A Visual Grammar for Face Detection

Augusto Meléndez, Luis Enrique Sucar, and Eduardo F. Morales

Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro No. 1, Tonantzintla, Puebla, Mexico
{amelendez,esucar,emorales}@inaoep.mx

Abstract. Several methods have been developed for face detection with certain success, however these tend to fail under difficult conditions such as partial occlusions and changes in orientation and illumination. We propose a novel technique for face detection based on a visual grammar. We define a symbol relational grammar for faces, representing the visual elements of a face and their spatial relations. This grammar is transformed into a Bayesian network (BN) structure and its parameters are obtained from data, i.e., positive and negative examples of faces. Then the BN is used for face detection via probabilistic inference, using as evidence a set of weak detectors for different face components. We evaluated our method on a set of sample images of faces under difficult conditions, and contrasted it with a simplified model without relations, and the AdaBoost face detector. The results show a significant improvement when using our method.

1 Introduction

Face recognition has been a problem of interest for many years. Several methods for face detection have been developed with certain success, however these tend to fail under difficult conditions such as partial occlusions and changes in orientation and illumination. Under these conditions, it is difficult to distinguish a face in an image. Although the full face is not distinguishable, we can locate some of its elements. Thus, an alternative is to recognize the *basic elements* in a face, and their relationships, and combine this information to build a more robust face detector.

A *visual grammar* can represent the hierarchical decomposition of a face, as well as the spatial and functional relations between its basic elements. Once we represent a face based on a visual grammar, this can be used to *parse* an image and detect the face or faces in it; even if some elements are missing. Although there is some recent work on the development of visual grammars for object recognition [3,11], these are not focused on face recognition. We propose an alternative representation and implementation of a visual grammar, that results in a simpler and more efficient face recognition system.

We propose a novel technique for face detection based on a visual grammar. For this we first define a symbol relational grammar [2] for faces, representing the visual elements of a face and their spatial relations. This grammar is then

A. Kuri-Morales and G.R. Simari (Eds.): IBERAMIA 2010, LNAI 6433, pp. 493–502, 2010.

transformed to a Bayesian network representation. The structure of the Bayesian network is derived from the grammar, and its parameters are obtained from data, i.e., from positive and negative examples of faces. Then the Bayesian network is used for face detection via probabilistic inference, using as evidence a set of weak detectors for different face components.

We evaluated our method on a set of sample images of faces under difficult conditions, and contrasted it with a simplified model without the spatial relations, and also with the AdaBoost face detector. The results show a significant advantage for our visual face grammar in terms of recognition rates, against several variants of the Viola and Jones face detector [7]. The results without including the spatial relations are also inferior, demonstrating the advantage of using a relational grammar.

2 Related Work

2.1 Visual Grammars

Visual grammars entered a hibernation state for some years due to difficulties that remain challenging even today: (i) there is a enormous amount of visual knowledge about the real world scenes that has to be represented in the computer in order to make robust inference, (ii) the computational complexity is huge, and (iii) one cannot reliably compute the symbols from raw images. Recent progress on machine learning, statistical models and more robust visual detectors; as well as the important and steady increase in available computational power; have make possible the rebirth of visual grammars, with several interesting approaches in recent years.

One of the most important contributions is presented by Song-Chun and Munford [11]. In this work they defined a *Grammar of Images* that is represented using an *And-Or Graph*. They propose a common framework for visual knowledge representation and object categorization. Hierarchic and structural composition is the key concept behind these grammars. They discuss three case studies where the principal objective is to represent the structure of the image.

Tian-Fu, Gui-Song and Sng-Chun [8] present a compositional boosting algorithm for detecting and recognizing 17 common image structures in low-middle level vision tasks. This algorithm has two steps: *bottom-up proposal* and *top-down validation*. The bottom-up step includes two types of boosting methods. First, generate hypotheses, often in low resolutions, for some part of the image using AdaBoost, through a sequence of image features. Then propose instances of an image, often in high resolution, by binding existing children nodes of the image. A top-down process validates the bottom-up hypotheses following the Bayesian posterior probabilities.

An attribute graph grammar is defined by Feng Hang and Song-Chu [3]. They used this grammar for image parsing on scenes with man-made objects, such as buildings, kitchens and living rooms. Six rules or productions are defined for the grammar, each production rule not only expands a node into its components, but also includes a number of equations that constrain the attributes of a parent node and those of its children. The grammar rules are used recursively to produce a large number of objects and patterns in images and thus the whole graph grammar is a type of generative model. The inference algorithm integrates bottom-up detection which activates top-down prediction using the grammar rules.

Previous approaches tend to define very complex models as they intend to represent many classes of objects, and in general this results in complex and inefficient object recognition systems. We propose a more restricted model focused on face recognition, with two important differences to previous models: (i) it is based on a symbol-relation grammar, which provides an adequate formalism for visual grammars by incorporating spatial relations, and (ii) the visual grammar is transformed into a Bayesian network representation that allows to incorporate well known and efficient methods for parameter learning and recognition based on probabilistic inference. Although currently restricted to face recognition, the model can be generalized to represent and recognize other classes of objects.

2.2 Face Detection

There are many approaches for face detection and recognition which we can not cover here in detail due to space limitations (see [10] for a recent survey). Most techniques tend to fail if there are partial occlusions or important variations in illumination and orientation. Most recent approaches for face detection are based in the work of Viola and Jones [7], which combines simple Haar features with a cascade classifier based on the AdaBoost algorithm to detect frontal faces. This results in a very efficient detector with very good results on standard test sets, for example they report 93.9% of correct detections in set CMU+MIT dataset. Note that this dataset has in general frontal images without occlusions. Although superior to previous approaches, these methods are sensitive to face orientation, illumination changes and partial occlusions; under these conditions there is an important degradation in their performance. There are other approaches that consider probabilistic graphical models for face detection [5,1], however these are not based on a visual grammar.

3 A Face Grammar

A common approach to the description of visual languages makes use of formal grammars and rewriting mechanisms able to generate visual sentences. We are using the formalism of *symbol–relation grammars* to define our face grammar.

3.1 Symbol–Relation Grammars

Symbol relation grammars [2] are a syntactic formalism to describe visual languages, where each sentence in a language is represented as a set of visual objects and a set of relationships among those objects. A *Symbol-Relation Grammar* is a 6-tuple

$$G = (V_N, V_T, V_R, S, P, R)$$

where:

- V_N is a finite set of nonterminal symbols.
- V_T is a finite set of terminal symbols.
- V_R is a finite set of (*relation symbols*).
- $S \in V_N$ is the starting symbol.
- P is a finite set of labelled rewriting rules, called *s-item productions* of the form:

  $$l : Y^0 \rightarrow< M, R >$$

  where:
  1. l is an integer uniquely labeling the s-production.
  2. $< M, R >$ is a sentence on V_R and $V_R \cup V_T$
     - M is a set of s-tems (v, i), $v \in V_T$ and i is a natural number. In a simple way each s-item is writing v_i.
     - R is a set of r-items of the form $r(X_i, Y_j), X_i, Y_j \in M, r \in V_R$ this indicate a relationship r between X_i and Y_j.
  3. $Y \in V_N, Y^0 \notin M$
- R is a finite set of rewriting rules, called *r-items productions*, of the form

  $$s(Y^0, X^1) \rightarrow [l]Qos(X^1, Y^0) \rightarrow [l]Q$$

  where:
  1. $s \in V_R$.
  2. l is the label of an s-production $Y^0 \rightarrow< M, R >$
  3. $X \in V_N \cup V_T$ and $X^1 \notin M$
  4. $Q \neq \phi$ is a finite set of r-items of the form $r(Z, X^1)$ o $r(X^1, Z)$, $Z \in M$

3.2 Relational Face Grammar

Based on the previous formalism we define a relational face grammar. This initial face grammar only considers frontal views and the basic elements of a face; although it could be easily extended to consider other elements (like hair) and other views (like lateral or back). This face grammar includes several spatial relationships which are important because they represent the spatial configuration of the elements of a face, and help to decrease the number of false positives, as *false* basic elements in general do not comply with the desired configuration.

The main elements of this face grammar (FG) are defined as follows:

$$FG = (HEAD, \{eyes, nose, mouth, face\}, \{above, inside\}, HEAD, P)$$

where:

- $HEAD$ is a nonterminal symbol that represents the *complete* face.
- $\{eyes, nose, mouth, face\}$ are the terminal symbol that represent the corresponding elements.
- $\{above, inside\}$ are relational symbols that represent this spatial relations.
- $HEAD$ is the starting symbol

- P are the s-item productions rules, defined as:

  1: $HEAD^0 \rightarrow< HEAD, \phi >$
  2: $HEAD^0 \rightarrow< eyes, \phi >$
  3: $HEAD^0 \rightarrow< nose, \phi >$
  4: $HEAD^0 \rightarrow< mouth, \phi >$
  5: $HEAD^0 \rightarrow< face, \phi >$
  6: $HEAD^0 \rightarrow< \{eyes, mouth\}, above\{eyes, mouth\} >$
  7: $HEAD^0 \rightarrow< \{nose, mouth\}, above\{nose, mouth\} >$
  8: $HEAD^0 \rightarrow< \{eyes, face\}, inside\{eyes, face\} >$
  9: $HEAD^0 \rightarrow< \{nose, face\}, inside\{nose, face\} >$
  10: $HEAD^0 \rightarrow< \{mouth, face\}, inside\{mouth, face\} >$

For this initial face grammar we do not include r-item productions.

In principle, this grammar can be used to parse an image and detect a "HEAD". However, the detectors we used to find the terminal elements are not reliable, and in general there is a lot of uncertainty in detecting the terminal symbols and relations in images. Thus, a representation that can cope with uncertainty is required, such as Bayesian networks [6]. So we transform the relational face grammar to a BN representation.

3.3 Bayesian Networks

Bayesian networks [6] (BNs) are graphical structures that allow for a compact representation of a multi-variable probability distribution. A BN is a directed acyclic graph, $G = (V, E)$, in which the nodes (V) correspond to random variables, and the arcs (E) represent probabilistic dependencies between variables. Associated to each variable their is a conditional probability table (CPT) that contains the probability for each value of the variable given the combinations of values of its parents in the graph (assuming discrete variables). There are several algorithms for learning the structure and parameters of a BN from data, and for estimating the probability of certain variable given a subset of instantiated variables (probabilistic inference) [6,4].

For transforming the face grammar to a BN we make the following considerations:

- Each symbol (terminal or nonterminal) is represented as a binary variable (node).
- Each relation is represented as a binary variable (node).
- Productions are represented as a set of dependency relations (arcs) in the BN. An arc is incorporated between the nodes representing a nonterminal symbol (HEAD) to the terminal symbols. For relations, an arc is added from each of nodes of the symbols in the relation, to the node that represents the relation.

Additionally, we incorporate *virtual* nodes that represent the information we obtain from the detectors for the terminal symbols. These encode the uncertainty inherent to all the detectors as a CPT, and are incorporated to the model by

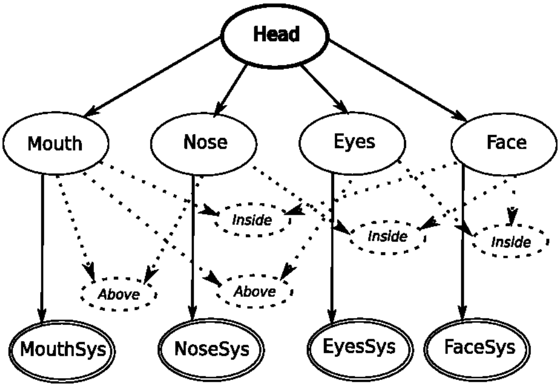

Fig. 1. Bayesian network that represents the face grammar. Simple ovals represent terminal and nonterminal elements, dashed ovals represent the relational elements, and double line ovals represent the detector variables.

adding an arc form the *symbol* node to the *system* node. For instance, we have a node that represents the *nose* symbol, with an arc to a node that represents the results of *nose* detector; which could be a "real nose" or a false detection.

In Figure 1 we depict the Bayesian network that we obtain to represent the face grammar based on the previous considerations. The root node represents the nonterminal symbol *HEAD*. In the next level, there are the nodes that represent the terminal elements of the face; and the nodes that represent the spatial relationship between these elements. In the last level, nodes represent the detectors for each element

3.3.1 Parameter Learning

Once the structure of the BN for the face grammar is defined, we have to obtain the corresponding parameters, i.e. the CPTs for each variable given it parents. We learn these probabilities from data using standard parameter learning for BNs [4].

For this we consider a set of sample images with positive and negative examples of faces. Positive examples are manually labeled indicating the terminal symbols in each image. We then apply the detectors for each terminal element, and count all correct and incorrect detections. Based on these statistics, we obtain the CPTs using a maximum likelihood estimator. Details of this process are given in the experiments.

The parameters for the relational elements could be estimated subjectively, as in general we expect that these relations are always satisfied in a human face. However, there could be special situations in which some relations are not satisfied, for instance if one of the elements is occluded or the face is inclined nearly 90 degrees, so we assigned a probability close to one for each relation being true (these could also be estimated from data but these will require additional effort in the training stage).

3.3.2 Recognition

To recognize a face we have to do a process analogous to parsing for each image using the visual grammar represented as a BN. For this, the terminal elements detected in the image and their spatial relations are instantiated in the BN, and via probabilistic inference we obtain the posterior probabilities of the nonterminal symbol, in this case $HEAD$.

Given that the detectors we use for the terminal symbols produce many false positives, there could be many *candidates* elements to instantiate the BN. In principle, we need to test all possible combinations, which might results in a combinatorial explosion. To avoid this problem, we implement this search process in an *optimal* order, by first testing the more reliable detectors (those that produce less false positives) and then the less reliable ones. Once certain instantiation produces a *high* probability of correct detection (above certain threshold), the search process is terminated; so in this way not all combinations have to be tested. This strategy gave good results in practice, as we will see in the next section.

Additionally we implemented a preprocessing stage to reduce the number of terminal elements detected in an image, as several detections tend to occur that correspond to the same element. If several terminal elements overlap, these were grouped into a single element (depending on the distance between the elements detected).

4 Experimental Results

4.1 Test Images

Although there are several standard face databases, in general they have images of faces in *good* conditions, without occlusions, rotations and other artifacts that make difficult recognition. As our method is focused for these situations, the existing databases are not a good testbed, so we build our own database. For these we downloaded images form Internet that contained faces in a great variety of conditions, most of them difficult for standard detection methods. Some of the images used in the experiments are depicted in Figure 2.

4.2 Terminal Element Detection

For detecting the terminal elements defined in our face grammar, we used detectors based on AdaBoost for faces, nose, mouth and eyes as implemented in the

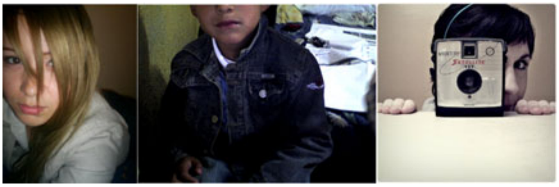

Fig. 2. Examples of face images used in the experiments

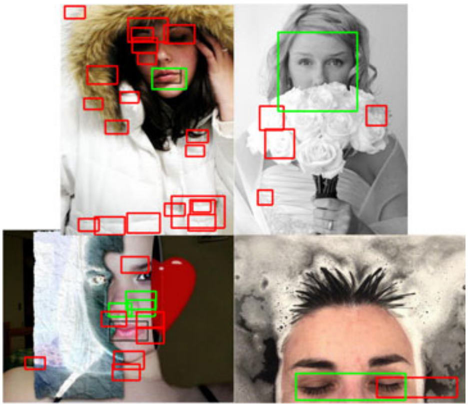

Fig. 3. Examples of the results obtained with different detectors. Green rectangles indicate true positives and red rectangles indicate false positives (best seen in color).

OpenCV library [9]. For the test images considered, these detectors give in general not very good results, tending to produce a large number of false positives, in particular the face, nose and mouth detectors; and also false negatives, specially the eyes detector. Some examples of detections are shown in Figure 3, where we observe a large number of false positives. After applying the preprocessing stage mentioned before to group detections, some elements are eliminated, but there are still several incorrect detections.

4.3 Training the Model

For training we used 200 images, 100 with faces and 100 without faces. For the images with faces with manually labeled the different terminal elements that were visible in the images. Based on these images we obtain the parameters for the BN.

As mentioned before, the CPTs for the relation nodes were estimated subjectively. Different probability values were considered, from 0.6 to 0.9 when the spatial-relation was *true*, and from 0.4 to 0.1 when the spatial-relation was *false*. We selected the values that produced the best results, although the differences are not significant.

4.4 Results

We applied the face grammar BN to a set of images, different from those used from training, 30 positive and 30 negative examples. For each test image, the BN model is applied several times, using the search process described before. The detected terminal elements are instantiated, and then we instantiate the relational nodes. To determinate if the elements detected satisfy the relations that we define, we consider the centroids of the elements detected. Once the

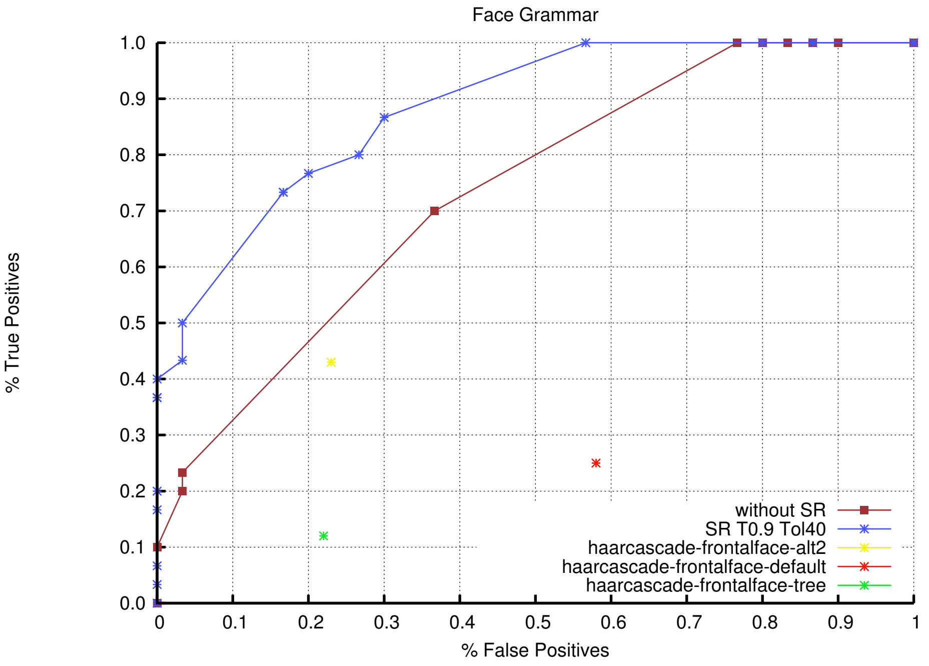

Fig. 4. The graph show the detection rate in terms of true positives vs. false positives varying the decision threshold for the proposed method. We compare the visual grammar with the full model (blue crosses) and the visual grammar with a partial model without relations (red rectangles). We also compare the proposed method against three variants of the Viola and Jones face detector with fixed thresholds (dots).

terminal and relation nodes are assigned, we estimate the probability of *HEAD* using probability propagation.

To evaluate the model we varied the decision thresholds (between 0.5 and 0.85) and compared the true positives vs. the false positives. In Figure 4 we summarize the results. We compared the full model against a partial model that does not include the relations. We observe a significant improvement when spatial relations are incorporated, which supports the choice of a symbol relation grammar. We also compared our method against three variants of the Viola and Jones face detector as implemented in OpenCV [9]. For these other methods we can not control the threshold, so we only show a point result. For the type of images considered, our method clearly outperforms these face detectors.

5 Conclusions

We have developed a novel method for face detection based on a symbol relation grammar for faces. We defined a face grammar that was then transformed to a Bayesian network, whose parameters are obtained from data. We applied this model for face detection under difficult conditions showing very good results compared against other face detectors.

Although the current grammar is restricted, we consider that this could be extended to provide a more complete description of a head from different viewpoints, and also for representing other classes of objects. In the future we want to explore learning the grammar from examples.

References

1. Candido, J., Marengoni, M.: Combining information in a bayesian network for face detection. Brazilian Journal of Prob. Stat. 23, 179–195 (2009)
2. Ferrucci, F., Pacini, G., Satta, G., Sessa, M., Tortora, G., Tucci, M., Vitiello, G.: Symbol-relation grammars: A formalism for graphical languages. Information and Computation 131, 1–46 (1996)
3. Han, F., Zhu, S.-C.: Bottom-up/top-down image parsing by attribute graph grammar. In: IEEE International Conference on Computer Vision, vol. 2, pp. 1778–1785 (2005)
4. Neapolitan, R.E.: Learning Bayesian Networks. Pearson, Prentice Hall, Chicago, Illiniois (2004)
5. Nefian, A.V.: A hidden Markov model based approach for face detection and recognition. PhD dissertation, Georgia Institute of Technology, Georgia, USA (1999)
6. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Francisco (1988)
7. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57, 137–154 (2004)
8. Wu, T.-F., Xia, G.-S., Zhu, S.-C.: Compositional boosting for computing hierarchical image structures. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
9. Yu, Q., Cheng, H.H., Cheng, W.W., Zhou, X.: Ch opencv for interactive open architecture computer vision. Adv. Eng. Softw. 35(8-9), 527–536 (2004)
10. Zhao, W.Y., Chellappa, R., Philips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Survey, 399–458 (2003)
11. Zhu, S.-C., Mumford, D.: A stochastic grammar of images. Found. Trends. Comput. Graph. Vis. 2(4), 259–362 (2006)