Visual Recognition of Dynamic Gestures Applied to Command Mobile Robots

Héctor Hugo Avilés-Arriaga[*], **Luis Enrique Sucar**[†], **Carlos Mendoza**[‡] **and Blanca Vargas**[⋆]
Tec de Monterrey, Campus Cuernavaca
Av. Paseo de la Reforma No. 182-A
Col. Lomas de Cuernavaca
C.P. 82589 Cuernavaca Morelos México
[*] 00374765@academ01.mor.itesm.mx, [†] esucar@campus.mor.itesm.mx
[‡] cmendoza@alumni.princeton.edu, [⋆] blanca.vargas@itesm.mx

Abstract

Visual recognition of dynamic gestures is a natural and effective way to command mobile robots. Probabilistic graphical models such as hidden Markov models (*HMMs*) and dynamic Bayesian networks *DBNs* are suitable techniques to describe uncertainty involved in the motion of a gesture. In this document, we argue that in addition to motion information, arm posture knowledge, in the form of spatial relations among human body parts like face and shoulders is important for gesture recognition. To incorporate relational knowledge, we propose an hybrid framework for knowledge representation called probabilistic-logic networks (*PLNs*). A PLN is composed of a set of logic and stochastic nodes combined into a Bayesian networks framework. Logic nodes consists of logic programs in the form of Horn clauses. Stochastic nodes are classic random variables. Tested in real-time recognition of dynamic gestures, PLNs show significantly better recognition results than DBNs.

1 Introduction

Visual recognition of gestures applied to command mobile robots is a challenging and interesting field of study in Human-Machine Interaction research. Gestures provides a natural form of communication with mobile robots and an alternative and complement to speech on noisy environments. When using gestures, it is possible to communicate information of the kind of "go there" or "this object". This enable us to consider mobile robots as *service robots*, useful to aid humans in everyday tasks.

Literature [Starner, 1995; Nam *et al.*, 1999; Martin and Durand, 2000] presents different alternatives to recognize gestures in terms of their motion, being hidden Markov models (*HMMs*) the standard technique to do so. Recently, dynamic Bayesian networks (*DBNs*) have been employed for gesture recognition with good performance. These techniques consider uncertainty, which is present in gesture executions due to possible variations in the form, force, amplitude or velocity, when it is executed even by the same user [Morris, 1977]. In this paper we argue that spatials relations among human body parts, like face and shoulders, can improve the performance of gesture recognition. First-order logic is a suitable formalism to describe this kind of relational knowledge. Because of this, it is reasonable to think in merging logic and probability to represent at the same time posture and motion information, respectively. We present a methodology to recognize visually a set of natural dynamic gestures with a new model called *probabilistic-logic networks* [Morales *et al.*, 2000]. PLNs are composed of a set of logic and stochastic nodes combined into a Bayesian networks framework. Logic nodes consists of logic programs in the form of Horn clauses. This new model was compared experimentaly with DBNs. In our experiments, we obtained better recognition results using PLNs than using dynamic Bayesian networks.

The outline of this document is as follows. Section 2 describes probabilistic-logic networks. In section 3 we present the vision techniques developed to extract posture and motion information of the user. Section 4 presents a PLN model that we employ to recognize gestures. Experiments and results are presented in section 5. Section 6 shows our conclusions and future work.

2 Probabilistic-logic networks

Several approches have been proposed to combine probability and logic, such as [Bacchus, 1990; Koller and Halpern, 1996; Poole, 1997], among others. In this document we propose an alternative model based on *probabilistic-logic networks* [Morales *et al.*, 2000]. A PLN extends Bayesian networks with *logic nodes*. Logic nodes are arbitrary logic programs in the form of Horn-clauses. The goal is to incorporate relations among variables (*e.g.*, an object is above another, two circles are concentric, *etc.*). A PLN is a representation with logic and stochastic nodes interacting in an hybrid mechanism. A logic node expresses a relation between its input variables. This node evaluates to true when the relation holds for certain values of its input nodes.

Formally, a PLN is a directed acyclic graph G with a set $X = (D, L)$ of nodes, where D and L are sets of stochastic and logic nodes, respectively. Each node $x \in X$ has associated a dependency structure and functional structure. The dependency structure is defined by the set of parents $Pa(x)$. For each stochastic node $d \in D$, the functional structure is defined by a conditional probability table (if d is discrete) or a probability density function (if d is continuous). The functional

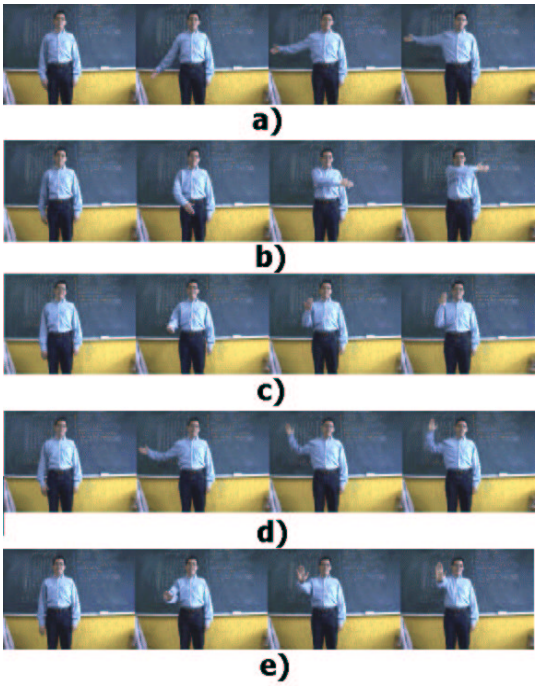

Figure 1: *Gestures considered by our system: a) go-right, b) go-left, c) come, d) attention and e) stop.*

structure of a logic node $l \in L$ corresponds to a deterministic conditional probability table, commonly constructed once it is known the instantiated values of $Pa(l)$. PLNs can employ binary or multi-valued logic nodes.

There are two basic approaches for inference in a PLN: (i) conversion to a standard BN and, (ii) stochastic simulation. When all the variables in the model are discrete, the logic nodes can be transformed into a conditional probability table (*CPT*) by evaluating the logic function for all the values of its arguments. In this way, the model is transformed to a BN, so standard probability propagation techniques can be employed [Pearl, 1988]. When there are continuous variables (or the CPT is too large), we use stochastic simulation techniques [Pearl, 1988], evaluating each logic node according to the values of its parent variables in each iteration. In case the parent variables of the logic nodes are all observed (known) root nodes, the logic nodes can be evaluated directly for the observed values and the simulation is not required.

The proposed representation has the following advantages. First, it mantains the full expressive power of logic programs (represented as Horn-clasuses) within a BNs framework. Second, it is an alternative to deal with continuous variables in BNs; which in some cases do not need to be discretized. Third, this model can be easily extended to represent dynamic processes by incorporating logic nodes into DBNs.

3 Vision techniques

In order to locate and track the motion of the user, we employ a skin pixel classifier based on the method developed by Jones [Jones and Rehg, 1996] and a radial scan segmentation algorithm proposed by SAVI Group [SAVI, 1999]. The skin pixel classifier is based on histogram color models information and the Bayes rule for skin or non-skin pixels detection. The segmentation algorithm traces lines over the image with certain angular distance among them, from the center of the image to its edges, classifying pixels over these lines, as skin or non-skin pixels. At the same time, it uses some segmentation conditions to grow skin regions. These algorithms are applied over the image to locate the user face and his right-hand. After the right-hand is localized, it can be tracked in the image sequence employing a search window around its previous position. Some images that show face segmentation and hand tracking are presented in figure 2. An extended explanation of these vision techniques can be found in [Avilés, 2000].

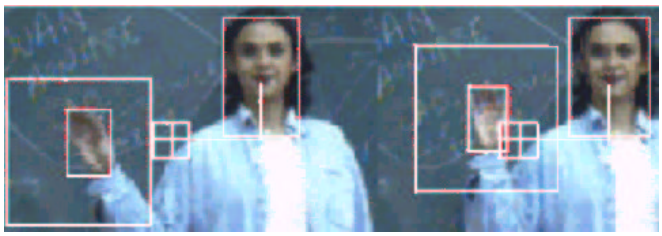

Figure 2: Tracking of the user's right hand.

4 PLNs for gesture recognition

In order to recognize dynamic gestures with PLNs, consider the topology as it is shown in figure 3. This topology corresponds to a dynamic PLN -i.e., a dynamic Bayesian network [Kjærulff, 1992] augmented with logic nodes- unrolled two times.

We divide observation nodes of the nerwork into two sets, motion observation nodes and posture observation nodes. In this topology, motion observation nodes correspond to four simple features used to describe the hand displacement: $\Delta area$ of changes in area of the hand, Δx or changes in hand position on the *x-axis* of the image, Δy or changes in hand position on the *y-axis* of the image and *form* or comparison between sides of the square region that segments the hand. Δx and Δy are *joined* in a single node because with this topology we obtained better results than using one node per variable. This operation has a direct relationship with Pazzani's work [Pazzani, 1995]. To estimate depth motion in a simple way, we use the $\Delta area$ feature. To evaluate the hand motion between two images, each of these features takes only one of three possible values: (+), (-) or (0) that indicate increment, decrement or no change, depending on the area, position and form of the hand in the previous image.

Posture observation nodes correspond to twelve nodes -$C1$ to $C12$- that define 2-D cartesian coordinates for right hand, head and torso of the user. We take these observations directly from the image, without discretization. These nodes are parents of the three logic nodes employed in this model. Given that these nodes are always observed, a probability distribution is not required and probability propagation can be done directly, by evaluating the logic nodes on-line.

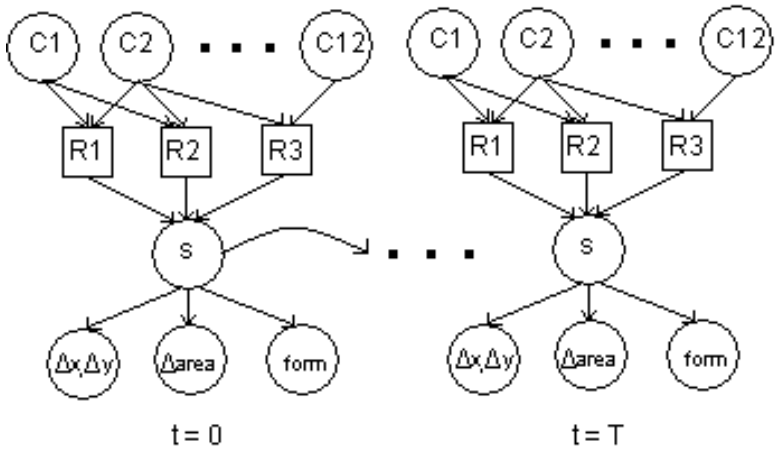

Figure 3: *Dynamic probabilistic-logic network for gesture recognition. Nodes C*1 *to C*12 *correspond to 2-D cartesian coordinates. s is the state of the system.* Δx, Δy, $\Delta area$ *and form are motion feature observations.*

Logic nodes *R*1,*R*2 and *R*3 describe logic programs that we call *right*, *above* and *torso*. They represent spatial relations such as if hand is to the right of the head, above or if the hand is over the user's torso, respectively. The combination of the evaluation of these relations provides spatial information about the arm posture. Given that these relations implicitly establish a reference system based on the user, it is less sensible to the distance between the user and the camera -or different users- than other systems based on a relative motion [Avilés, 2000]. Nodes *s* represents the state of the system at each time *t*.

4.1 Experiments and results

In this section we describe the recognition results using both PLNs and DBNs, and our initial experiments for telecontrolling a mobile robot. Our visual interface runs on a Silicon Graphics O2 5000. When a gesture is recognized, it is sent to the robot using two programs running into a socket client/server arquitecture.

The dynamic Bayesian network used for gesture recognition [Avilés, 2001] is shown in figure 4. The experiments for the visual recognition of gestures using PLNs and DBNs can be divided in two parts: training and recognition stages. For training, we employed an average of 150 sequences of observations for each gesture taken from one person. To test the recognition rates of our system, one user made an average of 54 executions for each gesture in front of the videocamera. The gestures we used to test the system were different from those used for training. In order to define the start and the end of a gesture, a small tolerance region was established around the initial position of the hand. The return of the hand to its initial position defines when the gesture is completed.

For training and testing PLNs and DBNs we employed the Expectation-Maximization algorithm [Dempster *et al.*, 1977] and the Forward algorithm [Rabiner, 1990].

Table 1 shows the recognition results using dynamic Bayesian networks with an *ergodic* (or fully connected) [Rabiner, 1990] topology of 2 states described in [Avilés, 2001]. Recognition results with PLNs are presented in table 2.

	Come	Stop	Go-left	Go-right	Attention
Come	92.45%	1.88%	5.66%		
Stop	3.45%	93.10%	1.72%		1.72%
Go-left			100%		
Go-right				100%	
Attention		74.07%			25.93%

Table 1: Recognition results using DBNs. Rows represent the percentage of correct classification in the execution of each gesture as well as the percentage of incorrect classification. The average recognition rate is 82.52%

The two models obtained the same recognition result in go-left and go-right gestures. These results are easily explained considering the nature of the necessary movements in one or another case. Whereas the gestures go-left and go-right require displacements that predominate towards the left and the right (parallel to the image plane), the gestures come, attention and stop involve movements in depth (perpendicular to the image plane). Direct depth information is not considered, however, with PLNs we obtained better recognition results in gestures attention and come. As in previous work [Avilés, 2001] with DBNs, the main error in recognition with PLNs appears between stop and attention gestures. We believe that the inclusion of new spatial relationships in PLNs can avoid this problem. For remote operation of a mobile robot, we have performed some initial experiments in a laboratory environment with promising results. Two different users executed a set of gesture repetitions to move the robot from one point to another. Each gesture that is recognized by the visual system, it is sent to the robot, and the predefined motion command is executed. At this initial stage, we are able to move the robot using the go-right, go-left, stop and come gestures. In the future, we will conduct additional experiments to evaluate this telecontrol gesture interface.

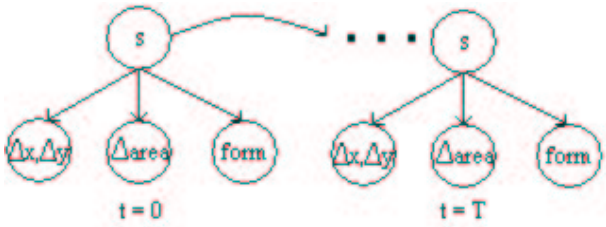

Figure 4: Topology for the dynamic Bayesian network employed in our system.

	Come	Stop	Go-left	Go-right	Attention
Come	100%				
Stop		75.86%			24.14%
Go-left			100%		
Go-right				100%	
Attention					100%

Table 2: Recognition results using PLNs. The average recognition rate is 95.14%

5 Discussion

We have presented our initial results in gesture recognition using probabilistic-logic networks. When using PLNs, it is possible to represent explicit spatial relations among different human body parts like face, hand and torso, in a more natural form than using dynamic Bayesian networks. In fact, with this model we can employ cartesian coordinates taken directly from the image, without discretization. Moreover, relational knowledge permit us to employ reference points based on the user, that is important if there exists variations in the distance from the user to the videocamera.

An alternative to include posture information within a standard DBN is to add the logical relations as additional observation nodes. However, if some coordinates are not observed, (*e.g.*, occlusions), in this case it will not be possible to estimate the relations. In the PLN this is not a problem, missing data can be considered using propagation via stochastic simulation.

The initial model presented for recognizing our 5 gestures does not take full advantage of the descriptive power of logic nodes in the form of Horn clauses. However, as we extend the gesture language or develop a grammar, the expressiveness of logic will become more important.

6 Conclusions and future work

This document describes a online methodology to recognize five dynamic gestures and compare recognition results between two techniques, probabilistic-logic networks and dynamic Bayesian networks. A PLN consists on a set of logic programs and stochastic variables combined into a Bayesian networks framework. The objective of applying PLNs for gesture recognition, is to capture posture and motion information involved in the execution of a gesture. Recognition results shows a significant improvement when posture information on the form of spatial relations of different body parts is combined with motion. As a future work we plan to include new gestures and include motion and posture information of other body parts like elbow and shoulder.

References

[Avilés, 2000] Héctor Avilés. Reconocimiento de gestos dinámicos aplicado a robots móviles. Master's thesis, Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Cuernavaca, 2000.

[Avilés, 2001] Héctor Avilés and Enrique Sucar Succar. Reconocimiento visual de gestos dinámicos empleando redes bayesianas dinámicas. In *VI Taller Iberoamericano de Reconocimiento de Patrones*, 2001. In Spanish.

[Bacchus, 1990] Fahiem Bacchus. On probability distributions over possible worlds. In *Uncertainty in Artificial Intelligence 4*, pages 217–227. Elsevier Science Publisher B.V., 1990.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[SAVI, 1999] Stereo Active Visual Interface Group. Available at: http://ww.cs.toronto.edu/˜herpers/projects.html, May 28, 1999.

[Jones and Rehg, 1996] Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. Technical Report CRL 98/11, Cambridge Research Laboratory, 1996.

[Kjærulff, 1992] Uffe Kjærulff. A computational scheme for reasoning in dynamic probabilistic networks. In *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*, pages 121–129, 1992.

[Koller and Halpern, 1996] Daphne Koller and Joseph Y. Halpern. Irrelevance and conditioning in first-order probabilistic logic. In *13th International Conference on Artificial Intelligence (AAAI)*, pages 569–576, 1996.

[Martin and Durand, 2000] Jerome Martin and Jean-Baptiste Durand. Automatic gesture recognition using hidden markov models. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[Morales *et al.*, 2000] R. Morales, E. Morales, and L.E. Sucar. Integrating bayesian networks with logic programs for music. In *2000 International Computer Music Conference*, pages 320–323, 2000.

[Morris, 1977] Desmond Morris. *El hombre al desnudo*, volume 1. Ediciones Orbis, Barcelona, España, 1977.

[Nam *et al.*, 1999] Yanghee Nam, Kwangyun Wohn, and Hyung Lee-Kwang. Modeling and recongnition of hand gesture using colored petri nets. *IEEE Transactions on Systems Man, and Cybernetics*, 29(5):514–521, 1999.

[Pazzani, 1995] Michael Pazzani. Searching for dependencies in bayesian classifiers. In *Fifth International Workshop on AI and Statistics*, 1995.

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan, 1988.

[Poole, 1997] David Poole. Independent choice logic for modelling multiple agents under uncertainty. *Artificial Intelligence*, 94:7–56, 1997.

[Rabiner, 1990] Lawrence R. Rabiner. *Readings in Speech Recognition*, chapter A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Morgan Kaufmann Publishers, 1990.

[Starner, 1995] Thad Eugene Starner. Visual recognition of american sign language using hidden markov models. Master's thesis, MIT. Program in Media Arts and Science, 1995.