Towards a General Vision System based on Symbol-Relation Grammars and Bayesian Networks

Elias Ruiz, Augusto Melendez, and L. Enrique Sucar

Computer Science Department,
National Institute of Astrophysics, Optics and Electronics.
Luis Enrique Erro 1, 72840 Tonantzitla, México
{elias_ruiz, amelendez, esucar}@inaoep.mx

Abstract. A novel approach to create a general vision system is presented. The proposed method is based on a visual grammar representation which is transformed to a Bayesian network which is used for object recognition. We use a symbol-relational grammar for a hierarchical description of objects, incorporating spatial relations. The structure of a Bayesian network is obtained automatically form the grammar, and its parameters are learned from examples. The method is illustrated with two examples for face recognition.

1 Introduction

Although there have been important advances in computer vision in the last decades, we are still far from a *general* vision system with capabilities similar to a human child. Most developments in object recognition have focused on high performance systems for particular applications; and lately mainly on recognizing specific object based on local features.

In the beginnings of the computer era, there were some intents to develop more general vision systems, but these were not successful due to several problems, including lack of computer power, and limited feature detection and recognition techniques. However, in recent years, with the development of very powerful and inexpensive computer platforms, and the advances in several areas of computer vision and artificial intelligence, the time for developing more general methods has arrived.

Some recent developments are based on visual grammars or biologically inspired. For instance, Zhu and Mumford [10] describe a general visual grammar representation using And-Or graphs. This model is limited in the sense that it does not consider the spatial relation between the visual elements, which are very important for recognition (e.g., the configuration of the elements of a face). On the other hand, models of the biological visual system have provided the basis for building computer vision models. Serre and Poggio [8] achieve a competitive recognition rate in real images, learning through examples of images using terminal elements called *patches*. However, it lacks a structure that allows to

incorporate prior knowledge, and it is not defined within a grammar or a formal representation.

We propose an approach that is also based on visual grammars with a biological inspiration, but trying to overcome some of the limitations of the previous works. Objects are represented using a visual hierarchy based on Symbol-Relational grammars that incorporate spatial relations between terminal and non-terminal elements. The terminal elements are biologically inspired, including edges and color patches. To incorporate uncertainty, the visual grammar is transformed to a Bayesian network (BN) [7], whose structure is generated automatically from the grammar and its parameters are learned from examples. Recognition is performed using standard BN inference techniques.

We present two preliminary examples of the proposed method for face recognition. One uses high-level elements and was compared with other state of the art methods for face detection. The other illustrates the low-level features for eye detection.

2 Representation and Recognition

2.1 Visual Grammar

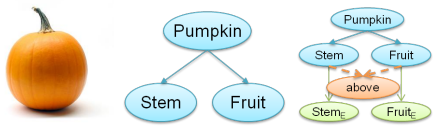
A visual grammar describes objects hierarchically. For our model, we need a grammar that allows us to model the decomposition of an object into its parts and how they relate with another parts. Symbol-Relation grammars *(SR grammars)* [2], provide this type of description and incorporate the possibility to add rewriting rules for relations between terminals and non terminals symbols.

In the productions of the grammar we can incorporate relations between elements. In our work, we incorporated spatial relations, which can determine the position of an object with respect to another object. Although there are different types of spatial relations, in our model we use *topological* and *order* relations, such as *inside_of* and *above*. Figure 1 shows a simple example of an object represented using *And-Or* graphs vs. a BN based on a SR grammar. Fig. 1b shows a simple *And-Or* graph. In the Fig. 1c, node *above* is added to represent the relations between *stem* and *fruit* nodes. This information is not clearly represented in an *And-Or* graph so we obtain a more expressive representation of the object.

2.2 Transforming a SR-Grammar to a Bayesian network

If we apply our model in real images, this will involve uncertainty in the detection of the elements and their relations. To manage uncertainty we transform the SR grammar to a Bayesian network, where a node can represent either a symbol or a relation. However, a visual grammar can lead to endless productions. To avoid this, we incorporate a restriction on SR grammar so they can be transformed into a Bayesian network. This restriction eliminates cyclic rules, for example: $A^0 \rightarrow \langle B^2 \rangle$ and $\mathrm{B}^0 \rightarrow \langle A^2 \rangle$, where A produces B and B produces A. The

restriction is that for every rule of the form $Y^0 \rightarrow \langle \mathbf{M}, \mathbf{R} \rangle$ and for all $m \in \mathbf{M}$ it holds that Y^0 is not *son* of m. Conversion is based on creating a Bayesian network with a structure similar to an *And-Or* graph, but incorporating spatial relations.



(a) Pumpkin (b) And-or Graph (c) Bayesian network

Fig. 1. Representations for the pumpkin object. (b) A simple *and-or graph* can not represent the topological relations between terminal or not terminal elements. (c) A Bayesian network representation obtained from the SR Grammar which incorporates an order relation (*above*), and additional virtual nodes that consider the uncertainty in the detectors.

Conversion algorithm We convert the SR grammar into a Bayesian network where the root node is the first element of the grammar, and the other nodes are terminals and nonterminals elements of the SR grammar. This is detailed in Algorithm 1. Briefly, the conversion algorithm performs the following steps:

1. Set the root node.
2. For each s-production rule, where the term on the left is the reference node (Nr), and for each symbol defined in each production, Add p_i as a child of Nr. If p_i is not terminal, perform a recursive call with p_i as the new Nr. If p_i is terminal, Add p_{i_E} (as evidence node) as a child of p_i.
3. For each relation $r(x, y)$, add the node as a child of his parents x and y.

3 Examples

We describe two initial examples of the application of our method for face representation, one based on high-level elements and other using low-level features.

3.1 Visual Grammar for Face Detection

The following visual grammar is used to describe high-level items in images of faces (front view), and we define it as follows:

$$FG = (\{FACE\}, \{eyes, nose, mouth, head\}, \{above, inside\_of\}, FACE, S, \emptyset)$$

The S-productions are defined by:

$$1 : FACE^0 \rightarrow < \{eyes^2, mouth^2\}, \{above(eyes^2, mouth^2)\} >$$
$$2 : FACE^0 \rightarrow < \{nose^2, mouth^2\}, \{above(nose^2, mouth^2)\} >$$

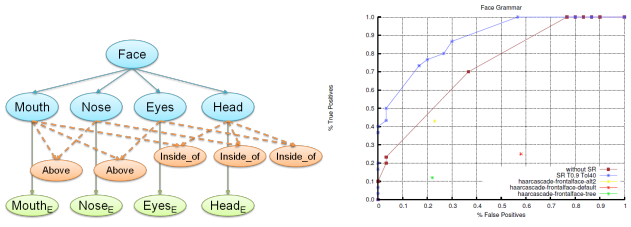
Algorithm 1 Convert SR-Grammar to Bayesian network.

Data: $G(V_N, V_T, V_R, S, P, R), Nr$; `/* Nr=Reference Node */`
Result: Bn

if $Nr = S$ **then**
  Set S as root node in Bn
foreach $p_i \in P$ ***where*** $Y^0 = Nr$ **do**
  `//` p_i `has the form` $l : Y^0 \rightarrow \langle \mathtt{M}, \mathtt{R} \rangle$
  foreach $m \in \boldsymbol{M}$ **do**
    Add p_i as child of Nr
    if $p_i \in V_N$ **then**
      `ConvertSRGtoBN`(G, p_i); `/* Recursion */`
    if $p_i \in V_T$ **then**
      Add p_{i_E} as child of p_i
  foreach $r_i \in \boldsymbol{R}$ **do**
    `//` r `has the form` $r(X, Y)$
    Add node r_i as child of X and Y.

$$3 : FACE^0 \rightarrow < \{eyes^2, head^2\}, \{inside\_of(eyes^2, head^2)\} >$$
$$4 : FACE^0 \rightarrow < \{nose^2, head^2\}, \{inside\_of(nose^2, head^2)\} >$$
$$5 : FACE^0 \rightarrow < \{mouth^2, head^2\}, \{inside\_of(mouth^2, head^2)\} >$$

From this SR grammar, and using the conversion algorithm, we obtained a BN representation (Fig. 2a). Once the structure is obtained from the grammar, the parameters are learned using standard parameter learning [6] from a set of training images of faces (in this case, 200 images). The elements of the face are obtained from object recognizers based on the AdaBoost algorithm [9]. As expected, the spatial relations helped significantly in the recognition task (Fig. 2b). More details of this work are described in [5].



(a) Bayesian Network Structure (b) Detection Rate Graph

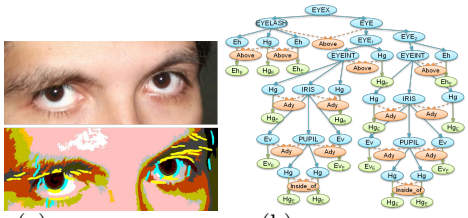
Fig. 2. Face detection using a SR grammar implemented as a BN. (a) The Bayesian network structure obtained from the SR grammar. (b) The graphs show the detection rate varying the decision threshold with and without spatial relations. The method is compared against three variants of the Viola and Jones face detector [1] with fixed thresholds (dots).

3.2 Low-level features of a Visual Grammar for Eyes

This grammar defines an eye based on bio-inspired features:
$G = (\{EYEX, EYELASH, EYE, EYEINT, IRIS, PUPIL\}\{Eh, Ev, Hg\}, \{above, ady, inside\_of\}, EYE, S, \emptyset)$
Where S is formed by the S-Productions:

$$
\begin{aligned}
1: EYEX^0 &\rightarrow < \{EYELASH^2, EYE^2\}, \{above(EYELASH^2, EYE^2)\} > \\
2: EYELASH^0 &\rightarrow < \{Eh^2, Hg^2, Eh^3\}, \{above(Eh^2, Hg^2), above(Hg^2, Eh^2)\} > \\
3: EYE^0 &\rightarrow < \{EYEINT^2, Hg^3\}, \{above(EYEINT^2, Hg^3)\} > \\
4: EYE^0 &\rightarrow < \{EYEINT^2, Eh^4\}, \{above(EYEINT^2, Eh^4)\} > \\
5: EYEINT^0 &\rightarrow < \{Hg^4, IRIS^2, Hg^5\}, \{ady(Hg^4, IRIS^2), ady(IRIS^2, Hg5)\} > \\
6: IRIS^0 &\rightarrow < \{Ev^2, PUPIL^2, Ev^3\}, \{ady(Ev^2, PUPIL^2), ady(PUPIL^2, Ev^3)\} > \\
7: PUPIL^0 &\rightarrow < \{Hg^6, Hg^7\}, \{inside\_of(Hg^6, Hg^7)\} >
\end{aligned}
$$

This visual grammar was specified manually from examples obtained by the segmentation algorithm as shown in Fig. 3a. The generated Bayesian network, is illustrated in Fig. 3b. In order to build a dictionary of low-level features that conform the terminal elements of the grammar, we built a simplified approach that considers some aspects of the visual system [8,3]. Once recognized certain edges of orientation (0° and 90°) with Gabor filters [4], we segment the rest of the image *homogeneous zones*, which are quantized to 32 colors.



(a) Terminal Elements (b) Structure of BN

Fig. 3. A visual grammar for eyes. (a) Terminal elements before (top) and after (bottom) segmentation. (b) BN generated from the SR grammar.

4 Conclusions and Future Work

A first stage in the design of a general vision system was described. This approach combines visual SR grammars and Bayesian networks to represent and recognize objects in an image. The model was tested for face recognition with high-level features as terminal elements with promising results. There are several avenues for future research. One is to develop a more complete grammar for faces from the low-level features to the high-level elements. Other is to explore alternative representations based on relational Bayesian networks. We plan in the future to apply this formalism to other classes of objects and to learn the visual grammar from images.

References

1. Qingcang Yu A, Harry H. Cheng A, Wayne W. Cheng B, and Xiaodong Zhou B. Ch opencv for interactive open architecture computer vision, 2004.
2. F. Ferrucci, G. Pacini, G. Satta, M. I. Sessa, G. Tortora, M. Tucci, and G. Vitiello. Symbol-relation grammars: a formalism for graphical languages. *Inf. Comput.*, 131(1):1–46, 1996.
3. Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.
4. D. Gabor. Theory of communication. *JIEE*, 93(3):429–459, 1946.
5. Augusto Melendez, Luis Sucar, and Eduardo Morales. A visual grammar for face detection. In Angel Kuri-Morales and Guillermo Simari, editors, *Advances in Artificial Intelligence - IBERAMIA 2010*, volume 6433 of *Lecture Notes in Computer Science*, pages 493–502. Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-16952-6-50.
6. R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2003.
7. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, 1988.
8. T. Poggio T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Technical Report CBCL-259, MIT Artificial Intelligence Laboratory, December 19 2005.
9. Paul Viola and Michael J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57:137–154, May 2004.
10. Song Chun Zhu and David Mumford. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, 2(4):259–362, 2006.