Recognizing Visual Categories with Symbol-Relational Grammars and Bayesian Networks

Elías Ruiz and L. Enrique Sucar

Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro 1, Tonantzintla, Puebla, México
{elias_ruiz,esucar}@inaoep.mx

Abstract. A novel proposal for a compositional model for object recognition is presented. The proposed method is based on visual grammars and Bayesian networks. An object is modeled as a hierarchy of features and spatial relationships. The grammar is learned automatically from examples. This representation is automatically transformed into a Bayesian network. Thus, recognition is based on probabilistic inference in the Bayesian network representation. Preliminary results in recognition of natural objects are presented. The main contribution of this work is a general methodology for building object recognition systems which combines the expressivity of a grammar with the robustness of probabilistic inference.

1 Introduction

Most current object recognition systems are centered in recognizing certain type of objects, and do not consider their structure. This implies several limitations: (i) the systems are difficult to generalize to any type of object, (ii) they are not robust to noise and occlusions, (iii) the model is difficult to interpret.

This paper proposes a model that achieves a compositional representation of a visual object in order to perform object recognition tasks, based on a visual grammar [4] and Bayesian networks (BNs) [8]. Thus, we propose the incorporation of a visual grammar in order to develop an understandable compositional model so that from basic elements (obtained by a patch-based approach) it will construct more complex forms by certain rules of composition defined in the grammar, in order to achieve object recognition in a limited context (e.g. images of natural objects). In addition, a model expressed as a symbolic grammar provides a transparent and understandable representation.

A Symbol-relation grammar (*SR grammar*) $\mathcal{G}_i$ is learned for each object class c_i, and then transformed automatically to a BN which incorporates the symbols and relations as nodes, and the arcs represent the structure derived from the grammar rules. Intermediate nodes in this BN structure are hidden, so we learn the parameters of the model using the Expectation-Maximization algorithm (EM). Once the structure and parameters of the BN are obtained, it can be used for recognizing a class of object using probabilistic inference.

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 540–547, 2014.

We test our model in a pair of classes from Caltech-256 [5] and the ETH-80 dataset [6] obtaining competitive results in terms of precision and recall; but with a significant reduction in the training and inference times. Also, fewer training examples are required to achieve a competitive performance.

Next we present a brief review of alternative hierarchical/compositional approaches for object recognition and contrast them with our approach. Then we describe in detail the proposed model, including the model building and recognition methods. We present experimental results in learning visual grammars for several object classes, and then using these for recognition. We conclude with a summary and directions for future work.

2 Related Work

There are several works using a hierarchical approach for object recognition based on visual grammars [1,3,10,9]. In these studies, there is a clear consensus in the usage of a certain kind of grammar to represent compositionally the terminal elements (lexicon). However, they differ in what terminal elements to use and how to handle the uncertainty in order to perform object recognition.

The proposed model differs in several aspects from previous work:

- It is based on a SR-grammar which incorporates spatial relationships.
- The grammar is induced automatically from example images.
- The terminal elements (*visual lexicon*) are patch-based and learned automatically, so they can be used for different visual objects.
- The grammar is automatically transformed to a BN which provides a robust and efficient technique for object recognition.

We consider the use of SR grammars because of the convenience of putting the relationships in predicate logic, which is natural in this kind of grammar. Also, it is desirable that the grammar is automatically learned from examples, for greater generality; the grammar is independent of the lexicon definition used. Finally, the transformation to a model that considers uncertainty must also be automatic.

We use BNs to represent the information given by the grammar incorporating uncertainty. Other studies use different schemas or even probabilistic grammars. Bayesian networks have several advantages, such as preserving the structure given by the grammar and providing efficient algorithms for parameter learning and probabilistic inference.

3 Object Recognition Model

The proposed method compromises two phases: (i) model construction and transformation to a BN (Fig. 1); and (ii) image pre-processing and object recognition using probabilistic inference (Fig. 2). Next we describe each phase in detail.

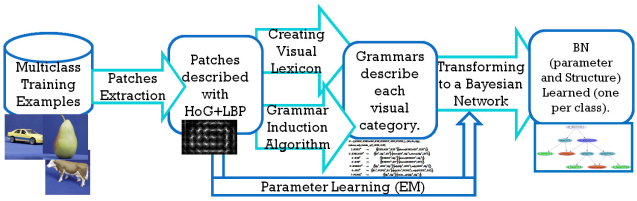

Fig. 1. Training phase. Starting with training images of several classes, we extract features using a grid in each image and describing each patch with HoG +LBP features. The lexicon is created by a clustering algorithm and several sets of rules are induced, one per each class. The grammar is learned from the terminal elements obtained by the lexicon. Finally the model is transformed to a BN, whose parameters are learned from examples. The structure is given by the grammar obtained in the previous stage.

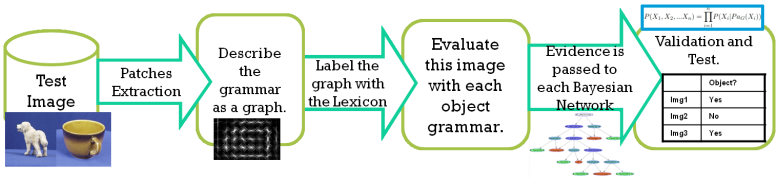

Fig. 2. Object recognition. The test image is described with patches and each patch is labeled with the visual lexicon. After that, the algorithm evaluates subsets of those terminal elements with similar structure to the grammar over the previously trained BN in order to do inference. At the end, we obtain results by probabilistic inference in the BN obtaining the probability of the presence of an object (for a specific class) in each image.

3.1 Model Construction

Features Extraction and Lexicon. Patch description is performed with a simple grid over each image and describing each region by using Local Binary Patterns (LBP) and Histogram of Gradient (HoG) descriptors [2]. This description is applied for all the classes that will be processed in the training phase. We use the *k-means* algorithm in order to label each patch to the centroid of the corresponding cluster. The number of clusters (k) is selected according to the number of classes; $k \geqslant 50$ when the number of the classes is greater than 10, otherwise k is fixed to 30. All the terminal elements constitute the *Visual lexicon*. Each patch can be related to another patch with six spatial relationships (Fig. 3). Note that the lexicon can be improved and this will not affect other layers of our model.

Learning the Grammar. For our model, we need a grammar that allows us to model the decomposition of a visual object into its parts and how they relate with other parts. SR-grammars, which are described in [4], provide this type of description.

Fig. 3. The six spatial relationships used in our grammar. *Above*, $Left$, and *Overlapped* (in four forms: Left, Above, -45° and +45°). Each patch is labeled with an element of the lexicon. Relationships with no adjacency between patches are not considered. We use these relationships because they preserve the coherence when we subsume two regions in another new one. The new non-terminal elements generated preserve all the relationships from its children with other elements, and loose their internal relationships.

The next step is to generate the rules that make up the grammar. Using the training images, we search the most common relationships between visual words obtained throughout the multiclass train set. Such relationships become candidate rules to build the grammar. This is an iterative process where the rules are subsumed and converted to a new non-terminal element of the grammar ($w_{c_i} \in V_N$), where V_N is the set of non-terminal elements of the grammar. If we repeat this process, the starting symbol of the grammar represents the object that we want to recognize. Each w_{c_i} is obtained by the formula:

$$w_{c_i} = \underset{T_a,T_b,r}{\arg\max} \left(F_{c_i}\left(R_{r,T_a,T_b}\right) + \max_{c_x \in C,\, c_x \neq c_i} \left(d\left(F_{c_x}\left(R_{r,T_a,T_b}\right),\, F_{c_i}\left(R_{r,T_a,T_b}\right)\right)\right)\right) \quad (1)$$

where w_{c_i} is the new non terminal element generated, F_{c_i} is defined as the frequency of the rule R over the c_i training dataset (in how many images the rule appears). d is the euclidean distance and C is all the training set. R_{r,T_a,T_b} is a rule holded by the spatial relationship r, and $T_a, T_b \in V_T \bigcup V_N$ (T_a,T_b can be terminal or non-terminal elements of the grammar). The stop criterion is a frequency threshold for the rule (the rule needs to be found in at least $F_{c_i} > n$ images of the training set. n is usually fixed to a half of the training set). This criterion avoids generating a highly complex grammar. As an example, the rule $Above(P_1, P_2)$ is subsumed into a new non-terminal element named NT_1. The rule obtained is: $1 : NT_1^0 \rightarrow < \{P_1^2, P_2^2\}, \{Above(P_1^2, P_2^2)\} >$. With this rule generation method, circular productions are avoided (the BN generated would have infinite structure).

Transformation of the Grammar. We transform the grammar into a BN, using the following procedure. For every production rule, $Y^0 \rightarrow \langle \mathbf{M}, \mathbf{R} \rangle$, we produce the node Y^0 in the grammar and connect this node with all $x \in \mathbf{M}$. For every relationship $r(a,\, b) \in \mathbf{R}$ we produce the node r connected with its parents $a, b \in \mathbf{M}$. Furthermore, for every terminal node $a \in V_T$ we create a leaf node a_E with parent a representing the evidence and the associated CPT its uncertainty.

The transformation procedure is illustrated with the following example. If we consider the next grammar:

$$\mathcal{G} = (V_N,\, V_T,\, V_R,\, S,\, P,\, R)\,;\, V_N = \{NTR6C7C27,\, NTR5C33C7\}\,;$$
$$V_T = \{TermC7,\, TermC27\}\,;\, V_R = \{LAOverlaps,\, LBOverlapps\}\,;\, S = NTR5C33C7;$$

where the production rules are defined by P :

1. $NTR5C33C7 \rightarrow \langle \{NTR6C7C27,\ TermC7\},\ \{LAOverlaps(NTR6C7C27,\ TermC7)\} \rangle$
2. $NTR6C7C27 \rightarrow \langle \{TermC7,\ TermC27\},\ \{LBOverlapps\,(TermC7,\ TermC27)\} \rangle$

The algorithm generates the structure of a Bayesian network illustrated in Fig. 4a.

OR Rules. In some cases, there are several candidate rules to be considered in our model. The choice of only one can represent a strict restriction (because other positive structures would be missed). Thus we have included the OR production in our grammar. The OR rule produces different ways for an object definition. A simple example with two OR-rules is illustrated in Fig. 4b.

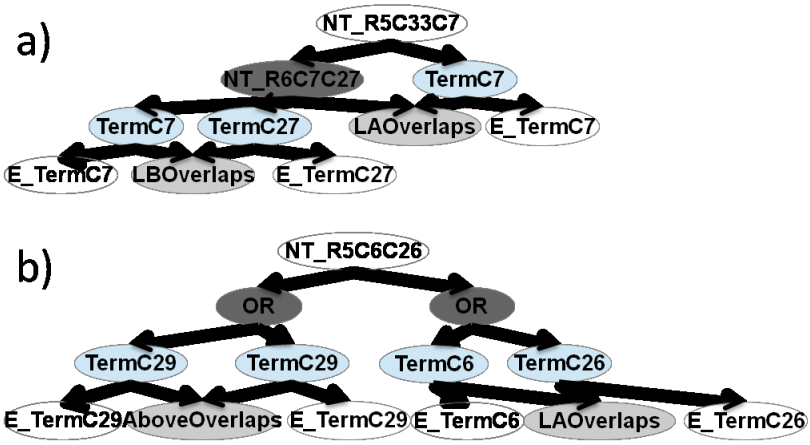

Fig. 4. a) Bayesian Network generated by the example grammar. Evidence is given only to leaf nodes. Leaf nodes with two parents (in light-gray) represent relationship nodes. Leaf nodes with only one parent (in white) represent evidence in terminal elements. b) BN representing an Or-grammar with two rules.

Parameter Learning. Once the BN is obtained, its parameters are learned initially with weights obtained from each rule. However we also have included the Expectation Maximization (EM) algorithm applied over the intermediate nodes in the BN in order to prevent overfitting with a validation set.

3.2 Object Recognition

For object recognition, an image is initially described with the visual lexicon. Finding a valid configuration means to discover a relationship that has a match with the grammar rule as represented in the BN. This match is converted to evidence in the BN. If the complete grammar is found, there is a high probability that the object learned with the grammar appears in the image.

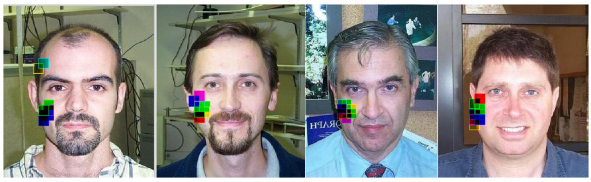

Fig. 5. Regions detected for a face example. The colored boxes are the terminal elements detected and provided as evidence to the BN. Different colors are used for a better differentiation only. In this case our model did not learn the whole face; instead the model learned an object part. Thus, the grammar helps to detect parts of the object. Best seen in color.

4 Results

To evaluate experimentally the proposed model, we consider the ETH-80 database [6].[1] We also tested the model using two classes of the Caltech-256 database [5]. The obtained model is tested using a different set of test images. Recognition is evaluated based on the posterior probability given by probability propagation in the BN. Examples from other categories are considered as negatives for this evaluation.

Examples of detected objects for the faces category are illustrated in Fig. 5; as we can see, the method discovered certain regions (terminal elements) that are related to the specific category.

Recognition results are evaluated in terms of: $Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$, $Precision = \frac{TP}{TP+FP}$, and $Recall = \frac{TP}{TP+FN}$. TP is the true positive, TN the true negative, FP the false positive and FN the false negative rate in each experiment. The model obtained for each class of object was evaluated with a set of test images that include positive and negative examples. The negative examples were obtained from the other classes for ETH-80 and clutter dataset for Caltech-256. The results are summarized in the Tables 1 and 2. Although these results are in general not superior to other methods in the state of the art [7], we consider that they are promising as the proposed method provides a general framework that still needs to be optimized. Nevertheless, there are some important advantages of the proposed approach: It can be trained with a few examples (20 per class), the training and the inference time is fast (around 0.2 seconds per image in inference for ETH database). The state of the art methods require several hours for training whereas we require minutes for the entire dataset. We evaluated the F-measure for ETH database (Fig. 6) and we can see how the model stabilizes its results after 80 examples. Each line represents one of the eight classes.

[1] Note that we are more interested if the grammar can recover positive test examples, because we want to know what the most invariant structure in each category is.

Table 1. Recognition results of the model in ETH database. Positive examples are obtained from the specified class and negative examples are obtained from the other classes. The training and validation sets from each category are of 100 elements (80 for training and 20 for validation). Time is for inference only. The models can be trained in less than one hour (all of them) without optimization of the code.

Class	Accuracy	Precision	Recall	Time	Num Examples
apple	82.9	82.69	83.22	2min	310
car	82.9	80.5	86.77	2min	310
cow	72.58	68.04	85.16	2min	310
cup	90.32	89.06	91.93	3min	310
dog	73.06	75.81	67.74	2min	310
horse	77.58	71.42	91.93	2min	310
pear	85.96	87.79	83.54	3min	310
tomato	87.58	91.75	82.58	2min	310

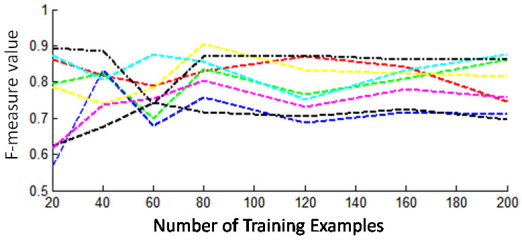

Fig. 6. Each line represents F-measure value for each class with different number of training images. The model is stable with more than 80 examples. We think that the fall for some classes around 60 examples was because the grammar was focusing on one class and loosing performance in the others.

Table 2. Recognition results of the model in two categories for Caltech 256 database (faces and motorbikes). Time is for inference. We do not expect a high accuracy in the clutter category. However, the model learned the "structure" for the selected categories.

Class	Accuracy	Precision	Recall	Time	Num Examples
motorbikes	89.2	89.4	89.1	3.5min	330
clutter	80.5	90.0	68.5		330
faces	90.7	79.8	97.7	4.5min	375
clutter	56.7	67.5	65.5		375

5 Conclusions and Future Work

A novel and general model for object recognition based on SR-grammars and BNs was described. We have performed experiments with natural object classes with promising results. The models can be learned with a few training examples, and the method is very efficient in terms of training and inference times.

The main contribution of this work is proposing a general methodology for developing object recognition systems that combines the richness and expressivity of formal grammars and the robustness and efficiency of Bayesian networks. We consider that this work contributes to the final goal of developing more general vision systems, analogous to those developed for voice and language.

There are several avenues for future research: (i) Improve the or-rules including a grammar book in order to reutilize rules discovered in other datasets (e.g. part-objects) in order to increase the coverage over positive examples. (ii) Improve the lexicon by using a flexible scheme including max pooling techniques and multilabels for terminal elements. (iii) Evaluate the model with other classes of objects or environments.

References

1. Chang, L., Jin, Y., Zhang, W., Borenstein, E., Geman, S.: Context, computation, and optimal roc performance in hierarchical models. IJCV 93(2), 117–140 (2011)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, vol. 1, pp. 886–893. IEEE Computer Society (2005)
3. Felzenszwalb, P.F.: Object detection grammars. In: ICCV Workshops, p. 691. IEEE (2011)
4. Ferrucci, F., Pacini, G., et al.: Symbol-relation grammars: a formalism for graphical languages. Inf. Comput. 131(1), 1–46 (1996)
5. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology (2007)
6. Leibe, B., Schiele, B.: Analyzing appearance and contour based methods for object categorization. In: CVPR, pp. 409–415. IEEE (2003)
7. Linde, O., Lindeberg, T.: Composed complex-cue histograms: An investigation of the information content in receptive field based image descriptors for object recognition. CVIU 116(4), 538–560 (2012)
8. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo (1988)
9. Ruiz, E., Sucar, L.E.: An object recognition model based on visual grammars and bayesian networks. In: Klette, R., Rivera, M., Satoh, S. (eds.) PSIVT 2013. LNCS, vol. 8333, pp. 349–359. Springer, Heidelberg (2014)
10. Zhu, S.C., Mumford, D.: A stochastic grammar of images. Foundations and Trends in Computer Graphics and Vision 2(4), 259–362 (2006)