MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Modeling High-Order Dependencies in Local Appearance Models

David Guillamet *
Baback Moghaddam †
Jordi Vitria *

TR2003-94 June 2003

Abstract

We propose a novel local appearance modeling method for object detection and recognition in cluttered scenes. The approach is based on the joint distribution of local feature vectors at multiple salient points and their factorization with Independent Component Analysis (ICA). The resulting densities are simple multiplicative distributions modeled through adaptative Gaussian mixture models. This leads to computationally tractable joint probability densities which can model high-order dependencies. Our technique has been initially tested with natural and cluttered scenes with some degree of occlusions yielding promising results. We also propose a method to select a reduced set of learning samples in order to mantain the internal structure of an object to be able to use high-order dependencies reducing the computational load.

Iberian Conference on Pattern Recognition & Image Analysis (IbPRIA'03)
Mallorca, Spain, June 2003

* Universitat Autonoma de Barcelona
† MERL - Research Laboratory

Published in: *Iberian Conference on Pattern Recognition & Image Analysis* (IbPRIA'03)

Modeling High-Order Dependencies in Local Appearance Models *

David Guillamet[1], Baback Moghaddam[2], and Jordi Vitrià[1]

[1] Computer Vision Center-Dept. Informàtica
Universitat Autònoma de Barcelona
08193 Bellaterra, Barcelona, Spain
{davidg, jordi}@cvc.uab.es
[2] Mitsubishi Electric Research Laboratories
Cambridge, MA 02139, USA
baback@merl.com

Abstract. We propose a novel local appearance modeling method for object detection and recognition in cluttered scenes. The approach is based on the joint distribution of local feature vectors at multiple salient points and their factorization with Independent Component Analysis (ICA). The resulting densities are simple multiplicative distributions modeled through adaptative Gaussian mixture models. This leads to computationally tractable joint probability densities which can model high-order dependencies. Our technique has been initially tested with natural and cluttered scenes with some degree of occlusions yielding promising results. We also propose a method to select a reduced set of learning samples in order to mantain the internal structure of an object to be able to use high-order dependencies reducing the computational load.

1 Introduction

For appearance based object modeling in images, the choice of method is usually a trade-off determined by the nature of the application or the availability of computational resources. Existing object representation schemes provide models either for global features [13], or for local features and their spatial relationships [10, 1, 12, 5]. With increased complexity, the latter provides higher modeling power and accuracy. Among various local appearance and structure models, there are those that assume rigidity of appearance and viewing angle, thus adopting more explicit models [12, 10, 9]; while others employ stochastic models and use probabilistic distance and matching metrics [5, 8, 1].

Recognition and detection of objects is achieved by the extraction of low level feature information in order to obtain accurate representations of objects. In order to obtain a good description of objects, extracted low level features

* This work is supported by Comissionat per a Universitats i Recerca del Departament de la Presidencia de la Generalitat de Catalunya and Ministerio de Ciencia y Tecnologia grant TIC2000-0399-C02-01.

must be carefully selected and it is often necessary to use as many salient features as possible. But one of the most common problems in computer vision is the computational cost of dealing with high dimensional data as well as the intractability of joint distributions of multiple features.

We propose a novel local appearance and color modeling method for object detection and recognition in cluttered scenes. The approach is based on the joint distribution of local feature vectors at multiple salient points and factorization with Independent Component Analysis (ICA). Taking this new statistically independent space to create $k = 3$ tuples ($k = 3$ salient points) of the most salient points of an object, we are able to obtain a set of joint probability densities which can model high-order dependencies. In order to obtain a good estimation of the tuple space, we use an adaptative Gaussian mixture model based on the Minimum Description Length (MDL)[14] criterion to optimally represent our data.

We have tested our method in a real and complex environment where we detect a real object (the US Pentagon building) after 9/11/01. We demonstrate that our technique is able to detect a complex object with a damaged portion of the building and under different natural conditions but we have to select a properly number of training tuples. Our method is based on high-order dependencies but, since the object consists of several keypoints, the number of possible tuples for learning is extremely huge. Thus, we propose a method to select the learning tuples in order to be able to work with high-order dependencies using a reasonable amount of computational resources.

2 Methodology

We propose to use an adaptative Gaussian mixture model as a parametric approximation of the joint distribution of image features of local color and appearance information at multiple salient points.

Let i be the index for elementary feature components in an image, which can be pixels, corner/interest points [3, 4], blocks, or regions in an image. Let x_i denote the feature vector of dimension n at location i. x_i can be as simple as {R,G,B} components at each pixel location, some invariant feature vectors extracted at corner or interest points [7, 10, 11], transform domain coefficients at an image block, and/or any other local/ regional feature vectors.

For model-based object recognition, we use the *a posteriori* probability defined as $max_l P(M_l|T)$ where M_l is the object model and $T = \{x_i\}$ represents the features found in the test image. Equivalently, by assuming equal priors, classification/detection will be based on maximum likelihood testing:

$$max_l P(T|M_l) \tag{1}$$

For the class-conditional density in equation (1), it is intractable to model dependencies among all x_i's (even if correspondence is solved), yet to completely ignore these dependencies is to severely limit the modeling power of the probability densities. Objects frequently distinguish themselves not by individual regions

(or parts), but by the relative location and comparative appearance of these regions. A tractable compromise between these two modeling extremes (which does not require correspondence) is to model the joint density of all k-tuples of x_i's in T. Figure (1) shows a general scheme of our methodology.

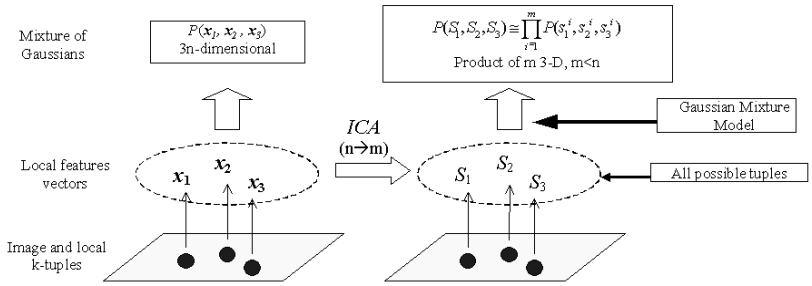


Fig. 1. System diagram for k-tuple density factorization using ICA and Gaussian mixture models.

2.1 Joint Distribution of k-tuples

Instead of modeling the total joint likelihood of all $x_1, x_2, \ldots x_I$, which is an $(I \times n)$-dimensional distribution, we model the alternative distribution of all k-tuples as an approximation:

$$P(\{(x_{i_1}, x_{i_2}, \ldots, x_{i_k})\}|M_l) \tag{2}$$

This becomes a $(k \times n)$-dimensional distribution, which is still intractable (Note: $k < n$ and $k << I$). We can use multi-dimensional histograms as an approximation of the joint distribution of image features with, i.e 20 histogram bins along each dimension, and such a framework would require $20^{(k \times n)}$ bins. Therefore, a factorization of this distribution into a product of low-dimensional distributions is required. We achieve this factorization by transforming x into a new feature vector S whose components are (mostly) independent. This is where Independent Component Analysis (ICA) comes in.

2.2 Density Factorization with ICA

ICA originated in the context of blind source separation [2, 6] to separate "independent causes" of a complex signal or mixture. It is usually implemented by pushing the vector components away from Gaussianity by minimizing high-order statistics such as the 4^{th} order cross-cumulants. ICA is in general not perfect therefore the IC's obtained are not guaranteed to be completely independent.

By applying ICA to $\{x_i\}$, we obtain the linear mapping

$$x \approx AS \tag{3}$$

and

$$P(\{(S_{i_1}, S_{i_2}, \ldots, S_{i_k})\}|M_l)$$
$$\approx \prod_{j=1}^{m} P(\{(s^j_{i_1}, s^j_{i_2}, \ldots, s^j_{i_k})\}|M_l) \quad (4)$$

where A is a n-by-m matrix and S_i is the "source signal" at location i with nearly independent components (Note: $m < n$). The original high-dimensional distribution is now factorized into a product of m k-dimensional distributions, with only small distortions expected. We note that this differs from so-called "naive Bayes" where the distribution of feature vectors is assumed to be factorizable into 1D distributions for each component. Without ICA the model suffers since in general these components are almost certainly statistically dependent.

After factorization, each of the k dimensional factored distributions becomes manageable if k is small, e.g., $k = 2$ or 3. Moreover, matching can now be performed individually on these low-dimensional distributions and the scores are additively combined to form an overall score.

Figure (2) is a graphical model showing the dependencies between a pair of 3-dimensional feature vectors x_1, x_2. The joint distribution over all nodes is 6-dimensional and all nodes are (potentially) interdependent. The basic approach towards obtaining a tractable distribution is to remove intra-component dependencies (vertical and diagonal links) leaving only inter-component dependencies (horizontal links). Simultaneously, we seek to reduce the number of observed components from $n = 3$ to a smaller number $m = 2$ of "sources". Ideally, a perfect ICA transform results in the graphical model shown in the right diagram where the pair S_1, S_2 only have pair-wise inter-component dependencies. Therefore, the resulting factorization can be simply modeled by 2D histograms or Gaussian mixture models[3].

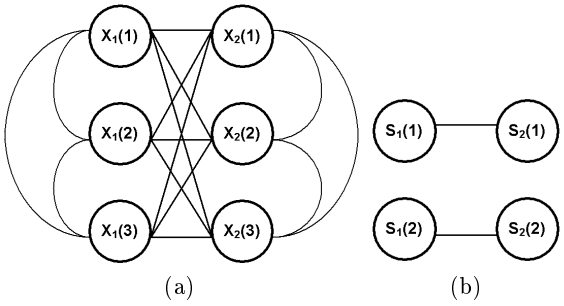


Fig. 2. Graphical models: (a) fully-connected graph denoting no independence assumptions (b) the ICA-factorized model with pair-wise only dependencies.

[3] We should note that in practice with an approximate ICA transform, the diagonal links of the original model are less likely to be removed than the vertical ones.

3 Experimental Results

Our experimental results have been focused on the use of $k = 3$ tuples in order to analyze the effect of choosing different learning tuples. We used a Harris operator [4, 11] to detect interest points and extracted the first 9 differential invariant jets [7] at each point as the corresponding feature vector x. Using these jets as our feature results in a local appearance model which is not only invariant to in-plane rotation (and translation) but is also robust with respect to partial occlusions. We must emphasize however that our methodology is not restricted to differential invariant jets and can in principal be used for any local set of features, for example, color, curvature, edge-intensity, texture moments. We then performed ICA to get $m < 9$ independent components for the feature vectors (jets). Using a $k = 3$ tuple model results in a set of $3D$ Gaussian mixture models which were used to model our 3-tuple joint component densities.

We tested our system with real and cluttered scenes where objects can be affected by different natural factors. This is the case presented in figure (3) which shows the modeling of the US Pentagon building before and after the September 11 terrorist attack. Figure (3.a) presents a real image of the pentagon building and figure (3.b) shows the extracted building used for our learning and modeling. Figure (3.c) depicts a test image which was taken after the bombing debris was cleared away by the cleanup crew (leaving a whole section of the building missing).

(a) (b) (c)

Fig. 3. (a) Satellite image of the US Pentagon building (prior to 9/11/01). (b) extracted building region used for learning. (c) a new test image of the same region taken after 9/11/01 under different natural conditions and with the damaged portion of the building missing (removed after site cleanup). (Note: All images have been rescaled for display purposes.)

Image of figure (3.b) has been used as training and the number of extracted keypoints is approximately 250. All possible $k = 3$ tuples that we can generate from 250 keypoints is extremely huge (like $250 \times 249 \times 248 = 15438000$) and it is impossible to learn a mixture of Gaussians with this huge number of training tuples. Our idea is to select a subset of them in order to find a representative

set of tuple candidates to learn the Gaussian mixture models and obtain a good representation of the natural object. In order to manage with natural occlusions, tuples must be carefully selected. Thus, we defined a radial threshold (R_{thr}) and we only consider those tuples that the distance between each keypoint of the tuple with respect to the middle point of the tuple is less than R_{thr}. This idea is represented in figure (4) where we can see three local features (x_1, x_2 and x_3) and the middle point of the tuple. When all the three distances (R_1, R_2 and R_3) between each feature and the middle point are less than R_{thr}, the tuple will be considered for training. As can be seen, this idea comes out in order to consider tuples with close keypoints to mantain the object structure.

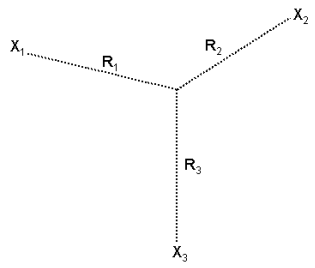
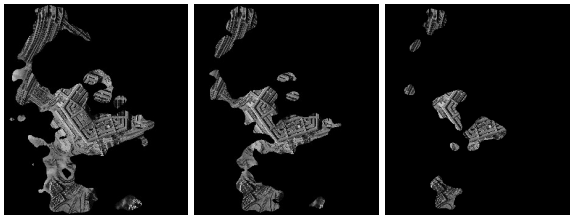


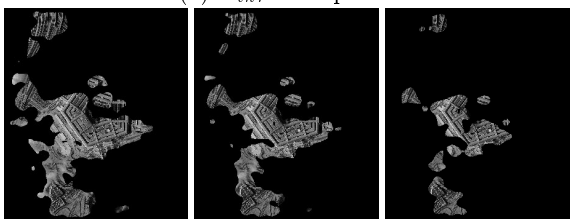
Fig. 4. Given 3 local features (x_1, x_2 and x_3) to create a $k = 3$ tuple, we obtain the middle point and when all the distances (R_1, R_2 and R_3) between the middle point to all the three features are less than a predefined radial value (R_{thr}), this tuple is considered for training.

This present work shows that a good criterion to choose a set of learning tuples is fundamental in order to obtain satisfactory results. Our pentagon object used for learning is about 120×120 pixels and, as seen in figure (3), it consists of several structured parts but repeated along the object. After obtaining all the pentagon keypoints, we have considered a set of learning tuples with a radial threshold of $25, 30, 35, 40$ and 45 pixels because we need to mantain the structure of the object. For example, a radial threshold of 45 pixels is about a quarter of the pentagon and, as seen, it should be enough because our pentagon contains a repeated structure. In case that a learning object consists of several and different structured parts, the radial threshold for our learning tuples should be analyzed more carefully. Detection maps corresponding to different radial thresholds can be seen in figure (5) where we can appreciate that small radial thresholds lead to bad detection maps and big radial thresholds lead to good (or acceptable) detection maps. We should state that the number of training tuples when we use big radial thresholds are really huge and our adaptative gaussian mixture model needs a considerable amount of computational resources.

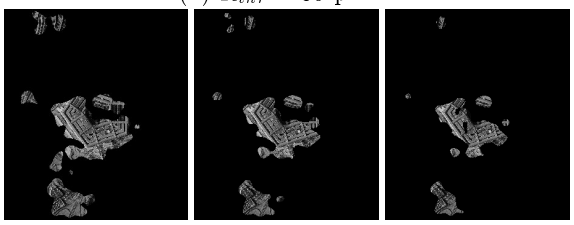
Since we are testing our method with an object with a missing part, see figure (3.c), detection maps of figure (5) are understandable in the sense that part of the pentagon may not be recognized properly. When using a $R_{thr} = 25$



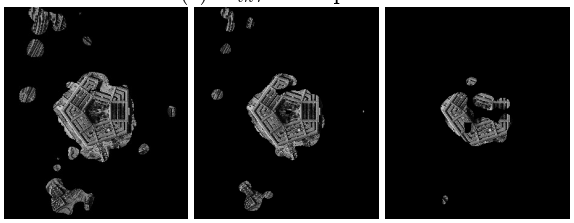
(a) $R_{thr} = 25$ pixels



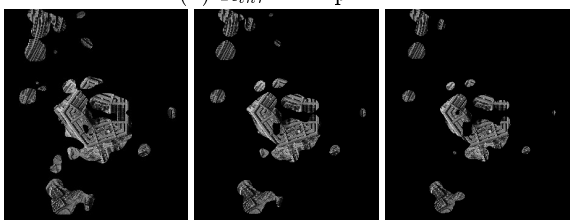
(b) $R_{thr} = 30$ pixels



(c) $R_{thr} = 35$ pixels



(d) $R_{thr} = 40$ pixels



(e) $R_{thr} = 45$ pixels

Fig. 5. Detection maps corresponding to different radial thresholds (from $R_{thr} = 25$ to $R_{thr} = 45$ pixels).

pixels, results are not acceptable since the pentagon is not correctly detected and a lot of external regions are considered as the pentagon. But, when using

$R_{thr} = 40$ pixels, pentagon is correctly detected and only a few external regions are considered as being part of the pentagon object.

4 Conclusions

A novel probabilistic modeling scheme was proposed based on the factorization of high-dimensional distributions of local image features. Our framework was tested using real imagery where the US Pentagon building is learned and detected in other natural conditions and with a damaged portion of the building missing. These experiments with complex and cluttered scenes demonstrate that this technique is well suited to object detection and localization tasks in natural environments. As seen, one of the problems is the huge number of training tuples obtained when considering high-order dependencies and the associated computational resources required that are extremely high. Thus, we propose a method to select a reduced set of learning tuples in order to mantain the internal structure of the object to be able to use high-order dependencies reducing the computational load.

References

1. Chang P., Krumm, J.: Object recognition with color cooccurrence histograms. In Proc CVPR, 1999
2. Comon P.: Independent component analysis - a new concept? Signal Processing, 36:287-314, 1994
3. Deriche R., Giraudon G.: A computational approach for corner and vertex detection. International Journal Computer Vision, 10(2): 101-124, 1993.
4. Harris C., Stephens M.: A combined corner and edge detector. In Alvey Vision Conf. 1988, pp. 147-151
5. Huang J., Kumar S.R., Mitra M., Zhu W.J., Zabih R.: Image indexing using color correlograms. In Proc. of International Conference in Computer Vision and Pattern Recognition, 1997
6. Jutten C., Herault J.: Blind separation of sources. Signal Processing, 24:1-10,1991
7. Koenderink J.J., van Doorn A.J.: Representation of local geometry in the visual system. Biological Cybernetics, 55: 367-375, 1987
8. Moghaddam B., Biermann H., Margaritis D.: Regions-of-Interest and Spatial Layout in Content based Image Retrieval. In Proc. of European Workshop on Content Based Multimedia Indexing, 1999
9. Moghaddam B., Pentland A.: Probabilistic Visual Learning for Object Representation. IEEE Transactions on PAMI, 19(7): 696-710, 1997
10. Schmid C., Mohr R.: Local grayvalue invariants for image retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence 19 (5), 530–534, 1997
11. Schmid C., Mohr R., Bauckhage C.: Comparing and evaluating interest points. In Proc ICCV, 1998.
12. Schneiderman H., Kanade T.: Probabilistic Modeling of Local Appearance and Spatial Relationships for Object Recognition. In Proc of CVPR, pp. 45-51, 1998.
13. Swain, M.J., Ballard, D.H.: Color Indexing. International Journal of Computer Vision, vol. 7, pp. 11-32, 1991
14. Tenmoto H., Kudo M., Shimbo M.: MDL-Based Selection of the Number of Components in Mixture Models for Pattern Recognition. In SSPR/SPR, pp. 831-836, 1998.