Learning Bidimensional Context-Dependent Models Using a Context-Sensitive Language

Miguel Sainz(*) and Alberto Sanfeliu(**)
(*) Instituto de Cibernética (**)Instituto de Robotica e Informatica Industrial
Universidad Politécnica de Catalunya
Diagonal 647, 08028 Barcelona
(*)sainz@ic.upc.es (**)sanfeliu@ic.upc.es

Abstract

Automatic generation of models from a set of positive and negative samples and a-priori knowledge (if available) is a crucial issue for pattern recognition applications. Grammatical inference can play an important role in this issue since it can be used to generate the set of model classes, where each class consists on the rules to generate the models. In this paper we present the process of learning context dependent bidimensional objects from outdoors images as context sensitive languages. We show how the process is conceived to overcome the problem of generalizing rules based on a set of samples which have small differences due to noisy pixels. The learned models can be used to identify objects in outdoors images irrespectively of their size and partial occlusions. Some results of the inference procedure are shown in the paper.

1. Introduction

Techniques for automatically acquiring shape models from sample objects are presently being researched. At present, a vision developer requires to select the appropriate shape representation, design the reference models using the chosen representation, introduce the information and program the application. This methodology is used in industrial applications, since there is not any other available. However, it is cumbersome and impractical when dealing with large set of reference models.

The recognition systems in the future must be capable of acquiring objects from samples with limited human assistance. There exist few approaches to automatically acquire generic models. Some of them are based on neural networks [6], appearance representation [8] and grammatical inference [3],[10]. In this paper we deal with grammatical inference methods to learn the grammar models from a set of positive and negative samples.

Grammatical inference (GI) methods have been basically developed for regular grammars (or finite state automata), but the potential descriptive power of these grammars is very restricted and seldom they have been used for learning complex models.

Computer vision models usually contain symmetries and structural relationships that are not describable by regular neither by context free grammars (languages); at least context sensitive grammars (languages) are required. Recently, Alquezar and Sanfeliu [1] presented a formalism to describe, recognize and learn a class of non-trivial context sensitive languages, denominated Augmented Regular Expressions (ARE). AREs augment the descriptive power of regular expressions by including a set of constraints that involve the number of instances of the operands of the star operations in each string of the language. The method for learning AREs consists of a regular grammatical inference step, aimed at obtaining a regular superset of the target language, followed by a constraint induction process, which reduces the extension of the inferred language transforming it into a context sensitive one.

In this paper, a new method is presented to learn bidimensional computer vision models from a set of positive and negative samples. Each model is represented by an pseudo-bidimensional ARE, where each row is represented by an ARE and the columns are all together represented by another ARE. The paper describes the learning process and the results of the application of this method to learn traffic signs.

2. Description of the bidimensional model

In previous work [10], we set some criteria of automatic model learning and recognition of

1015-4651/96 $5.00 © 1996 IEEE
Proceedings of ICPR '96

bidimensional objects in outdoor scenes, from true color images and through a two step process based on the **Active Grammatical Inference** methodology. In this work we define the bidimensional model and the necessary steps to learn it from a set of sample images. The output of the process is a two level context sensitive language which represent each model class.

The formal representation of a bidimensional model is:

Definition 2.1 A pseudo-bidimensional Augmented Regular Expression (or PSB-ARE) is a four-tupla (Σ_R, *V*, *T*, *L*), where Σ_R is the set of the row *ARE's* [1], *V* is the associated set of *star variables*, *T* is the associated *star tree*, and *L* is a set of independent linear relations $l_1 \dots l_{nc}$, each envolving the variables in *V*. If the set *L* is given by partitioning the set of star variables *V* into two subsets V^{ind}, V^{dep} of independent and dependent star variables, respectively, and expressing the latter as linear combinations of the former, for $I \leq i \leq nc$:

$$l_i \equiv v_i^{ind} = a'_{i1} \cdot v_1^{ind} + \dots + a'_{ij} \cdot v_{1j}^{ind} + \dots + a'_{i(ni)} \cdot v_{ni}^{ind} + a'_{i0}$$

where *ni* and *nc* are the number of independent and dependent star variables, respectively. *R*, *V* and *T* are described in detail in [1].

The definition of *V* restricts the allowed values for the star variables to natural numbers, $\forall\ k \in [1,ns]$: $v_k \in N$. Consequently, the set of linear relations *L* is only well-defined when the involved variables take natural numbers as values. This also implies that some of the star variables may be implicitly constrained to a smaller range inside the natural numbers (e.g. $v_k \geq z$, $z \in N$; v_k always odd; v_k always even; etc.). Moreover, the coefficients a_{ij} (or a'_{ij}) of the linear relations will always be rational numbers.

The PSB-ARE can be seen as a column ARE of the ARE's of each row. Fig. 1 shows the PSB-ARE of a traffic sign. We call this representation as pseudo-bidimensional ARE due that there are two levels of ARE. The first level it is the row level where each row of the model is represented by an ARE. The second level is the column level, where all the columns are represented only by one ARE which terminal symbols are the row ARE's. This type of representation is a nonsymmetric representation which can be automatically generated by a string grammatical inference method.

This representation describes the models irrespectively of the scale and position of the object in the scene. Moreover it permits to identify partially occluded objects as well as objects with distorsions. However the representation does not allow to represent models at different angles of rotation. The generation of the PSB-ARE models is based in a four step procedure which will be explained in the next section.

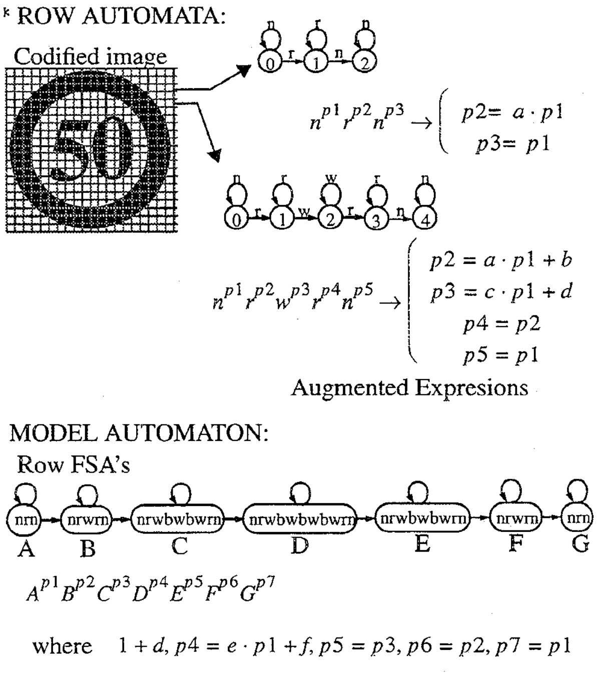

ROW AUTOMATA:
Codified image
50
n
r
n
0
1
2
$n^{p1} r^{p2} n^{p3} \rightarrow \begin{cases} p2 = a \cdot p1 \\ p3 = p1 \end{cases}$
n
r
w
r
n
0
1
2
3
4
$n^{p1} r^{p2} w^{p3} r^{p4} n^{p5} \rightarrow \begin{cases} p2 = a \cdot p1 + b \\ p3 = c \cdot p1 + d \\ p4 = p2 \\ p5 = p1 \end{cases}$
Augmented Expresions
MODEL AUTOMATON:
Row FSA's
nrn
nrwrn
nrwbwbwrn
nrwbwbwbwrn
nrwbwbwrn
nrwrn
nrn
A
B
C
D
E
F
G
$A^{p1} B^{p2} C^{p3} D^{p4} E^{p5} F^{p6} G^{p7}$
where $1 + d, p4 = e \cdot p1 + f, p5 = p3, p6 = p2, p7 = p1$
Pseudobidimensional Expresion (PSB-ARE)

Figure 1. Model Description

3. Context sensitive language description

3.1 Global description

The process of learning bidimensional computer vision models has to deal with model learning from a set of non ideal examples which are extracted from a set of images of diverse scenes. The goal of the learning process is to obtain the bidimensional model from the areas selected of each image (A_j^i, where *i* is the number of the image and *j* the number of the area in a image). Let $S^+ = \{A_1^{1+},\dots, A_j^{i+},\dots,A_p^{m+}\}$ the set of positive samples and $S^- = \{A_1^{1-},\dots, A_j^{i-},\dots,A_p^{q-}\}$ the set of negative samples.

As it has been explained in the previous section, each bidimensional model is represented by its rows and columns. A row is denoted by $row_k^{A_j^i}$ where *k* is the number of the row. The learning process is based in the generalization of the constructive rules of the rows and the columns from the set of positive and negative samples to obtain the bidimensional model. In order to learn a good set of rules for each model, the learning process can not be done in one step since usually the samples have noisy rows and columns. For this reason the process requires the following sequence of operations:

- Learn the set of basic row models (mod^t_{row} where t is the number of the row model) of the A_j^i as FSAs (finite state automata) from S^+ and S^-.
- Identify each $row_k^{A_j^i}$ as one of the mod^t_{row}.The result is a $mod_k^{A_j^i}$
- Learn the ARE model of every $row_k^{A_j^i}$ (as a row model) for each S^+.
- Learn the column model of the codified A_j^i.

The first step serves to guarantee that every mod^t_{row} matches the ideal structure of each class of row as a FSA. The noisy transitions are eliminated due that they show low probability. The second step is also used to eliminate noisy elements, since only model rows will be identified. The third step is used to learn the context sensitive language of each row. Finally, the last step serves to learn the context sensitive model of the columns.

3.2 FSA and Row ARE extraction

As it was described in the previous section, the first step is to learn the basic FSA models, mod^t_{row} , from S^+ and S^-. In this process some $row_k^{A_j^i}$ are selected from the chosen A_j^i. Negative rows are introduced to limit the generalization of the learned transitions.

There exists in the literature several methods to learn by induction FSA from a S^+ [3][7] and from S^+ and S^- [12]. We have used the AGI (Active Grammatical Inference) methodology [9][2] which permits to learn the FSA from a set of positive and negative samples in one or several cycles, without imposing any predefined induction rule.

Active Grammatical Inference is a methodology which allows to guide the learning process of a grammar by using the acquired knowledge and/or the demanded constraints imposed by the external knowledge. The whole process is conceived as a sequence of learning cycles, each one including a combination of neural and symbolic techniques, where the control of the next neural training is dynamically modified by the acquired information and/or by the imposed external information. See [9] for details of the methodology.

After obtaining all possible FSA that appear in the set of objects to modelize, the learning process of a specific model begins by obtaining the Augmented Regular Expression of each $row_k^{A_j^i}$. As explained in [1], an ARE can be obtained from a FSA and a set of string samples. The basic row model (mod^t_{row})of an A_j^i are obtained from only one of the images. The criteria to selected that image are first, the minimum noise amount and second, the maximum size to get a good detail resolution.

Once the A_j^i is selected, the first step is to identify every $row_k^{A_j^i}$ as one of the mod^t_{row} . To do so, we have to calculate the distance from each $row_k^{A_j^i}$ to every mod^t_{row}, and choose the one with the minimum distance (if it is below a predefined threshold). For this process we use a fast error correcting parser. The algorithm is an extended version of Viterbi's parser with error correction .

The second step is to pick some $row_k^{A_j^i}$ from different images to apply the ARE extraction algorithm. We select two $row_k^{A_j^i}$ for calculating the star variables of an ARE row. We apply the ARE algorithm [1] and the result is a linear system of the star variables:

$$\begin{bmatrix} a^k_{11} & \cdots & a^k_{1j} & \cdots & a^k_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a^k_{i1} & \cdots & a^k_{ij} & \cdots & a^k_{in} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a^k_{m1} & \cdots & a^k_{mj} & \cdots & a^k_{mn} \end{bmatrix} \cdot \begin{bmatrix} v^k_1 \\ \cdots \\ v^k_i \\ \cdots \\ v^k_n \end{bmatrix} = \begin{bmatrix} b^k_1 \\ \cdots \\ b^k_i \\ \cdots \\ b^k_m \end{bmatrix}$$

where a^k_{ij} are the coefficients of the linear combination of the star variables, v^k_n are the star variables and b^k_i are dependent coefficients. Since we are using only two $row_k^{A_j^i}$ for calculating the star variables n-m=1.

Initial Row string:

nnnnbbwnnnnrrrrrnnrrrrrrrrrrrrrrr rrrrrrrrrrrrnnnnnwrnn

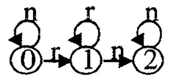
$\overset{10}{\text{n}}\quad\overset{36}{\text{r}}\quad\overset{10}{\text{n}}$

Corrected string: n r n

Second sample generated: nnnnnrrrrrrrrrrrrrrrrrrnnnnn

$\overset{5}{\text{n}}\quad\overset{18}{\text{r}}\quad\overset{5}{\text{n}}$

Best FSA:



ARE linear system equations:

$${}_np1\,{}_rp2\,{}_np3 \qquad \text{where} \qquad \begin{bmatrix} \frac{20}{9} & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} p1 \\ p2 \\ p3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Figure 2. Row ARE extraction sample

One of the sample string is the result of the Extended Viterbi error corrector parser. This is done to reduce the amount of noise in the samples. The other sample string is generated from the corrected one by dividing by two the number of times that each symbol appears and using the integer value as the new number of times.

3.3 Column extraction

The final step to finish the model learning is the column extraction. The aim of this extraction is to find the ARE of the row's ARE set of the A_j^i.

This structure will be an ARE expression where each language terminal will be a row FSA class. We have named this structure as PSB-ARE. It is generated from a sequential list of the FSA that belongs to the sample rows.

The first step is to get the FSA structure of the rows sequence. The system starts labelling each row with a number corresponding to the FSA type of that row. Then considers each label as a state on the PSB-FSA and then fusionates all the same type adjoining states into one and adds a transition from that state to himself. This extra transition is a self-loop of that state. This process is shown in Figure 2.

Initial string 00011111111112222222333344444222222111111111000

String skeleton 0-1-2-3-4-2-1-0

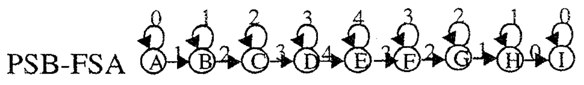


Figure 3. PSB-FSA extraction.

The second step is to calculate the star variables of the FSA to obtain the final PSB-ARE structure of the model. This is done in similar way than the row ARE extraction, using the PSB-FSA as the automaton and the row labels string as the sample to apply the ARE algorithm. The last step is to calculate the relation between the width and the height of the sample. This is done in order to allow the sample model generation from the width hypotheses generated by the recognition system.

4. Model pattern seeds

We represent each reference model by means of a context sensitive language, which have the language structure to generate the samples of a specific model. However the language does not incorporate the potential distortions at the low and high level (for example due to noise or partial occlusion of the reference model). In order to take into account these distortions, the parser or the matching process that uses this model description must compute a similarity measure [3], [4]. Since our language is context sensitive and at present there is not an efficient error correcting parser, we have developed a new strategy. The strategy consist on three steps : (1) look in the image for the candidate pattern seed transitions; (2) generate the best reference models in accordance with the first step; and (3) do the matching process using the Leveshtein [5] measure distance. With this strategy we overcome the problem of distortions and partial occlusions with a low computational processing time.

There are two problems associated to this matching strategy. The first one is how to generate a robust hypotheses to calculate the right size of the model sample (considering that there can be noise or/and partial occlusion of the object in the image analysed). And the second one is how to find the proper coordinates to superpose the sample model to the object image and do the distance measuring.

To solve these two inherent problems we define the candidate pattern seed transitions, which is a portion of the context sensitive language of the reference model and encompasses several rows. It is based on finding robust transitions in the image (robustness is required since that the image noise is very common in images).

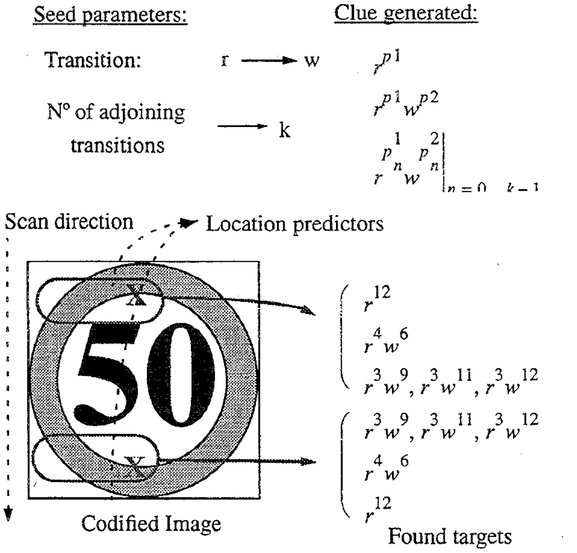


Figure 4. Seed transitions of an ARE candidate.

The selection process of a seed transition starts by manually choosing one or several symbol transitions. Then the system will select the mod_k^{Aj} of the model which have the manually selected symbol transition and accomplish the following two criteria:

The mod_k^{Aj} under (below) has to be different to the selected mod_k^{Aj}. This means that we are looking for a mod_k^{Aj} class change between two rows.

There has to be a minimum number of rows below (under) with the same mod_k^{Aj} that the one with the selected transition. We are looking for a portion of the model that has the same mod_k^{Aj} description.

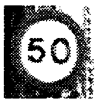
The information that the model keeps is the location of the selected mod_k^{Aj}.

In Figure 4. there is a simplified example of a seed transition and its location in the candidate image.

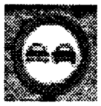
5. Results

In this section we will show some results of the intermediate steps of a bidimensional model learning.

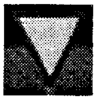
We have learned the context sensitive language of the three following traffic signs:



S1: Speed Limit



S2: No passing



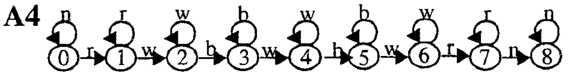
S3: Give the way

We have used 3 different images of each object to select the samples for the FSA extraction and we have find a set of eight predominant mod^{t}_{row}. The allowed terminal symbols are {r,b,w,n}(red, black, white and 'other' colors). The automata found are the following:

A0 -> n	A4 -> n-r-w-b-w-b-w-r-n
A1 -> n-r-n	A5 -> n-r-w-b-w-b-w-b-w-r-n
A2 -> n-r-w-r-n	A6 -> n-r-w-b-w-b-w-b-w-b-w-r-n
A3 -> n-r-w-b-w-r-n	A7 -> n-r-w-r-w-b-w-r-n

A8 -> n-r-w-r-w-r-w-b-w-b-w-r-n

Grafical representation of A4:



(All mod^{t}_{row} the have a similar lineal topological structure. See the graphical representation)

	S1		*S2*		*S3*	
FSA	N° Occurr.	Medium Error	N° Occurr.	Medium Error	N° Occurr.	Medium Error
A0	2	0.04%	2	0.00%	2	0.00%
A1	62	0.02%	57	0.48%	44	0.18%
A2	78	0.03%	88	1.73%	118	0.06%
A3	2	0.23%	0	---------	0	----------
A4	25	0.13%	1	0.45%	0	----------
A5	15	0.36%	1	8.56%	0	----------
A6	26	0.52%	0	---------	0	----------
A7	1	0.45%	32	2.66%	0	----------
A8	0	---------	19	4.84%	0	----------
TOTAL	211	0.13%	200	1.83%	164	0.09%

Table 1: FSA row assignation

The results of the FSA row assignation over a sample model are shown in Table 1.

6. Conclusions

In this paper we present a method to learn bidimensional context dependent models using a context sensitive language inference method image from a set of positive and negative samples. A model is conceived as a pseudo-bidimensional Augmented Regular Expression which consists on the rules that generate the rows and the columns. The method is general enough to be applicable to structured objects of scene images, allowing to impose restrictions and a priori knowledge. The method works although the objects have noisy pixels or rows in the learning set. Some results are shown on the outdoors scenes where the objective was to learn the traffic signs from a set of color images.

7. References

[1] R. Alquezar and A. Sanfeliu, "Augmented regular expressions: a formalism to describe, recognize and learn a class of context -sensitive languages", *Pattern Recognition* (In press) (1996).

[2] R. Alquezar and A. Sanfeliu, "An algebraic framework to represent finite-state machines in single-layer recurrent neural networks", *Neural Computation*, **7**, Sept.(1995).

[3] K.S. Fu, *Syntactic Pattern Recognition and Applications*, Prentice-Hall, New York, (1982).

[4] H.Bunke and A. Sanfeliu, *Syntatic and Structural Pattern Recognition: Theory and Applications*, World Scientific, (1990).

[5] V.I. Levensthein, "Binary codes capable of correcting deletions, insertions and reservals", *Sov. Phys. Dokl, 10 (8), 707-10*, Feb (1966).

[6] W. Lei and N. M. Nasrabadi, "Invariant object recognition on neural network of cascaded RCE nets", *Int. Journal of Pattern Recognition and Artificial Intelligence*, Vol. 7, No.4, pp 815-829, (1993).

[7] L. Miclet, "Grammatical inference," in *Syntatic and Structural Pattern Recognition: Theory and Applications*, H.Bunke and A.Sanfeliu, Eds., World Scientific, 1990.

[8] H. Murase and S.K. Nayar, "Visual learning and recognition of 3D objects from appearance", *International Journal of Computer Vision*, Vol. 14, No.1, pp 5-24, January, 1995.

[9] A. Sanfeliu and R. Alquezar, "Active Grammatical Inference: a new learning methodology", in *Shape and Structure in Pattern Recognition*, D. Dori and A. Bruckstein (eds.), World Scientific Pub., Singapore (1995).

[10] M. Sainz and A. Sanfeliu, "A first approach to learn the model of traffic signs using connectionist and syntactic methods", *Proceedings of the VI Simposium de Reconocimiento de Formas y An*alisis de *Im*agenes, Cordoba, 3-6 April (1995).

[11] W.A. Woods, "Transition networks grammars for natural language analysis", CACN 13, pp.591-606, 1970.

[12] J. Oncina and P. Garcia, "Identifying regular languages in polynomial tiem", in *Advances in Structural and Sysntactic Pattern recognition*, H. Bunke (ed.), World Scientific, Singapore, 1992, pp.99-108.