A dynamic naive Bayesian classifier has the joint probability function:

$$P(A,S)=P(S_1)\prod_{t=1}^{T-1}P(S_{t+1}\mid S_t)\prod_{t=1}^{T}\prod_{m=1}^{M}P(A_t^{(m)}\mid S_t) \quad (1)$$

where $P(S_1)$ is the prior probability value of being at state S_1 at time $t=1$, $P(S_{t+1}\mid S_t)$ is the transition probability between classes S_t and S_{t+1}, and, $P(A_t^{(m)}\mid S_t)$ is the probability function of the observed feature m at time t given the class S_t. DNBCs follow two main assumptions: i) the first-order *Markov property*, and ii) the process is *stationary*.

A DNBC is denoted as $\lambda=\{P(S_1),P(S_{t+1}\mid S_t),P(A_t^{(m)}\mid S_t)\}$. The main difference between the DNBCs and HMMs probability functions [17] is the product $\prod_{m=1}^{M}P(A_t^{(m)}\mid S_t)$ that stands for the assumption of conditional independence among attributes given the class -HMMs implicitly assume a joint $P(A_t\mid S_t)$. If only a single attribute is considered or this attribute is product of the concatenation of several features, $M=1$ and (1) reduces to the joint probability function of a standard HMM. Figure 2 shows a DNBC unrolled two times with three attributes.

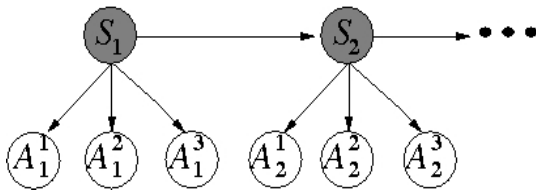


Figure 2. Graphical representation of a DNBC unrolled 2 times with 3 attributes.

3.1 Parameter learning

As usual in HMMs applications, the complete data pair (A,S) is not available and only A is accessible. Maximum likelihood estimation (ML) [17] is a common criterion for the selection of the parameters λ that best explain the observed and unseen data. This parameter learning process can be performed for DNBCs by means of the Baum-Welch algorithm [52] to iteratively improve $P(A)$ until no relevant difference in consecutive likelihoods of the model is found. Equations to compute new expectations λ' can be derived using the Baum's auxiliary function as described in [53]. Following this procedure, re-estimation formulas are

$$P(S_1 \mid \lambda') = \frac{P(A, S_1 \mid \lambda)}{P(A \mid \lambda)}, 1 \le S_1 \le N, \qquad (2)$$

for prior states' probabilities. Transition probabilities are calculated as

$$P(S_{t+1} \mid S_t, \lambda') = \frac{\sum_{t=1}^{T-1} P(A, S_{t+1}, S_t \mid \lambda)}{P(A, S_t \mid \lambda)}, 1 \le S_t, S_{t+1} \le N, \qquad (3)$$

and finally, for each attribute $A^{(m)}$:

$$P(A_t^{(m)} = k^{(m)} \mid S_t, \lambda') = \frac{\sum_{t=1}^{T} P(A, S_t \mid \lambda)\delta_{A_t^{(m)}, k^{(m)}}}{\sum_{t=1}^{T} P(A, S_t \mid \lambda)},$$

$$1 \le k^{(m)} \le K^{(m)}, 1 \le S_t \le N$$

(4)

where $\delta_{A_t^{(m)}, k^{(m)}} = 1 iff A_t^{(m)} = k^{(m)}$, and 0 otherwise. Given that parameters $P(S_{t+1} \mid S_t)$ and $P(k^{(m)} \mid S_t)$ do not depend on time t; hence $P(S_2 = j \mid S_1 = i) = P(S_{t+1} = j \mid S_t = i)$ for $t \in [2, T-1]$ and $\forall i, j$, and $P(k^{(m)} \mid S_t = i)$, for $t \in [2, T]$ and $\forall i$. The estimation of the previous distributions is based on the well-known variables forward $\alpha_{t,i}$, backward $\beta_{t,j}$, and $\xi_{t,i,j}$, the joint probability of moving from state i to state j at time . The computation of these variables must be modified to reflect the fact that $P(A_t, S_t \mid \lambda) = \prod_{m=1}^{M} P(A_t^{(m)} \mid S_t, \lambda)$. The forward variable is reformulated as

$$\begin{aligned} \alpha_{t,i} &= P(A_t, \ldots, S_t = i \mid \lambda) \\ &= [\sum_{j=1}^{N} \alpha_{t-1,j} P(S_t = i \mid S_{t-1} = j, \lambda)] P(A_t \mid S_t = i, \lambda) \\ &= [\sum_{j=1}^{N} \alpha_{t-1,j} P(S_t = i \mid S_{t-1} = j, \lambda)] \prod_{m=1}^{M} P(A_t^{(m)} \mid S_t = i, \lambda) \end{aligned}$$

(5)

The backward variable is computed as follows:

$$\begin{aligned} \beta_{t,j} &= P(A_t, A_{t+1}, A_{t+2}, \ldots, A_T \mid S_t = j, \lambda) \\ &= \sum_{i=1}^{N} P(S_{t+1} = i \mid S_t = j) P(A_{t+1} \mid S_{t+1} = i) \beta_{t+1,i} \\ &= \sum_{i=1}^{N} P(S_{t+1} = i \mid S_t = j) [\prod_{m=1}^{M} P(A_{t+1}^{(m)} \mid S_{t+1} = i) \beta_{t+1,i}], \end{aligned}$$

(6)

and finally, $\xi_{t,i,j}$ is

$$
\begin{aligned}
\xi_{t,i,j} &= P(S_{t+1}=j, S_t \mid A, \lambda) \\
&= \frac{\alpha_{t,i} P(S_{t+1}=j \mid S_t=i) P(A_{t+1} \mid S_{t+1}=j) \beta_{t+1,j}}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_{t,i} P(S_{t+1}=j \mid S_t=i) P(A_{t+1} \mid S_{t+1}=j) \beta_{t+1,j}} \\
&= \frac{\alpha_{t,i} P(S_{t+1}=j \mid S_t=i) \prod_{m=1}^{M} P(A_{t+1}^{(m)} \mid S_{t+1}=j) \beta_{t+1,j}}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_{t,i} P(S_{t+1}=j \mid S_t=i) \prod_{m=1}^{M} P(A_{t+1}^{(m)} \mid S_{t+1}=j) \beta_{t+1,j}}
\end{aligned}
\tag{7}
$$

Parameter adjustment must be done for each attribute independently; however, factorization does not affect considerably the number of operations to compute the intermediate parameters of the Baum-Welch algorithm. For example, the number of multiplications performed to compute forward variables is $N(N+1)(T-1)$ for standard HMMs; for DNBCs, this number increases only to $N(N+M)(T-1)$. Notwithstanding, as we will show below, attribute factorization reduces importantly the training time required by DNBCs in comparison to HMMs. Scaling and multiple observations sequences can be considered by the method proposed in [54].

3.2 Classification

Classification of a sequence of attribute observations A is as usual. Given a set of L DNBCs $\{\lambda_i \mid i=1,\ldots,L\}$, each of them trained with samples of a particular gesture class, compute $P(A \mid \lambda_i) = \sum_{j=1}^{N} P(A, S_j \mid \lambda_i) = \sum_{j=1}^{N} \alpha_{T,j}$ for each λ_i, by using the Forward algorithm. It is assumed that the λ_i with higher probability, *i.e.*, corresponds $argmax_{\lambda_i} P(A \mid \lambda_i)$, to the gesture class that has been executed.

4. Visual system

A monocular visual system based on an adaptive skin detection scheme to deal with different users was developed. The system is initiated with a person standing in a rest position, at a distance between 1.5m and 4m in front of the video camera. Face detection is performed using the face detector algorithm presented in [55]. The image subregions where it is expected to find the right-hand and torso of the person are estimated with body proportions based on face dimensions [56]. Figure 3 shows a result of this procedure.

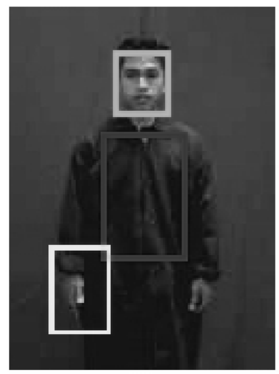


Figure 3. Estimation of the torso and hand positions of the person.

Hand segmentation and tracking proceeds as follows: We constructed a Bayes classifier accordingly to [57] to label pixel colors in the rgb space as skin or non-skin. We sampled 1,975,242 skin pixels taken from 30 people and 19,552,655 non-skin pixels under various lighting conditions to build general skin and non-skin probability functions, $P_g(rgb \mid skin)$ and $P_g(rgb \mid \neg skin)$, respectively. Thirty-two class intervals were defined for each rgb color channel. Once the hand is detected, a small skin-color search window is applied for its tracking. A direct likelihood comparison rule $P_g(rgb \mid skin) > P_g(rgb \mid \neg skin)$ was used to speed up the system. This approach worked well over four years in several demonstrations in our Lab [58]. A video that shows the application of this system for telecontrolling a

mobile robot can be found at *http://www.youtube.com/watch?v=opAUo0zJHGY*. However, a new camera and testing environment with intense white lighting and white walls caused the camera to perceive low-saturated colors that did not correspond to the probability distributions initially created. Because of this, the visual system was unable to locate the hand accurately, even with other color models such as *Hue-Saturation-Value.*
To deal with this problem and different users, an adaptive scheme was developed by combining the general $P_g(rgb\,|\cdot)$ skin and non-skin probability functions, with "personal" ones, $P_p(rgb\,|\cdot)$, created on-line by sampling randomly the face and torso of the user. These color functions are combined by the *independent likelihood pool* [59] rule defined as [16]

$$ILP(rgb\,|\cdot) = P_g(rgb\,|\cdot) * P_p(rgb\,|\cdot) \quad (8)$$

This way, one pixel is classified as skin *iff:*

$$ILP(rgb\,|\,skin) > ILP(rgb\,|\,\neg skin) \quad (9)$$

The CAMSHIFT algorithm [60] is used to track the hand motion over the rest of the image sequence. This strategy allows the visual system to track the hand effectively in our experimental conditions. An example of the tracking system is shown in Figure 4. A video showing this visual system is available at *http://www.youtube.com/watch?v=dFff01Tjvww.*

An intuitive explanation of the positive results with the ILP approach is that it weights general color distributions with precise information obtained from the images on-line. Other rules such as linear combination of probabilities: $ILP(rgb) = w_1 P_g(rgb\,|\cdot) + w_2 P_p(rgb\,|\cdot)$ did not generated the same results, probably because of the need to select accurate weights w_1 and w_2. However, a deeper analysis of the potential of this rule is beyond the scope of this document, and more experimentation is required to provide conclusive arguments on the application of this scheme for skin detection.

5. Gesture database

We propose 9 dynamic gestures oriented to interact with a mobile robot -see Fig. 5. Gestures were performed by 10 men and 5 women with the right arm at 3m in front of the video camera. To minimize the adverse effects of the visual processing errors over the feature extraction step, a blue-screen background was set and each participant was asked to use long-sleeved clothes of colors different from skin color. A short video was used to instruct people to perform each gesture class before starting its corresponding sampling round and no special recommendations were given afterwards. Except for one person named here man10, none of the other people had experience with the visual system or previous training executing gestures. The complete set of examples is composed of 7308 gestures. Every person contributed with a different number of samples; however, there are recorded at least 50 samples of each gesture per person.

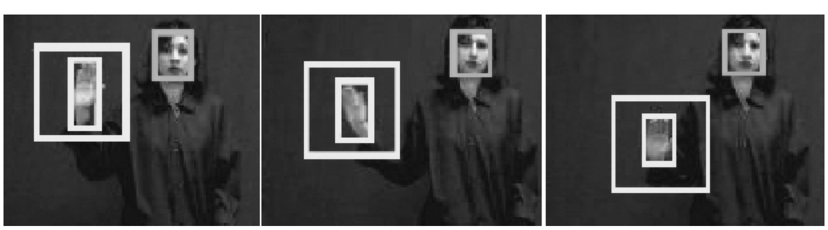


Figure 4. Example of the results of hand tracking through a sequence of 3 images.

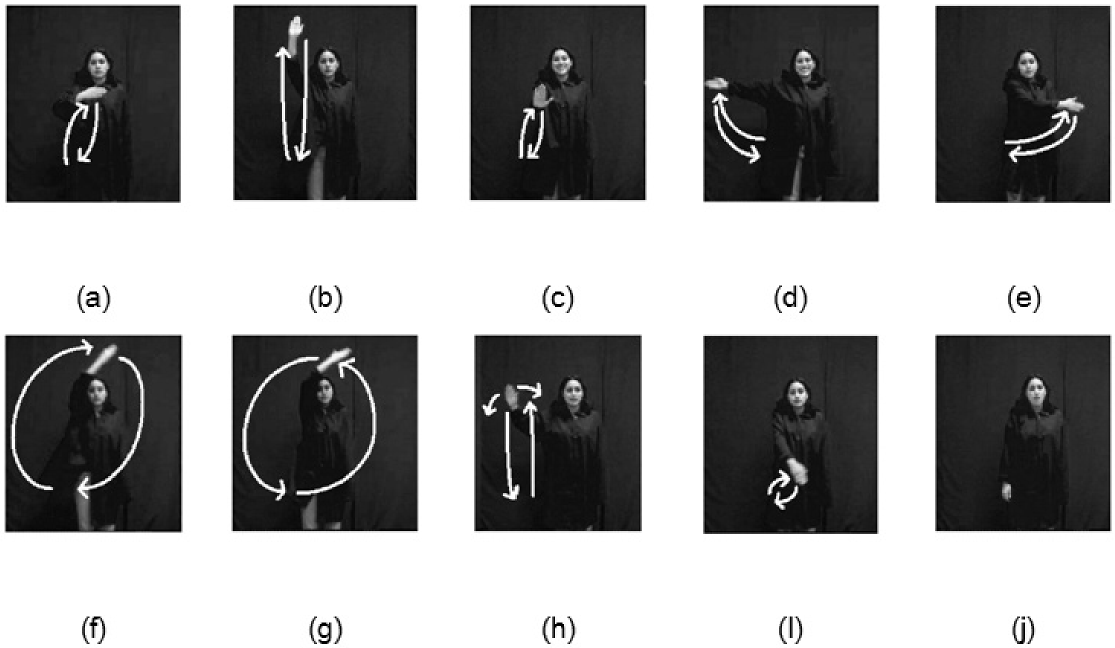


Figure 5. Gesture set: (a) come, (b) attention, (c) stop, (d) right, (e) left, (f) turn left, (g) turn right, (h) waving-hand and (i) pointing; (j) initial and final position for each gesture.

Each sample is composed by the length T of the observation sequence -that ranges from 6 to 42 observations- and the gesture data itself. Every observation is composed by i) *(x, y)*-coordinates of the upper and lower corners of the rectangle that segments the right hand, ii) *(x, y)*-coordinates of the upper and lower corners of the rectangle that segments the user's torso, and iii) *(x, y)*-coordinates of the center of the user's face. This coarse posture data enable us to easily transform the information to different feature sets. All coordinates are relative to the usual upper-left corner of the image. Data was recorded on plain text files. Spatial criterion about the position of the hand was used to start and end the capture of each gesture example. Observations were sampled every 4 images at a frame rate of 30 images per second approximately. This database can be downloaded from *http://sourceforge.net/projects/visualgestures/.* Additionally, two more sets of gestures were constructed. One person -labeled as man10- executed the 9 gestures at distances of 2m and 4m from the video camera. The same person performed again the 9 gestures with rotations of $\pm 45^{\circ}$ around the vertical axis at a distance of 3m. In the rotated sampling round we noted that the visual system worked well, although it was not originally designed for that purpose. The total number of gesture samples is 1081 for the database with distance variations, and 1052 for the database with rotation changes. Again, there are at least 50 samples per gesture at each distance and orientation.

6. Gesture attributes

From the coarse posture information described in Section 5, we extracted the following 7 gesture attributes: a) 3 features to describe motion, and b) 4 to describe posture. Motion features are $\Delta area$ -or changes in hand area-, Δx and Δy -or changes in hand position of the *XY*-plane of the image. The conjunction of these three attributes let us estimate hand motion in the Cartesian space *XYZ*. Each one of these features takes only one of three possible values: {*+, -, 0*} that indicate increment, decrement or no change, depending on the area and position of the hand in a previous image of the sequence.

For example, if the hand moves to the right, then Δx = +, if its motion is to the left, Δx = - and if there is no motion in the -axis, Δx = 0. An example on how these variables are instantiated accordingly to the user's hand motion is presented in Figure 6.

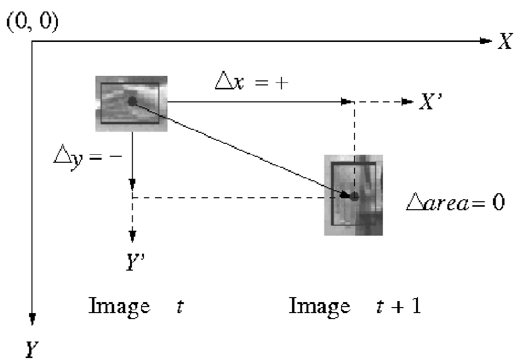


Figure 6. Figure 6: Example of motion features. In this image, the hand motion is performed to the right of the observer and downwards, so Δx = + and Δy = -; red points indicate the center of the hand. Given that the hand area does not change significantly, $\Delta area$ = 0.

Posture features named *form*, *right, above*, and *torso* describe hand orientation and spatial relations between the hand and other body parts, such as the face and torso. Hand orientation is represented by *form*. This feature is discretized into one of three values: + if the hand is vertical, - if the hand is horizontal, or 0 if the hand is leant to the left or right over the *XY* plane. *right* indicates if the hand is to the right of the head, *above* if the hand is above the head, and *torso* if the hand is in front of the torso. These three latter attributes take binary values, **true** or **false**, that represent if their corresponding condition is satisfied or not. An example of posture extraction in terms of these variables is depicted in Figure 7. This feature set does not make explicit use of magnitude components as usual on other approaches. The intention is rather to represent gestures through qualitative descriptions such as "The hand is moving to the user's right and upwards" or "The gesture is performed above the head".

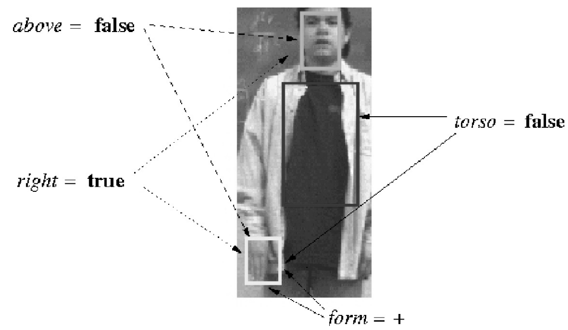


Figure 7. Example of the posture features. The image shows that the hand has a vertical position, below the head, to the right of the user and not over the user's torso, so the attribute values are *above* = **false**, *right* = **true**, *torso* = **false**, and *form* = +.

7. Experiments and results

We conducted three main experiments to compare classification and learning performances of DNBCs and HMMs. In the first experiment, gestures taken from the same person are used for recognition. In the second experiment, we evaluate the generalization capabilities of the classifiers by training and testing with gestures from different people. Experiment three considers gestures with variations on distance and rotation. First, we describe our experimental setup.

7.1 Experimental setup

Our visual system processes up to 30 f.p.s. The hardware is an IBM PC Intel Pentium 1.6 GHz, 512Mb RAM, a Sony EVI-D30 camera and a WinTV frame grabber. The image resolution is $640*480$ pixels. Sample code of the visual system is available at http://sourceforge.net/projects/visualgestures/.

All the experiments were carried out with DNBCs and HMMs with posture-motion on the one hand, and motion features only on the other. Figure 8 shows a graphical description of the 4 models. For DNBCs -Figs. 8a and 8b- instead of assuming statistical independence between Δx and Δy given the class variable, they were joined as a single attribute. Doing this, we obtained better classification results for these classifiers. All models

$$\text{recognition rate} = \frac{\text{Number of gestures correctly classified}}{\text{Number of testing samples}} * 100 \qquad (10)$$

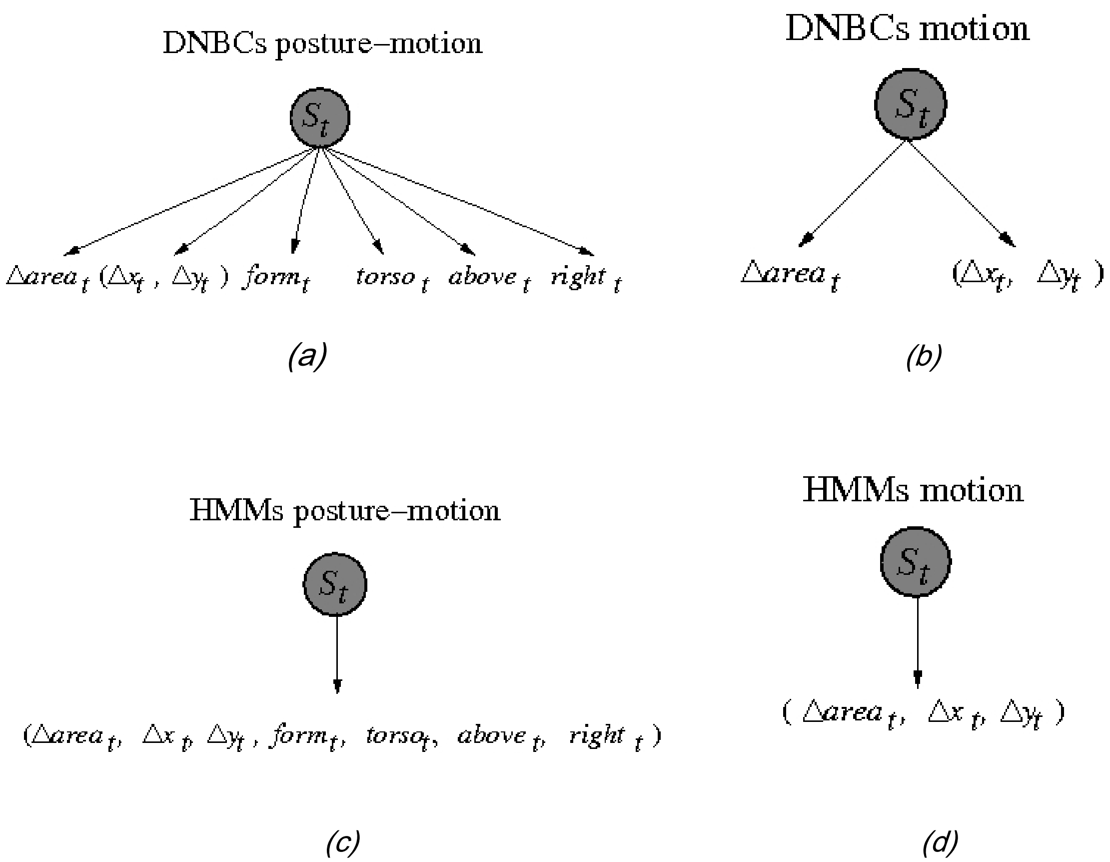


Figure 8. Graphical representation of DNBCs and HMMs considered in our experiments. (a) DNBCs with posture-motion features, (b) DNBCs with motion features, (c) HMMs with posture-motion information and (d) HMMs with motion attributes. Attributes within parenthesis conform a single joint probability distribution.

were set to follow standard "linear" transition topologies without skip transitions, initialized to an uniform discrete probability distribution.

The number of parameters to specify state observation distributions of HMMs with posture-motion features is 648 and with motion data only is 27. With DNBCs, parameters are 21 in the former case, and to 12 in the latter case. For training, stopping criterion is achieved if the absolute difference of $\log P(A \mid \cdot\,)$ of two consecutive models in an EM iteration is less than 1.0E-1. The whole gesture sequence was used without preprocessing. We use a modified version of the Tapas Kanungo's HMMs Toolkit [61] for training and testing HMMs and DNBCs. Recognition rate is calculated as follows:

7.2 Individual recognition

We use gesture samples performed at 3m in front of the video camera in the experiments presented in this section. For a single participant, 50 gestures of each class were selected randomly. From this pool, 20 gesture examples were chosen at random to construct a training data set. The remaining 30 samples compose the test data set. Training and testing examples are the same for all classifiers. The experiment was performed for each one of the 15 participants and repeated 10 times to average results. Table 1 shows the average error rate, total training time and the number of EM iterations of the four models as a function of the number of states of the model. As it is shown, the error rate tends to decrease as the number of states increases for all models. This indicates that commonly suggested topologies that range from 3 to 6 states [49,62,63,64] could not be adequate enough in all situations in gesture recognition. However, performance is not improved importantly beyond 12 states and slightly decreases with 18 states. Except for the experiment with a 3-state transition topology, DNBCs outperform recognition performances of HMMs. It also shows that DNBCs benefit training time significantly, without compromising recognition rates. In particular, training time of DNBCs with posture-motion data is consistently around one-tenth of the time required for HMMs. This difference is due to the number of possible observations of the models that is higher for HMMs. The number of iterations required for HMMs and DNBCs with and without posture does not vary considerably on each trial. This is because $\log P(A \mid \cdot\ )$ of these models are similar, as it is shown below. For the rest of the experiments, we selected a 12-state transition topology as a compromise between training time and recognition results.

It is useful to measure how erroneous responses are distributed among classes by the classifiers. We follow the method introduced by R. van Son to calculate error dispersion measures d_s and d_r from confusion matrices [65]. This method relies on entropy-base measure perplexity [66]. is the mean number of wrong responses per correct class; is the mean number of samples incorrectly classified on each possible class. These indices account for dispersion through the horizontal and vertical dimensions of the confusion matrix, respectively. The higher the dispersion is, the higher the value of these measures should be. To obtain these

	Posture-motion models						Motion models					
	Average error rate (%)		Total				Average error rate (%)		Total			
			Training time (*Sec*)		Number of iterations				Training time (*Sec*)		Number of iterations	
Number of states	DNBCs	HMMs	DNBCs	HMMs	DNBCs	HMMs	DNBCs	HMMs	DNBCs	HMMs	DNBCs	HMMs
3	3.81	3.02	36.77	322.82	13964	25007	28.43	28.71	28.26	44.28	13238	16548
6	2.37	2.6	126.7	1047.3	23305	28449	14.28	18.66	99.63	134.03	24719	25126
9	1.94	2.3	288.19	2344.5	29296	31676	13.28	16.73	217.35	303.96	32306	32707
12	1.78	2.18	516.63	4360.37	33270	34540	13.03	15.75	380.59	556.04	37183	38472
15	1.78	2.14	778.02	7805.28	34773	35940	13.42	15.84	599.79	868.54	41116	41258
18	1.72	2.2	1135.12	11854.8	36885	38696	13.82	16.09	897.48	1272.06	44718	44166

Table 1. Average training time, total training time and number of iterations for DNBCs and HMMs with and without posture data, as a function of the number of states in the transition topology.

measures, we calculated cumulative confusion matrices by pooling matrices of DNBCs and HMMs classifiers generated in the experiment with 12-states, for all the participants. Table 2 shows the values of these measures. For comparison purposes, consider a $9*9$ confusion matrix with a uniform distribution. For this matrix, error rate is 88%, and d_s = d_r = 8, i.e., 8 is the mean number of entries in which misclassifications are distributed. DNBCs with motion data provide lower error dispersion in comparison to HMMs with the same data. By contrast, DNBCs with posture and motion attributes generate slightly higher values than the corresponding HMMs. Notwithstanding, this latter difference does not seem to be significant in comparison to the dispersion values obtained from the uniform error distribution.

Figure 9a shows recognition rates for each person following this setup. Classifiers with posture and motion attributes improve recognition rates significantly in comparison to classifiers with motion attributes in all cases. Figures 9b and 9c depict average number of iterations and average training time to construct the classifiers, respectively. Figure 9d presents $\log P(A|\cdot)$ for each model.

In order to take a closer look on the performance of these models, an independent experiment was performed by varying the number of training samples for man10. In all trials, 30 examples selected at random are used for testing. Figure 10 shows the average recognition rate on 10 runs of the experiment as a function of the number of training examples. The only trial where HMMs clearly outperform DNBCs is the one training sample case with motion models, with a difference of 8.74%. However, this difference could not be meaningful at all, since HMMs rate is hardly above 50% and, it is somewhat unrealistic to expect reliable recognition rates with one or two training examples using motion only. Figure 11a shows the progression of $\log P(A|\cdot)$ of each classifier as a function of the number of EM iterations for the same gesture example. It is shown that DNBCs converge faster than HMMs. To evaluate how DNBCs reflect the evolution of a gesture, we computed the most probable states path of each model *via* the Viterbi algorithm -Figure 11b. It can be seen that paths are quite similar among all models and also observations spread uniformly over the 12 states in all cases.

7.3 Experiments with multiple people

It is common to construct and validate gesture models with samples taken from a single person. We agree with previous discussions that it is difficult to correctly recognize gestures from people not considered on training. However, in various applications, recognition must be performed with gestures from people not previously presented to the classifiers. Few systematic work has been done to test the behavior of the classifiers under this situation. To evaluate this, we use the classifiers constructed in the previous experiment for each person, to classify gestures from the remaining 14 people. For testing, we randomly extracted 2 samples per gesture from each personal database, excluding gestures from the person for whom the classifiers were constructed. In this form, a test set of 48 samples per gesture was generated.

	d_s	d_r
DNBCs motion-posture	1.69	1.69
HMMs motion-posture	1.53	1.54
DNBCs motion	3.06	3.09
HMMs motion	3.23	3.28

Table 2. Error rate and error dispersion indices and for DNBCs and HMMs with motion and posture-motion attributes.

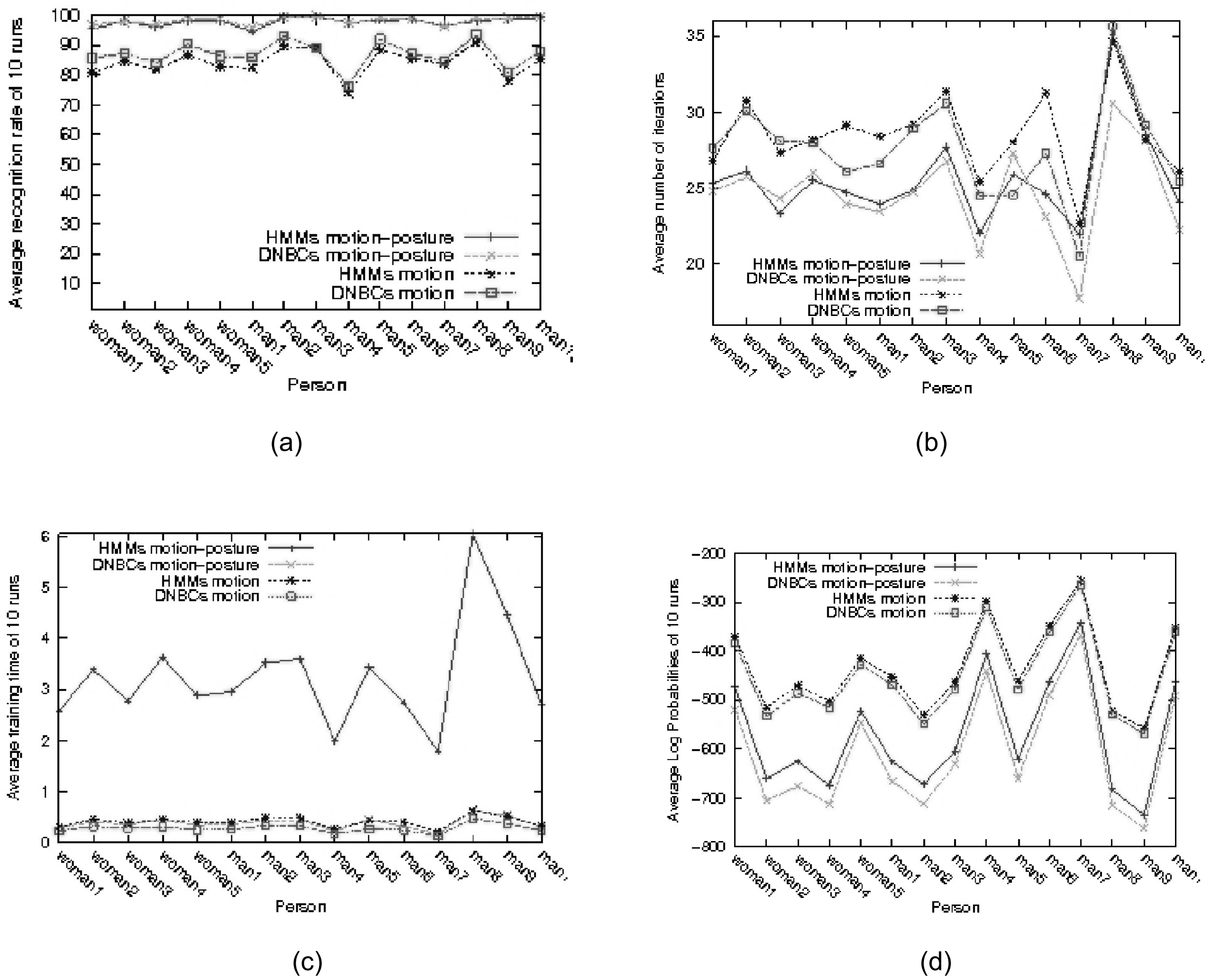


Figure 9. Results of the individual recognition of gestures using DNBCs and HMMs with posture-motion and motion attributes: (a) average recognition rates for each participant, (b) the average number of iterations for training, (c) average training time, and, (d) average log probabilities of observations given the model.

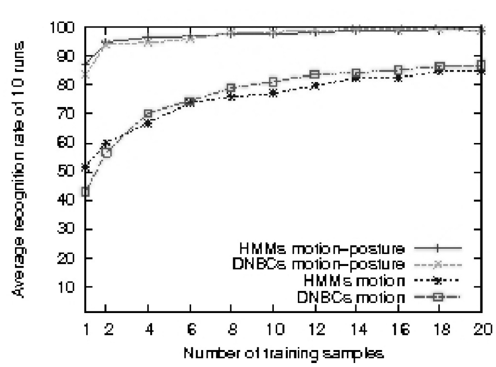


Figure 10. Average recognition rates of man10 as a function of the number of training examples.

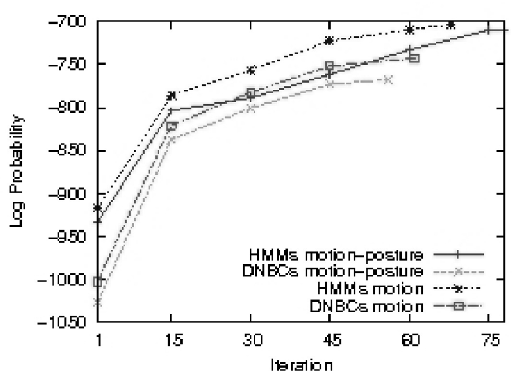
Figure 12a presents average recognition results of 10 repetitions of this experiment as a function of the personal classifiers used for testing. Average recognition rates for DNBCs with posture and motion features is 73.85%; for HMMs is 74.80%. Average recognition rates for DNBCs and HMMs with motion data is 52.80% and 51.60%, respectively.

Another experimental setting with models trained with 2 samples per gesture of 14 people and tested with 30 samples per gesture of the fifteenth person is depicted in Figure 12b. Percentages were obtained by averaging 10 instances of this experiment. Horizontal axis indicates the person to whom the testing set belongs to. Average recognition percentages for posture-motion models is 85.79% for DNBCs and 86.45% for HMMs. DNBCs with motion attributes obtained 67.73% and HMMs 64.18%. Recognition performance of the DNBCs and HMMs counterparts are closer in these two experiments, evincing the competitiveness of DNBCs for this problem.

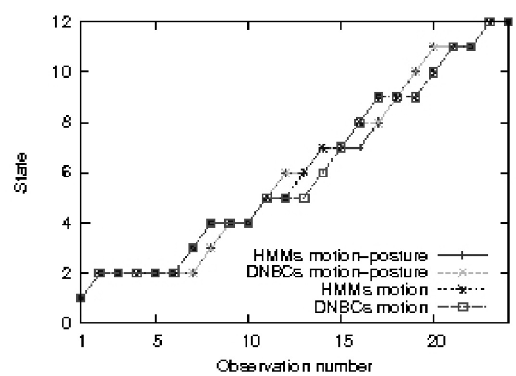
7.4 Variations on distance and rotation

In many applications, gestures are always executed at the same distance and orientation from the capture devices. In other application domains, - such as in human-robot interaction in which both the person and the robot can move- this restriction may not hold all the time. For the experiment on distance variation, 15 samples were randomly extracted for each gesture performed at 2m and 4m, giving a test set of 30 samples per gesture. The classifiers constructed in the first experiment for man10 were used. Figure 13a shows average recognition results of 10 runs of the experiment, as a function of the number of training samples. DNBCs provide competitive classification results in comparison to HMMs, with posture-motion and motion features.

The recognition of rotated gestures is a difficult problem in gesture recognition. The selection of accurate invariant features is one of the most evasive goals in this area. Although it is usually suggested that 3D information [49,67,68] or multiple views are necessary [51], we decided to evaluate the recognition performance of our models on this problem. The setting of this experiment is similar to the previous one. Fifteen samples of each gesture class executed at a $\pm 45^{\circ}$ were extracted at random to conform 30 testing samples per gesture. Again, we used the models constructed in the first experiment. Figure 13b shows these average recognition results of 10 runs of the experiment, as a function of the number of training samples. HMMs with posture-motion features outrange their DNBCs counterpart with an average difference of 4.61%. However, HMMs recognition rate is 72.55% in its best case, showing this is also a complex problem for HMMs. Motion models performed poorly in all cases.

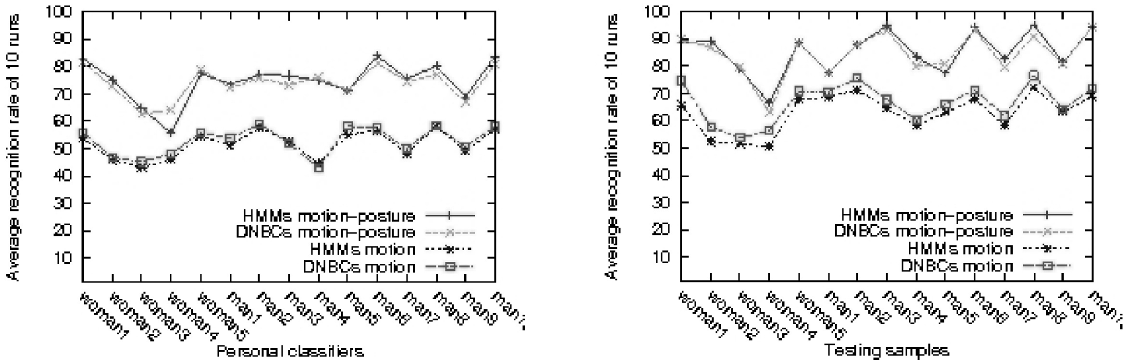


(a)



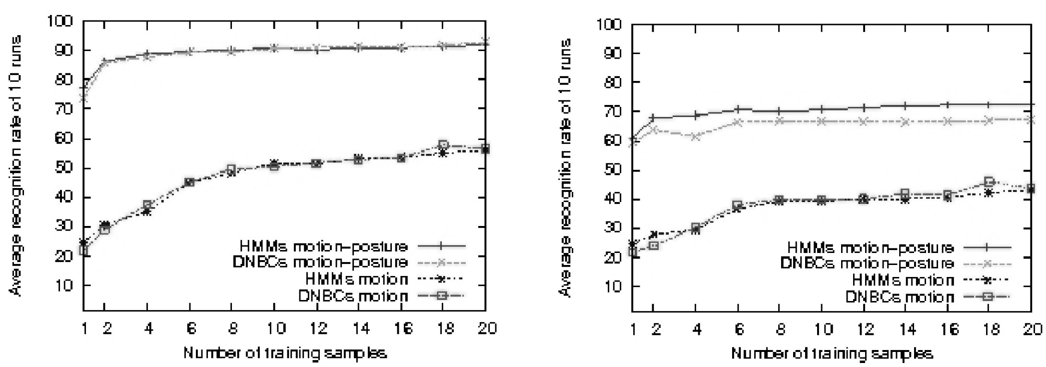
(b)

Figure 11. Examples of a single training and testing trial: a) convergence graph, and b) state transition through an observation sequence.



(a) (b)

Figure 12. Recognition rates of the experiments with multiple people: a) results with "personal" classifiers that are used to recognize gestures from the other 14 people, and b) with testing examples of each person to evaluate classifiers constructed with gestures from the other participants.



(a) (b)

Figure 13. Recognition results of gestures executed at (a) 2m and 4m, and (b) $\pm 45^{\circ}$.
The classifiers constructed in the first experiment for man10 were used.

7.5 Discussion

Results presented in the previous sections show the competitiveness in terms of recognition rates of DNBCs in comparison to standard HMMs in various issues in gesture recognition, using two sets of attributes. Attribute factorization allows an important decrease on training time for discrete motion and posture-motion models that benefits on-line learning of gestures. The recognition of rotated gestures with posture and motion information is the only experiment in which HMMs clearly outperform DNBCs. We believe this could be due to the large number of observation symbols required by HMMs that allows HMMs to handle such strong variations slightly better. We also show that the models with posture and motion data surpass the classifiers with motion features in all the experiments, in particular, when considering changes in distance and rotation.

Notwithstanding these positive results, we apply a single DNBCs structure to all our gestures, as usual in gesture recognition. However, besides classification, a complete gesture analysis requires also the development of models that effectively describe attributes and their statistical dependence relationships for each gesture class. We have shown that conditional independence assumptions decrease recognition performance only in complex situations that are difficult even for HMMs, and yet allow us to explore structural learning [69] and feature selection techniques [70]. For example, in [15], an evolutionary learning approach to cope with feature selection is proposed, searching for dependencies between attributes and the number of hidden states for each gesture using DNBCs structures with our database and feature set. Their results suggest the possibility to improve recognition rates with different attribute sets, associations and number of states for each gesture. We believe that these findings could lead to a fruitful research field in gesture recognition in the near future that may help us to improve our knowledge of gestures and to develop more accurate models for this purpose.

8. Conclusions

In this paper, an empirical comparison of DNBCs and standard HMMs was presented. DNBCs incorporate conditional independence among gesture features given the state into HMMs framework. DNBCs i) provide competitive error dispersion and recognition rates in various problems in gesture recognition, ii) require fewer parameters, iii) improve training time, and iv) permit structural learning and feature selection techniques to construct such dependences. In addition, we showed that a set of natural and simple posture and motion gestures allows us to correctly classify gestures. We also showed that classification performance of recognizers with these posture-motion data surpass motion-based ones. Also, an adaptive skin-color scheme to track the right hand of multiple people with different skin tones under different lighting conditions was described, and its implementation made available for other research groups. An extensive and comprehensive set of experiments was carried out with gestures taken from a single person, from multiple people, and with variations on distance and rotation. An additional product of this work is a freely accessible gesture database with more than 7000 samples of 9 gesture classes performed by 15 people. Our results show the effectiveness of the proposed approach and that DNBCs are a suitable alternative that opens the way to important issues such as feature selection and on-line learning.

References

[1] Starner T., Weaver J. & Pentland A., Real-Time American Sign Language Recognition Using Desk and Wearable Computer-Based Video, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 20, No. 12, Dec 1998, pp. 1371-1375.

[2] Lee H.K. & Kim J.H., An HMM-Based Threshold Model Approach for Gesture Recognition, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 21, No. 10, Oct 1999, pp. 1371-1375.

[3] Inoue M. & Ueda N., Exploitation of Unlabeled Sequences in Hidden Markov Models, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 25, No. 12, pp. Dec 2003, pp. 1570-1581.

[4] Pavlovic V., Sharma R. & Huang T.S., Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, 1997, pp. 677-695.

[5] Domingos P. & Pazzani M., On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Machine Learning, Vol. 29, No. 2-3, 1997, pp. 103-130.

[6] Friedman N., Geiger D. & Goldszmidt M., Bayesian Network Classifiers, Machine Learning, Vol. 29, No. 2-3, 1997, pp. 131-163.

[7] Avilés H. & Sucar L.E., Dynamic Bayesian networks for visual recognition of dynamic gestures, Journal of Intelligent and Fuzzy Systems, Vol. 12, No. 3–4, 2002, pp. 243-250.

[8] Hannaford B., Multi-dimensional hidden Markov model of Telemanipulation Tasks with Varying Outcomes, Proc. IEEE International Conference on Systems, Man and Cybernetics, 1990, pp. 127-133.

[9] Frasconi P., Soda G. & Vullo A., Text Categorization for Multi-page Documents: A hybrid Naive Bayes HMM Approach, Proc. ACM/IEEE Joint Conference on Digital Libraries, 2001, pp. 11-20.

[10] Pavlovic V., Garg A. & Kasif S., A Bayesian Framework for combining gene predictions, Bioinformatics, Vol. 18, No. 1, 2002, pp. 19-27.

[11] Xiang T. & Gong S., Incremental and adaptive abnormal behaviour detection, Computer Vision and Image Understanding, 2008, pp. 59-73.

[12] Lester J., Choudhury T., Kern N., Borriello G. & Hannaford B., A hybrid discriminative/generative approach for modeling human activities, Proc. Nineteeth International Joint Conference on Artificial Intelligence, 2005, pp. 766-722.

[13] Ahmad M. & Lee S.W., Human action recognition using shape and CLG-motion flow from multi-view image sequences, Pattern Recognition, Vol. 41, No. 7, 2008, pp. 2237-2252.

[14] Palacios M.A., Brizuela C.A. & Sucar L.E., Evolutionary Learning of Dynamic Naive Bayesian Classifiers, Proc. 21th International FLAIRS Conference, 2008, pp. 655-659.

[15] Palacios M.A., Brizuela C.A. & Sucar L.E., Evolutionary Learning of Dynamic Naive Bayesian Classifiers, Journal of Automated Reasoning, Vol. 45 No. 1, 2009, pp. 21-37.

[16] Avilés H., Sucar L.E. & Mendoza C.E., Visual Recognition of Similar Gestures, 18th International Conference on Pattern Recognition, 2006, pp. 1100-1103.

[17] Rabiner L.E., A tutorial on hidden Markov models and selected applications in speech recognition, Readings in speech recognition, Alex Waibel, Kai-Fu Lee Editors, Morgan Kaufmann, 1990, pp. 267-296.

[18] Wilson A. & Bobick A., Using Hidden Markov Models to Model and Recognize Gesture Under Variation, International Journal on Pattern Recognition and Artificial Intelligence, Special Issue on Hidden Markov Models in Computer Vision, Vol. 15, No. 1, 2000, pp. 123-160.

[19] Brand M., Olivier N. & Pentland A., Coupled hidden markov models for complex action recognition, Proc. IEEE Conference on Computer Vision and Pattern Recognition, 1999, pp. 994-999.

[20] Marcel S., Bernier O., Viallet J.E. & Collobert D., Hand gesture recognition using input-output hidden Markov models, Proc. Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 456-461.

[21] Vogler C. & Metaxas D.N., Parallel Hidden Markov Models for American Sign Language Recognition, Proc. International Conference on Computer Vision, 1999, pp. 116-122.

[22] Chambers G.S., Venkatesh S., West G.A.W. & Bui H.H., Hierarchical recognition of intentional human gestures for sports video annotation, Proc. 16th International Conference on Pattern Recognition, Vol. 2, 2002, pp. 1082-1085.

[23] Pavlovic V., Frey B.J. & Huang T.S., Variational learning in mixed-state dynamic graphical models, Proc. Uncertainty in Artificial Intelligence (UAI), 1999, pp. 522-530.

[24] Natarajan P. & Nevatia R., Hierarchical Multi-channel Hidden Semi Markov Models, Proc. International Joint Conference on Artificial Intelligence (IJCAI'07), 2007, pp. 2562-2567.

[25] Duong T., Bui H.H., Phung D.Q. & Venkatech S., Activity Recognition and Abnormality Detection with the Switching Hidden semi-Markov Model, Proc. 9th IEEE International Conference on Computer Vision, Vol.1, 2005, pp. 838-845.

[26] Artieres T., Marukatat S. & Gallinari P., Online Handwritten Shape Recognition Using Segmental Hidden Markov Models, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 29, No. 2, Feb 2007. pp. 205-217.

[27] Cassandra A.R., Kaelbling L.P. & Littman M.L., Acting optimally in partially observable stochastic domains, Proc. Twelfth National Conference on Artificial Intelligence (AAAI), Vol. 2, 1994, pp. 1023-1028.

[28] Hoey J. & Little J.J., Value-Directed Human Behavior Analysis from Video Using Partially Observable Markov Decision Processes, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 29, No. 7, Jul 2007, pp. 1118-1132.

[29] Rubine D., Specifying Gesture by Example, Computer Graphics, Vol. 25, No. 4, July 1991, pp. 329-337.

[30] Mardia K.V., Ghali N.M., Hainsworth T.J., Howes M. & Sheehy N., Techniques for online gesture recognition on workstations, Image and Vision Computing, Vol. 11, No. 5, 1993, pp. 283-294.

[31] Montero J.A. & Sucar L.E., Feature Selection for Visual Gesture Recognition Using Hidden Markov Models, Proc. Fifth Mexican International Conference in Computer Science, (ENC'04), 2004, pp. 1-8.

[32] Cui Y., Swets D. & Weng J., Learning-based hand sign recognition using SHOSLIF-16, Proc. 5th Int. Conf. Computer Vision, 1995, pp. 631-636.

[33] Matthews I., Cootes T.F., Bangham J.A., Cox S. & Harvey R., Extraction of Visual Features for Lipreading, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 24, No. 2, Feb 2002, pp. 198-213.

[34] Shanableh T., Assaleh K. & Al-Rousan M., Spatio-Temporal Feature-Extraction Techniques for Isolated Gesture Recognition in Arabic Sign Language, IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetics, Vol. 37, No. 3, June 2007, pp. 641-650.

[35] Johansson G., Visual Perception of Biological Motion and a model for its analysis, Perception and Psychophysics, Vol. 14, No. 2, 1973, pp. 201-211.

[36] Webb J.A. & Aggarwal J.K., Structure from motion from rigid and jointed objects, Artificial Intelligence, Vol. 19, No. 1, 1982, pp. 107-130.

[37] Shah M., Understanding human behavior from motion imagery, Machine Vision and Applications, Vol. 14, No. 1, 2003, pp. 210-214.

[38] Giese M.A. & Poggio T., Morphable Models for the Analysis and Synthesis of Complex Motion Patterns, International Journal of Computer Vision, Vol. 38, No. 1, 2000, pp. 59-73.

[39] Bobick A.F. & Davis J.W., The recognition of human movement using temporal templates, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 23, No. 3, Mar 2001, pp. 257-267.

[40] Waldherr S., Gesture Recognition on a Mobile Robot, Diploma thesis, Carnegie Mellon University. School of Computer Science, 1998.

[41] Beintema J.A. & Lappe M., Perception of Biological motion without local image motion, Proc. of the National Academy of Sciences, Vol. 4, No. 8, 2002, pp. 5661-5663.

[42] Sigala R., Serre T., Poggio T. & Giese M., Learning Features of Intermediate Complexity for the Recognition of Biological Motion, International Conference on Artificial Neural Networks (ICANN), 2005, pp. 241-246.

[43] Casile A. & Giese M., Roles of motion and form in biological motion and recognition, International Conference on Artificial Neural Networks (ICANN), 2003, pp. 854-862.

[44] Casile A. & Giese M., Critical features for the recognition of biological motion, Journal of Vision, Vol. 5, 2005, pp. 348-360.

[45] Stokoe W., Sign Language Structure, University Buffalo Press, 1960.

[46] Just A., Bernier O. & Marcel S., Recognition of Isolated Complex Mono and BiManual 3D Hand Gestures, Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004, pp. 571-577.

[47] Ren H., Xu G. & Kee S.C., Subject-independent Natural Action Recognition, Proc. Sixth IEEE Conference on Automatic Face and Gesture Recognition, 2004, pp. 523-528.

[48] Corradini A. & Gross H.M., Camera-based Gesture Recognition for Robot Control, IEEE-INNS-ENNS International Joint Conference on Neural Networks, Vol. 4, 2000, pp. 133-138.

[49] Campbell L.W. , Becker A.D., Azarbayejani A., Bobick A.F. & Pentland A., Invariant features for 3-D Gesture Recognition, Technical report 379, M.I.T. Media Laboratory Perceptual Computing Section, 1996.

[50] Vogler C. & Metaxas D., ASL Recognition based on a Coupling Between HMMs and 3D Motion Analysis, Proc. International Conference on Computer Vision (ICCV'98), 1998, pp. 363-369.

[51] Ahmad M. & Lee S.W., Human Action Recognition Using Multi-View Image Sequences Features, Seventh International Conference on Automatic Face and Gesture Recognition, pp. 523-528, 2006.

[52] Baum L.E., Petrie T., Soules G. & Weiss N., A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Ann. Math. Stat., Vol. 41, No. 1, 1970, pp. 164-171.

[53] Bilmes J.A., A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, U.C. Berkeley, TR-97-021, *http://citeseer.ist.psu.edu/1570.html*, 1998.

[54] Rabiner L. & Juang B.H., Fundamentals on Speech Recognition, Prentice-Hall Signal Processing Series, New Jersey, 1993.

[55] Viola P.A. & Jones M.J., Robust Real-time Object Detection, International Journal of Computer Vision, Vol. 57, No. 2, May 2004, pp. 137-154.

[56] Azpeitia L.G.G., *Con la Vara que Midas*. Universidad de Colima, Colima, México. 1987. (In Spanish).

[57] Jones M.J. & Rehg J.M., Statistical Color Models with Application to Skin Detection, Technical report CRL-98/11, Cambridge Research Laboratory, 1996.

[58] Avilés H. & Sucar L.E., Real-Time Visual Recognition of Dynamic Arm Gestures, Video-Based Surveillance Systems: Computer Vision and Distributed Processing, P. Remagnino, P., G.A. Jones, N, Paragios, C.S. Regazzoni, Editors, Kluwer Academic, 2002, pp. 227-238.

[59] Manyika J. & Durrant-Whyte H., Data Fusion and Sensor Management: A descentralized Information-Theoretic Approach, Ellis Horwood, NY-London, 1994.

[60] Bradski G.R., Real Time Face and Object Tracking as a Component of a Perceptual User Interface, Proc. 4th IEEE Workshop on Applications of Computer Vision (WACV'98), 1998, pp. 214-219.

[61] Kanungo T., Hidden Markov Models Software, Available at: *http://www.kanungo.com/*. Last retrieved: May 26, 2008

[62] Kendon A., An agenda for gesture studies, Semiotic Review of Books, Vol. 7 No. 3, pp. 8-12, 1996. Available at: *http://www.univie.ac.at/Wissenschaftstheorie/srb/srb/gesture.html.*

[63] Yang H.D., Park A.Y. & Lee S.W., Gesture Spotting and Recognition for Human-Robot Interaction, IEEE Trans. in Robotics, Vol. 23, No. 2, Apr 2007, pp. 256-279.

[64] Elmezain M., Al Hamadi A., Appenrodt J. & Michaelis B., A Hidden Markov Model-based continuous gesture recognition system for hand motion trajectory, 19th International Conference on Pattern Recognition, 2008, pp. 1-4.

[65] van Son R.J.J.H., The Relation Between the Error Distribution and the Error Rate in Identification Experiments, Proc. European Conference on Speech Communication and Technology, 1995, pp. 2277-2280.

[66] Shannon C.E., A Mathematical Theory of Communication, Bell System Technical Journal, Vol. 27, 1948, pp. 379-423 and 623-656.

[67] Wu Y. & Huang T.S., Vision-Based Gesture Recognition: A Review, Gesture-Based Communication in Human-Computer Interaction, A. Camurri, G. Volpe, Springer Berlin / Heidelberg, Vol. 1739/1999, 1999, pp. 103-115.

[68] Parameswaran V. & Chellappa R., Human action-recognition using mutual invariants, Computer Vision and Image Understanding, Vol. 98, 2005, pp. 295-325.

[69] Friedman N., Murphy K. & Russell S., Learning the Structure of Dynamic Probabilistic Networks, Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI), 1998, pp. 139-147.

[70] Bressan M. & Vitria J., On the Selection and Classification of Independent Features, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 25, No. 10, Oct 2003, pp. 1312-1317.

Authors´ Biographies

Héctor AVILÉS-ARRIAGA

Héctor Avilés has a bachelor's degree in computer science from the Instituto Tecnológico de Ciudad Madero in 1997, master's and doctor's degrees in computer science (2000 and 2006, respectively) from the Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Cuernavaca. He held postdoctoral appointments from the Instituto Nacional de Astrofísica, Óptica y Electrónica (2006-2007) and from the Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas of the Universidad Nacional Autónoma de México (2008-2010). He is currently a member of the academic staff of the Computer Department at the Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas. Dr. Avilés has written more than 25 papers in journals, book chapters, international conferences and workshops. His research interests include visual recognition of gestures, multimodal human-robot interaction combining gestures and speech, and design and implementation of intelligent service robots.

L. Enrique SUCAR-SUCCAR

L. Enrique Sucar has a Ph. D. in computing from Imperial College, London, UK, 1992; an M.Sc. in electrical engineering from Stanford University, California, USA, 1982; and a B.Sc. in electronics and communications engineering from ITESM, Monterrey, Mexico, 1980. He has been a researcher at the Electrical Research Institute and professor at ITESM Cuernavaca and is currently a senior researcher at INAOE, Puebla, Mexico. He has been an invited professor at the University of British Columbia, Canada; Imperial College, London; and INRIA, France. He has more than 100 publications in journals and conference proceedings, and has directed 15 Ph.D. Thesis. Dr. Sucar is a member of the National System of Researchers, the Mexican Science Academy, AAAI, SMIA and senior member of the IEEE. He has served as president of the Mexican AI Society, has been a member of the Advisory Board of IJCAI, and is associate editor of the journals Computación y Sistemas and Revista Iberoamericana de Inteligencia Artificial. His main research interests are in graphical models and probabilistic reasoning, and their applications in computer vision, robotics and biomedicine.

Carlos Eduardo MENDOZA-DURÁN

Carlos Eduardo Mendoza-Durán holds a bachelor's degree in actuarial science from UNAM (1976), an M.A. in statistics from Princeton (1982), a Ph. D. in statistics from Princeton (1984). He is currently a full-time professor at Universidad Anáhuac México Norte at the School of Engineering. His main interests are data analysis, artificial intelligence and applied statistics.

Luis A. PINEDA-CORTÉS

Luis A. Pineda has a bachelor's degree in electronics from Universidad Anáhuac in Mexico City, an M. Sc. in computer science from ITESM, Campus Morelos and a Ph. D. in cognitive science from the University of Edinburgh (1986-1989). He has been the data center manager of NCR in Mexico City (1981-1983), a researcher at Instituto de Investigaciones Eléctricas (IIE) in Cuernavaca, Mexico (1983-1986 and 1992-1998) and also a research associate at the Human Communication Research Centre (HCRC) at the University of Edinburgh (1989-1992). Since 1998, he has worked as an associate researcher in the Department of Computer Science at Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS) of the Universidad Nacional Autónoma de México (UNAM), where he has been the head twice (1998-2002 and 2005-2010). He has published extensively on computational linguistics and artificial intelligence. Dr. Pineda is a regular member of the Mexican Academy of Science, a member of the National System of Researchers (SNI), level II, and since January 2010 he has been the coordinator of the Mexican Network for Research and Development in Computer Science (REMIDEC).