MVA2002 IAPR Workshop on Machine Vision Applications, Dec. 11 - 13, 2002, Nara- ken New Public Hall, Nara, Japan

8—27

Joint Distribution of Local Image Features for Appearance Modeling

David Guillamet*
Computer Vision Center
Universitat Autònoma de Barcelona

Baback Moghaddam†
Mitsubishi Electric Research Laboratories

Abstract

We propose an improved local appearance and color modeling method, as an extension of Moghaddam & Zhou [10], 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). We we are able to obtain a tractable set of joint probability densities which can model high-order dependencies in local image features. In this work we replace multi-dimensional histograms with Gaussian mixture models with model-order selection based on the Minimum Description Length (MDL) criterion. Furthermore, a hybrid color/appearance modeling scheme is introduced which significantly increases performance.

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 [15], or for local features and their spatial relationships [12, 1, 14, 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 [14, 12, 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 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, an extension of Moghaddam & Zhou [10], 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 this paper, we focus exclusively on object/image modeling with Gaussian mixture models (as opposed to histograms) which are optimally tuned using the Minimum Description Length (MDL) criterion to address the model-order selection problem. A detailed description of our model is given in Section 2. Experimental results with a subset of the COIL-100 database and real cluttered scenes are described in Section 3.

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, 12, 13], 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

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

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:

$$max_l P(T|M_l) \tag{2}$$

For the class-conditional density in equation (2), 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.

*Address: 08193 Bellaterra, Barcelona, Spain. E-mail: davidg@cvc.uab.es

†Address: 201 Broadway 8th floor, Cambridge, MA 02139, USA. E-mail: baback@merl.com

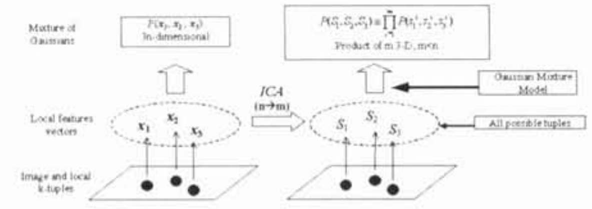


Figure 1: Graphical representation of a k-tuple density factorization using ICA [10].

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:

$$P(\{(x_{i_1}, x_{i_2}, \ldots, x_{i_k})\}|M_l) \tag{3}$$

This becomes a $(k \times n)$-dimensional distribution, which is still intractable (Note: $k < n$ and $k << I$). As in our previous work [10], using multi-dimensional histograms as an approximation of the joint distribution of image features with 20 histogram bins along each dimension, 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 Data factorization based on 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

$$x \approx AS \tag{4}$$

and

$$\begin{aligned} &P(\{(S_{i_1}, S_{i_2}, \ldots, S_{i_k})\}|M_l) \\ &\approx \prod_{j=1}^{m} P(\{(s^j_{i_1}, s^j_{i_2}, \ldots, s^j_{i_k})\}|M_l) \end{aligned} \tag{5}$$

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 1-D 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.

3 Experimental results

This paper is an extension of the previous work of [10] where multi-dimensional histograms of joint distributions of image features have been replaced by adaptative Gaussian mixture models, tuples have been extended from $k = 2$ to $k = 3$ points and we have added local color information to our model, all of which have resulted in great improvement with respect to previous results. As we will see, this new model representation is more suited for "cluttered" imagery where complex objects need to be modeled.

For our experiments, we used a Harris operator [4, 13] to detect interest points and extracted the first 9 differential invariant jets [7] at each point as the corresponding feature vector x. Our previous study [10] analyzed $k = 2$ tuples by using joint 2D histograms as a non-parametric approximation of the joint distribution of image features at multiple image locations. Now, an extension of this study has considered $k = 3$ tuples using 3D histograms and results demonstrate that this framework is quite powerful. Using the first 20 objects of COIL100 [11] as a reduced database of objects in order to analyze our technique, we have compared results using multi-dimensional histograms as described in [10] with $k = 2$ and $k = 3$ tuples. Results are presented in table (1).

	9D k tuples			3D k tuples		
Instances	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
1 Instance	3	17	20	16	20	20
2 Instance	1	4	4	8	7	6
3 Instance	1	4	4	6	5	2
4 Instance	1	4	5	4	5	4
5 Instance	1	4	3	6	4	3
Train	15%	85%	100%	80%	100%	100%
Test	5%	20%	20%	30%	26.25%	18.75%

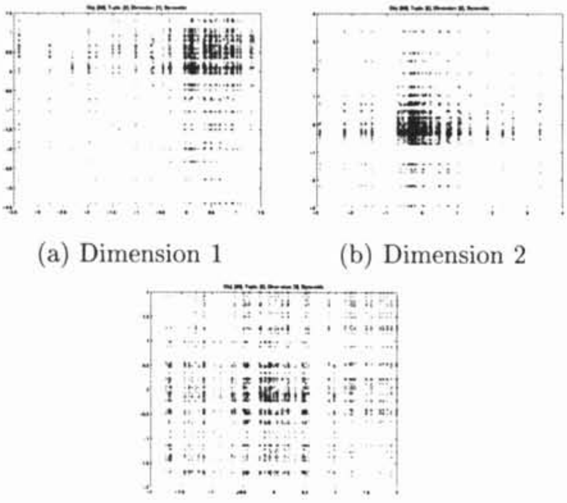
Table 1: No. of correct matches (out of 20) and resulting recognition rates when considering the original $9D$ invariant jets (appearance model) assuming independence and $3D$ ICA transformed vectors that are really independent. First instance was used as training, and 4 new instances were used to test.

This first experiment depicted in table (1) shows that when we only use an apperance model based on the original $n = 9$ dimensional vectors, recognition rates are lower than considering a $m = 3$ independent feature space obtained using ICA. In our previous study [10], the $m = 3$ dimensional independent space obtained using ICA was the optimal space in terms of recognition rates. But it is interesting to note that as k is increased (more points per tuple are considered), recognition rates are decreased when using multi-dimensional histograms. We note that we have used 32 bins per dimension, so that, when using $k = 3$ tuples, we are using a total $32^3 = 32768$ bins. As pointed out in our previous study [10], quite a few internal parameters of histograms must be fine-tuned in order to obtain a reliable representation, but it seems that when this space gets huge (as when working with $k = 3$ tuples), this representation is not appropriate.

A sample data distribution projected in our $m = 3$ dimensional ICA space is shown in figure (2) where we can appreciate that our tuple space is highly complex. This figure (2) shows the $k = 2$ tuple case but when

$k = 3$ tuples are considered, data distributions are more localized and complex. Using multi-dimensional histograms to represent this space would mean that we must have a very high precision in order to capture specific behaviours. Another drawback of using multi-dimensional histograms is that each model would need a lot of resources to be saved in disk.

Gaussian mixture models are a reliable alternative to represent this space for two specific reasons: simplicity and adaptability. Using a mixture of Gaussians would mean that local concentrations will be captured by each Gaussian, thus resulting in an adaptative model with fewer parameters. The typical problem of using a mixture of Gaussians as a model is the choice of the number of components to be used to represent data (also known as "model-order selection"). In our particular case, we used and adaptative mixture model [9] based on the Minimum Description Length (MDL) [16] optimality criterion to fit our data.


(a) Dimension 1

(b) Dimension 2

(c) Dimension 3.

Figure 2: Three 2 dimensional spaces representing one of the objects in the database.

As an illustration of mixtures of Gaussians using different number of components, we show table (2) where recognition rates using 5 and 10 Gaussians per model are used. Now, we can appreciate that depending on the number of Gaussians, results may vary. But it is important to note that the expected behaviour of incrementing the number of points per tuple (k) with an increasing of the recognition rate is reflected in this results (with the multi-dimensional histograms we did not have this behaviour). Using a correct estimation of the number of components through the Minimum Description Length (MDL) criterion, we are able to obtain better recognition results as shown in table (3).

	5 Gaussians 3D k tuples			10 Gaussians 3D k tuples		
Instances	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
1 Instance	20	17	16	20	19	18
2 Instance	13	14	14	14	14	15
3 Instance	11	14	15	10	13	14
4 Instance	12	12	12	9	11	12
5 Instance	9	10	12	7	10	12
Train	100%	85%	80%	100%	95%	90%
Test	56.25%	62.5%	66.25%	50%	60%	66.25%

Table 2: Recognition rates when considering a mixture of Gaussians in a $m = 3$ dimensional ICA space. We present a mixture of Gaussians using 5 and 10 Gaussians.

	MDL Estimation 3D k tuples		
Instances	$k = 1$	$k = 2$	$k = 3$
1 Instance	20	20	20
2 Instance	14	15	16
3 Instance	13	14	15
4 Instance	12	13	15
5 Instance	11	12	14
Train	100%	100%	100%
Test	62.5%	67.5%	75.0%

Table 3: Recognition rates when the number of Gaussians is estimated using the Minimum Description Length (MDL) criterion.

Since these low recognition rates are not very satisfactory, we introduced a hybrid appearance/color model by introducing the mean color of each normalized channel (red, green and blue channels) obtained from a circular region defined around each interesting point. Local color histograms can also be considered as reliable local color information but since our appearance local descriptors are defined by 9 dimensional vectors, we only have introduced a 3 dimensional color descriptor.

But the addition of color introduces essentially one degree of freedom (information) to the model and we would expect that $m = 4$ dimensional ICA spaces would suffice (indeed $m = 3$ results in poor performance). Recognition results considering a projected space of 4 and 5 dimensions are shown in table (4). We note that going to higher dimensions ($m = 5$) reduces performance due to the fact that the ICA factorization and modeling in higher dimensions is more difficult (esp. if the data has a lower intrinsic dimensionality of $m = 4$).

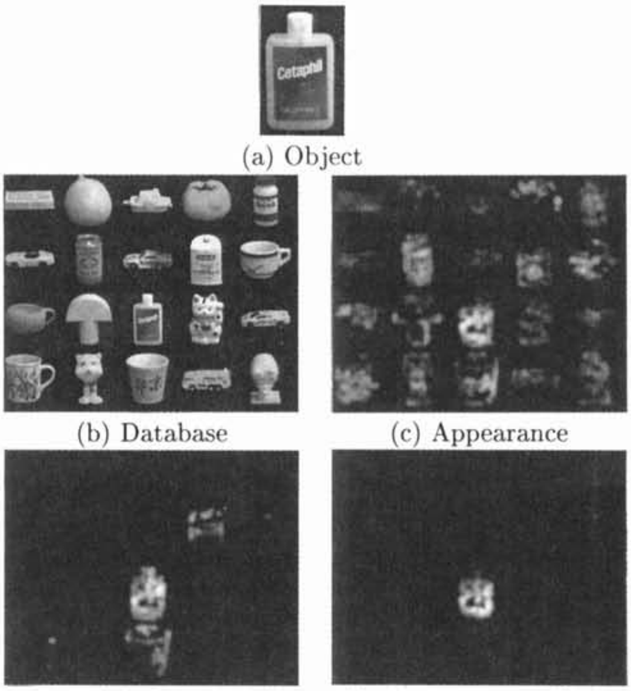
In conclusions, we found that the hybrid appearance/color model provides the best classification results when the statistically independent space obtained by ICA is defined using 4 dimensions. In this particular case, we found that using MDL adaptive mixtures did not improve the (already high) recognition performance in Table — *ie.* 15 components sufficed in capturing the distributions). With larger databases, this is not necessarily true and the use of MDL estimation is critically important.

	Appearance and Color Model 15 Gaussians 4D k tuples			Appearance and Color Model 15 Gaussians 5D k tuples		
Instances	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
1 Instance	20	20	20	20	20	20
2 Instance	20	20	20	16	18	19
3 Instance	20	20	20	15	18	18
4 Instance	18	19	20	15	17	17
5 Instance	19	19	19	14	16	18
Train	100%	100%	100%	100%	100%	100%
Test	96.25%	97.5%	98.75%	75%	86.25%	90%

Table 4: Recognition rates when 4 and 5 dimensional feature spaces are used to build our appearance and color model. As it was initially expected, using a 4 dimensional feature space we are able to obtain the best recognition results.

For an illustration of our current object classification framework, some visual results are presented in figure (3) where different likelihood maps of our joint density functions are shown when the particular object model of figure (3.a) is used for object detection.

We also tested our new approach under real and cluttered scenes where objects can be affected by different natural factors. This is the case presented in figure (4) which shows the modeling and subsequent detection of the US Pentagon building before and after the September 11 terrorist bombing. Figure (4.a) presents a real image of a pentagon building and figure



(a) Object

(b) Database

(c) Appearance

(d) Color

(e) Appearance + Color

Figure 3: (c), (d) and (e) are the likelihood maps obtained from image (b) when the object presented in (a) is used for object detection using different models.

(4.b) shows the extracted building used for our learning and modeling. Figure (4.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). This test image was also taken at a different time of day and under different weather conditions. Figure (4.d) shows the graphical likelihood map thresholded and multiplied by the original test image in order visualize the detected region (where the model likelihood is very high). We can see that our improved local appearance models (which are quite general in formulation) were found to be satisfactory for satellite/aerial imagery.

References

[1] P. Chang and J. Krumm. Object recognition with color cooccurrence histograms. CVPR'99, Colorado, June. 1999.

[2] P. Comon. Independent component analysis - a new concept? Signal processing 36:287-314, 1994.

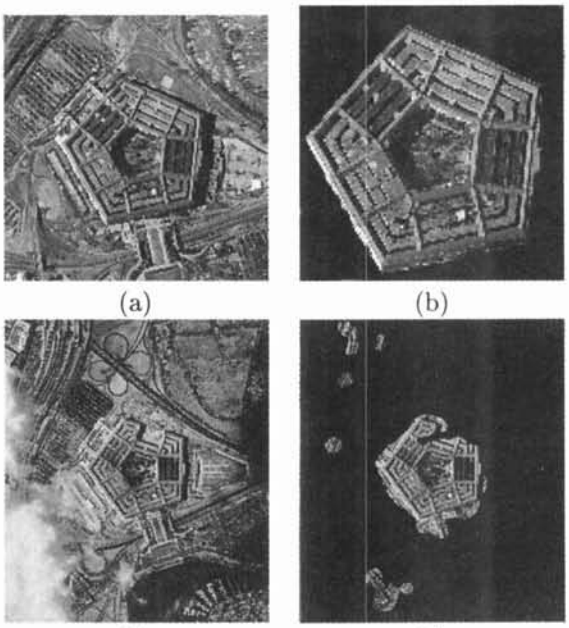
[3] R. Deriche and G. Giraudon. A computational approach for corner and vertex detection. International Journal of Computer Vision, vol. 10, n. 2, pp. 101-124, 1993.

[4] C. Harris and M. Stephens. A combined corner and edge detector. In Alvey Vision Conf. 1988, pp. 147-151.

[5] J. Huang, S.R. Kumar, M. Mitra, W.J. Zhu and R. Zabih. Image indexing using color correlograms. CVPR'97, San Juan, Puerto Rico.

[6] C. Jutten and J. Herault. Blind separation of sources. Signal processing, 24:1-10, 1991.

[7] J.J. Koenderink and A.J. van Doorn. Representation of local geometry in the visual system. Biological Cybernetics, vol. 55, pp. 367-375, 1987.



(a)

(b)

(c)

(d)

Figure 4: (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). (d) the highest probability target area of the test image given our local appearance model of (b). (Note: All images have been rescaled for display purposes.)

[8] B. Moghaddam, H. Biermann and D. Margaritis. Regions-of-Interest and Spatial Layout in Content based Image Retrieval. European Workshop on Content based Multimedia Indexing, CBMI'99, France, Oct. 1999.

[9] B. Moghaddam and A. Pentland. Probabilistic Visual Learning for Object Representation. PAMI 19(7), pp. 696-710, Jul. 1997.

[10] B. Moghaddam and X.S. Zhou. ICA-based Probabilistic Local Appearance Models, ICIP'01, October 2001.

[11] S.A. Nene, S.K. Nayar and H. Murase. Columbia Object Image Library: COIL-100. Technical report CUCS-006-96, Dept. Computer Science, Columbia University, February 1996.

[12] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. PAMI, 19(5):530-534, 1997.

[13] C. Schmid, R. Mohr and C. Bauckhage. Comparing and evaluating interest points. Proc ICCV, 1998.

[14] H. Schneiderman and T. Kanade. Probabilistic Modeling of Local Appearance and Spatial Relationships for Object recognition. CVPR'98, pp. 45-51. 1998. Santa Barbara, CA.

[15] M.J. Swain and D.H. Ballard. Color Indexing, International Journal of Computer Vision, vol. 7, pp. 11-32, 1991.

[16] H. Tenmoto, M. Kudo and M. Shimbo. MDL-Based Selection of the Number of Components in Mixture Models for Pattern Recognition. In SSPR/SPR, pp. 831-836, 1998.