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

Local Appearance-Based Models Using High-Order Statistics of Image Features

Baback Moghaddam David Guillament* Jordi Vitria†

TR-2003-085 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 factorization with Independent Component Analysis (ICA). The resulting densities are simple multiplicative distributions modeled through adaptive Gaussian mixture models. This leads to computationally tractable joint probability densities which can model high-order dependencies. Furthermore, different models are compared based on appearance, color and geometry information. Also, the combination of all of them results in a hybrid model which obtains the best results using the COIL-100 object database. Out technique has been tested under different natural and cluttered scenes with different degrees of occlusions with promising results. Finally, a large statistical test with the MNIST digit database is used to demonstrate the improved performance obtained by explicit modeling of high-order dependencies.

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Copyright © Mitsubishi Electric Research Laboratories, Inc., 2003
201 Broadway, Cambridge, Massachusetts 02139

* Universitat Autonoma de Barcelona
† Universitat Autonoma de Barcelona

Publication History:

1. First printing, TR-2003-085, June 2003

Local Appearance-Based Models using High-Order Statistics of Image Features

Baback Moghaddam
Mitsubishi Electric
Research Laboratories
201 Broadway, Cambridge, MA 02139
baback@merl.com

David Guillamet, Jordi Vitrià
Computer Vision Center, Dept. Informàtica
Universitat Autònoma de Barcelona
08193 Bellaterra, Barcelona, Spain
{davidg,jordi}@cvc.uab.es

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 factorization with Independent Component Analysis (ICA). The resulting densities are simple multiplicative distributions modeled through adaptive Gaussian mixture models. This leads to computationally tractable joint probability densities which can model high-order dependencies. Furthermore, different models are compared based on appearance, color and geometry information. Also, the combination of all of them results in a hybrid model which obtains the best results using the COIL-100 object database. Our technique has been tested under different natural and cluttered scenes with different degrees of occlusions with promising results. Finally, a large statistical test with the MNIST digit database is used to demonstrate the improved performance obtained by explicit modeling of high-order dependencies.

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 [17], or for local features and their spatial relationships [14, 3, 16, 7]. 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 [16, 14, 12]; while others employ stochastic models and use probabilistic distance and matching metrics [7, 11, 3].

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 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 adaptive Gaussian mixture model based on the Minimum Description Length (MDL)[18] criterion to optimally represent our data. Once $k = 3$ tuples can be used to model high-order dependencies, we add geometry information to our model resulting in a hybrid model that is able to improve initial results.

We have tested our method in a closed environment where we recognize objects taken under different points of view (COIL-100 [13] database). We also tested our technique with different levels of occlusions demonstrating that our technique is able to deal with moderate amounts of occlusion due to its inherent local representation of apperance and geometry. Furthermore, our technique is also generalizable to real, complex and cluttered environments and we present some results of object detection in these scenarios with promising results. Finally, a very large statistically significant test (using the MNIST database) is used to illustrate the generality of feature representations in our scheme as well as explicitly demonstrating the advantage of modeling higher-order statistics in our tractable joint distributions.

2. Methodology

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

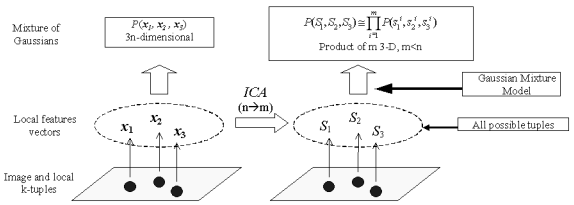


Figure 1. Diagram of our methodology.

Let i be the index for elementary feature components in an image, which can be pixels, corner/interest points [5, 6], 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 [9, 14, 15], 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: $\mathrm{max}_l P(M_l|T)$, 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:

$$\mathrm{max}_l P(T|M_l) \tag{1}$$

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.

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: $P(\{(x_{i_1}, x_{i_2}, \ldots, x_{i_k})\}|M_l)$. 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 [4, 8] 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 $x \approx AS$ and

$$\begin{aligned} & P(\{(S_{i_1}, S_{i_2}, \ldots, S_{i_k})\}|M_l) \\ \approx & \prod_{j=1}^{m} P(\{(s^j_{i_1}, s^j_{i_2}, \ldots, s^j_{i_k})\}|M_l) \end{aligned} \tag{2}$$

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[1].

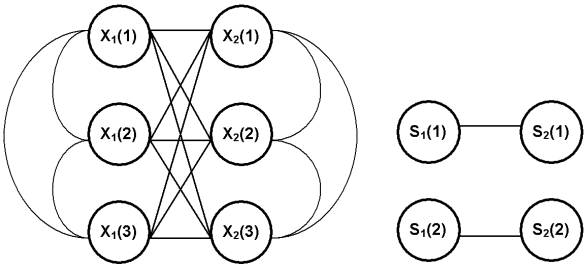


Figure 2. Graphical models: (a) fully-connected graph denoting no independence assumptions (b) the ICA-factorized model with pair-wise only dependencies.

2.3. Class-Conditional ICA

When object recognition consists of having r different classes and each class represented using a specific ICA model, it turns out that the combination of all ICA models must be normalized. In [2] a class-conditional ICA (CC-ICA) model is introduced that, through class-conditional representations, ensures conditional independence. The basic CC-ICA model is estimated from the training set for each class. If W_r and s_r are the projection matrix and the independent components for class C_r with dimensions $M_r \times N$ and M_r respectively, then $s^r = W^r(x - \overline{x}^r)$ where $x \in C_r$ and $\overline{x}^r$ is the class mean, estimated from the training set. Most ICA methods require, or at least advise, data whitening as preprocessing. Since some simple denoising is also recommended, dimensionality reduction and whitening through PCA is very common practice as a preprocessing stage for ICA. In this case, W^r can be decomposed as $W^r = B^r E^r$, where E^r is the $M_r \times N$ PCA whitening matrix and B_r the ICA unmixing matrix. Also $v^r \stackrel{def}{=} E^r(x - \overline{x}^r)$ is the whitened data. Assuming the class-conditional representation actually provides independent components, we have that the class-conditional probability in transformed space noted as $p^r(s) \stackrel{def}{=} p(s^r)$ can now be expressed in terms of unidimensional densities, $p(v|C_r) = \nu_r p^r(s) = \nu_r \prod_{m=1}^{M_r} p^r(s_m)$, with $\nu_r = (\int p^r(s)ds)^{-1}$, a normalizing constant. See [2] for more information.

[1]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

For our experiments, we used a Harris operator [6, 15] to detect interest points and extracted the first 9 differential invariant jets [9] at each point as the corresponding feature vector x. Using these jets as our features 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 as we shall see later. 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 or even shape descriptors (see Section 3.5). We then performed ICA to get $m < 9$ independent components for the feature vectors (jets). We then used $k = 1, 2, 3$, resulting in a set of 1D, 2D and 3D Gaussian mixture models which were used to model 1-tuple, 2-tuple and 3-tuple joint component densities. Initial experiments were done using multi-dimensional histograms as a non-parametric approximation of the joint distribution of tuples but results were not as satisfactory as parametric mixture models. Once an ICA space is defined, we used the definition of class-conditional ICA as described in the previous section in order to obtain the probability of a tuple belonging to each training class.

3.1. Appearance + Color Models

Experiments are based on 100 objects from the Columbia Object Image Library (COIL-100) [13]. Each object model is trained only using one instance per object and we have tested our method considering four new instances per each object captured from other points of view (each testing instance of the object is rotated 5 degrees in azimuth). Experiments demonstrate that appearance-based models (ie. using monochrome-based invariant jets) are not very satisfactory (see table (1)) therefore we introduced a hybrid appearance/color model by introducing the mean color of each normalized channel (R,G,B) obtained from a circular region defined around each interest point. Although color histograms [17] can also be used, given the local nature of the representation sought after, we limited it to the main dominant color in the surrounding region. Recognition rates using $k = 1$ tuples using appearance, color and a hybrid appearance/color model are presented in table (1). Table (1) can be understood as follows: row 1 with label *Instance 1* is the training instance used to create our object models and rows 2 to 5 are four new testing instances of each object — each of them rotated by 5 degrees from the previous instance. *Train* row is the recognition rate corresponding to the training set and *Test* row indicates the recognition rate obtained with the testing set.

As noted in Table (1), appearance model (first column

Table 1. Results using $k = 1$ tuples and a mixture of 10 Gaussians.

$k = 1$ tuples				
	Appearance (3D)	Color (3D)	Appearance + Color (3D)	Appearance + Color (4D)
Instance 1	85	31	93	89
Instance 2	38	25	52	66
Instance 3	33	23	50	55
Instance 4	28	18	49	57
Instance 5	28	17	42	52
Train	85%	31%	93%	89%
Test	31.75%	20.75%	48.25%	57.5%

of results) is reduced from a $n = 9$ dimensional space to a $m = 3$ dimensional ICA space since this ICA dimensionality was obtained by evaluating all possible values and selecting the best. In light of other experiments (not presented here for brevity), we strongly believe that our 9-dimensional jets have an intrinsic dimensionality of 3 components. The addition of (dominant) color introduces essentially one degree of freedom (information) to the model and we would expect that $m = 4$ dimensional ICA spaces would be the best (as in fact they were found to be). In Table (1) we present recognition results considering a projected ICA space of 4 dimensions. Since considering $k = 1$ tuples is the same as evaluating the probability of a single point to appear in one object model, recognition results are poor. Tables 2 and 3 show recognition rates when considering higher-order models with $k = 2$ and $k = 3$ tuples.

Table 2. Results using $k = 2$ tuples and a mixture of 10 Gaussians.

$k = 2$ tuples				
	Appearance (3D)	Color	Appearance + Color (3D)	Appearance + Color (4D)
Instance 1	98	54	99	99
Instance 2	44	32	70	73
Instance 3	33	30	68	71
Instance 4	28	33	58	63
Instance 5	29	29	62	64
Train	98%	54%	99%	99%
Test	33.5%	31%	64.5%	67.75%

Table 3. Results using $k = 3$ tuples and a mixture of 10 Gaussians.

$k = 3$ tuples				
	Appearance (3D)	Color	Appearance + Color (3D)	Appearance + Color (4D)
Instance 1	100	83	100	100
Instance 2	53	68	69	85
Instance 3	42	63	67	81
Instance 4	41	60	64	78
Instance 5	36	58	58	74
Train	100%	83%	100%	100%
Test	43%	62.25%	64.5%	79.50%

It is quite clear as seen in Tables 1, 2, and 3, that as the number of interest points per tuple is increased and hence more mutual information about the local appearance jets are modeled, the recognition rates are improved. Also, we should point out that results without an ICA factorization lead to recognition rates of 35%.

3.2. Incorporating Local Geometry

When considering $k = 3$ tuples, we can also take into account the geometry of a tuple — *i.e.* how the three points in a tuple are arranged spatially and use this information to perhaps increase our recognition capacity. Rank ordering the interest points in a tuple as (p_1, p_2, p_3), we can use the following geometrical descriptors:[2] (i) Distance L_{12}: Distance between p_1 and p_2. (ii) Distance L_{13}: Distance between p_1 and p_3. (iii) Angle α: Angle between the connecting line p_1, p_2 and the connecting line p_1, p_3. However, a feature space using these three geometric descriptors alone ($L_{12}, L_{13}, \cos(\alpha)$), results in a rather poor recognition rate of 50%. We thus conclude that geometry information alone is not sufficient for recognition. However, tuple geometry may help to disambiguate confusing appearance-based cases and therefore it helps to create a hybrid apperance/geometry mixture model.

Since the combined color and appearance model was found to be the best in previous experiments, we use this model as a first step classifier in our experiments and then used the geometric model as a second step classifier to resolve any possible ambiguities found by the first classifier. Table (4) shows the recognition rates when considering this new two-stage classifier design. We can see that the resulting (two-stage) recognition rates are better than those obtained with the (single stage) appearance/color model alone (see table (3)).

3.3. Model-Order Selection

So far we have reported results based on Gaussian mixture models which used an experimentally determined but fixed number of mixture components. The central problem of using a mixture of Gaussians as a model is of course the choice of the number of components (also known as "model-order selection"). Results with and without a Minimum Description Length (MDL) estimator can be seen in table (4) where we use an adaptive mixture model based on the MDL [18] optimality criterion used to fit the data in each case with the "right" number of components.

Comparing both columns of table (4), we see that using MDL has definitely enhanced the recognition performance of our system, undoubtedly through an increase in accuracy in modeling the joint distributions. First column of this table is using a mixture of 10 Gaussians for the fist classifer and 2 Gaussians for the geometry model.

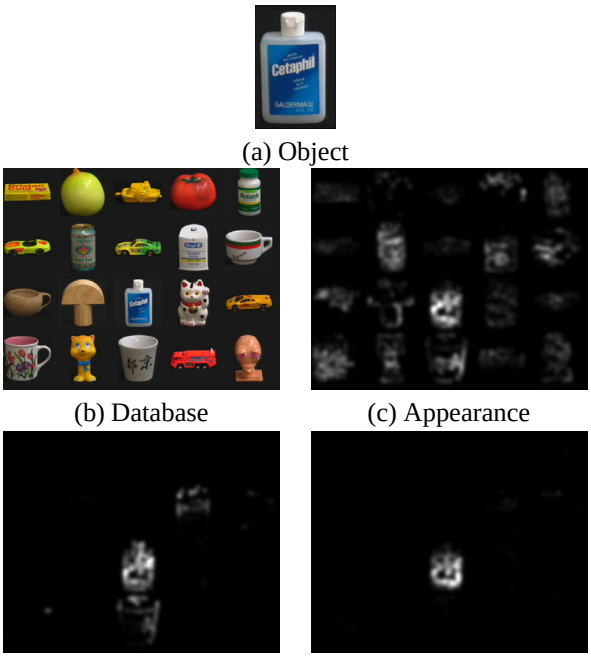
[2] Keypoints are ordered according to the output of the Harris operator, which is proportional to the principal curvatures of the intensity surface.[6]

Table 4. Results using a two-stage classifier and a MDL estimator.

	Classifier 1: Appearance + Color Model (4D) Classifier 2: Geometry Model	
	No MDL estimation	MDL estimation
Instance 1	100	100
Instance 2	88	96
Instance 3	86	94
Instance 4	81	91
Instance 5	77	86
Train	100%	100%
Test	83%	91.75%

3.4. Invariance to Partial Occlusion

As an illustration of our object classification framework, a representative visual example is shown in figure (3) where different likelihood detection maps (based on joint density functions) are shown when the particular object model of figure (3.a) is the search target. Note that the hybrid appearance/color model correctly localizes the target object from among the 20 object candidates arranged in the test image.



(a) Object

(b) Database (c) Appearance

(d) Color (e) Appearance + Color

Figure 3. Likelihood maps when using object (a) as the search target in the database (b).

Object detection and classification techniques should be robust under the presence of occlusions. Since our technique is based on local tuples obtained from a set of interest points, occlusions should be easy to deal with. In this particular case, we occluded parts of the test objects using the various quadrants assuming that the rest of the object would be sufficient to recover the original identity. Results are presented in table (5) where we use a first step classifier model based on appearance and color information and a second step classifier based on geometry, both using the MDL criterion to set the number of Gaussians.

Table 5. Results using $k = 3$ tuples and various quadrant occlusions (Q_1,Q_2,Q_3 and Q_4).

	Occlusions considered							
	Q_1	Q_2	Q_3	Q_4	Q_1+Q_2	Q_2+Q_3	Q_3+Q_4	Q_4+Q_1
Inst. 1	100	100	100	100	94	97	92	93
Inst. 2	83	83	70	80	60	55	62	54
Inst. 3	80	82	67	82	55	51	49	49
Inst. 4	75	74	59	76	57	57	54	58
Inst. 5	71	65	65	67	51	48	45	54
Train	100%	100%	100%	100%	94%	97%	92%	93%
Test	77.25%	76%	65.25%	76.25%	55.75%	52.75%	52.5%	53.75%

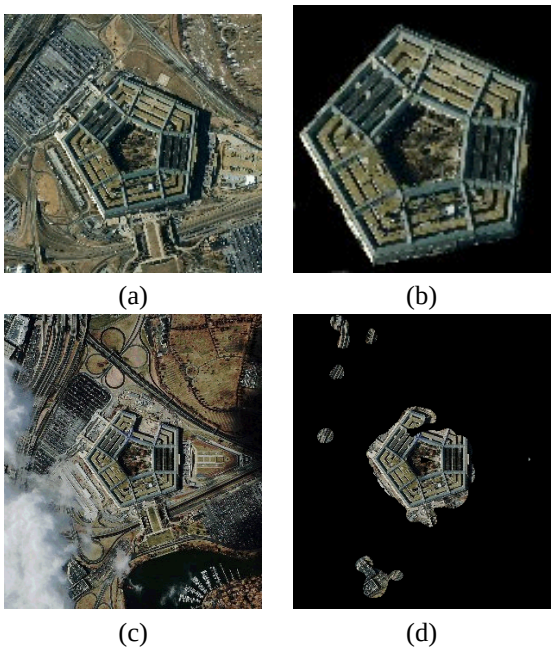
Table (5) shows that when one quadrant is missing, recognition rates are acceptable since we obtain an average rate of 75%. When two quadrants are missing, recognition rates decrease to an average of 52% but this is still a good trade-off between the level of occlusion and recognition rates by taking into account that we are testing object instances that are not previously learned (ie. images taken from other points of view).

We also evaluated our approach in real laboratory scenes where deformable objects can appear under various configurations, poses and occlusions. Figure (4) shows the objects and images used for this experiment. Two different objects with similar colors but different shapes were learned in order to detect them in a complex environment. As noted in this figure (4), objects can be hard to recognize since they contain different levels of occlusions and can be seen under different poses. Despite these difficulties, the objects are correctly detected indicating the level of robustness in our system.

Finally, we have also tested our system with real and cluttered scenes where objects can be affected by different natural factors. This is the case presented in figure (5) which shows the modeling and subsequent detection of the US Pentagon building before and after the September 11 terrorist attack. Figure (5.a) presents a real image of a pentagon building and figure (5.b) shows the extracted building used for our learning and modeling. Figure (5.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 (5.d) shows the graphical likelihood map thresholded and multiplied by the original test image in order to visualize the detected region (where the model likelihood is very high). We can see that our hybrid color/appearance model (which is quite general in formulation) is found to be satisfactory for such satellite/aerial imagery.

(a) (b) (c) (d) (e)

Figure 4. Columns (d) and (e) show the detection maps for objects (a) and (b), respectively for the scenes of column (c).



(a) (b) (c) (d)

Figure 5. Likelihood map of the US Pentagon with a damaged portion of the building missing. (Note: All images have been rescaled for display purposes.)

3.5. Modeling Higher-Order Dependencies

We next apply our object recognition scheme in a totally different context in order to demonstrate (1) how to integrate mutiple instances into a single model, (2) that our scheme can be used with other kinds of features and data representations and (3) that increasing the tuple order does in fact lead to improved performance. In this experiment, we chose the MNIST [10] digit database because it contains a huge number of training and testing samples ($60,000$ training samples and $10,000$ testing samples), so we can statistically verify that incrementing the order of our models will lead to better recognition rates. We must note that our scheme is not especially adapted to work with the MNIST database (which for one thing, is not even in color and has little in the way of appearance texture) rather it is a general technique for use in complex and cluttered scenes with the presence of occlusions. Our main goal here is to explore how increasing tuple order affects to the recognition rates using a well-known and large database.

In particular, features were extracted from hand-written MNIST digits using the same technique as in [1] where they obtain a set of shape histograms for each digit. In our case, each digit is represented by a set of 75 points sampled from the shape contour (75 pixel locations sampled from the output of the Canny dectector). Having 75 pixel locations, we have represented each location using a shape histogram (exactly the same as in [1]), so that each digit is represented by 75 shape histograms of 60 dimensions. In order to find the "right" ICA dimension to reduce our feature vectors, we did a k-NN (with k=5) based classification using the original shape histograms taking a reduced set of training and testing samples (200 training samples per each digit and the first $5,000$ testing samples) using the χ^2 test statistic (as in [1]) as a distance metric. Also, a k-NN (with k=5) based classification was done using the ICA projected feature vectors between $d = 5$ to $d = 50$ ICA dimensions with the same training and testing set as before using the L_1 norm as a distance metric in order to evaluate which is the ICA dimension that preserves the same recognition rates of the original space. The dimension found by the experiments to be the most suitable one to be used with our ICA scheme was 25, which was used thereafter.

We have tested two different approaches: (1) learn an adaptive mixture model per each training instance and (2) learn an adaptive mixture model per each digit class. Our factored $k = 2$ and $k = 3$ high-order models generate a huge number of tuples. In this particular case, when using $k = 2$ tuples, we generate an order of $5,000$ tuples per each digit and when using $k = 3$ tuples, $100,000$ possible tuples are generated. We randomly selected $1,000$ $k = 2$ tuples and $5,000$ $k = 3$ tuples to learn our adaptive Gaussian mixture models. For our experimental tests, we used

500 training samples per each digit ($5,000$ in total) and all the testing MNIST set ($10,000$ digits). Experimental results are shown in Table (6) where we can clearly see that incrementing the order of our models leads to an improvement in the recognition rates. Interestingly enough, we note also that there seems to be little difference between the two different approaches of handling multiple training instances: using one model/instance *vs.* one model/class.

Table 6. Results when using the MNIST database and our factorization model.

	k tuples		
Method	$k=1$ tuple	$k=2$ tuple	$k=3$ tuple
1 Model / Instance	74.23%	83.14%	91.57%
1 Model / Class	71.85%	82.03%	91.13%

Using the nearest neighbor classifier (k-NN with k=3) in the original space of shape histograms with the χ^2 test statistic, we obtain a recognition rate of 75.87% without using any point matching technique as in [1] and it is obvious that our method is not best-suited for the MNIST database (that is not the point here) but we do notice the improvement of our factored distribution models from $k=1$ to $k=3$. We should emphasize that even though we do not achieve the best reported recognition rates for the MNIST, our factored models with $k=3$ are not only significantly better than $k=1$ but also better than using k-NN in the original space of shape histograms (a recognition rate of 75.87%).

With this last experiment we can make the following observations: (1) Our technique can be extended to different data representations but in doing so, because it is a general technique, we can not expect it to obtain the best recognition rates, (2) by incrementing the order of our factored models, recognition rates can almost certainly be expected to improve, (3) our technique may be more suited to complex and cluttered scenes than to recognition of objects in closed object databases such as MNIST or COIL-100.

4. Conclusions

A novel probabilistic modeling scheme was proposed based on factorization of high-dimensional distributions of local image features. Our framework was tested using appearance and color information as well as using geometry information. An hybrid classifier based on all these local image features achieved the best recognition results. Our factored distributions were modeled using Gaussian mixture models based on the Minimum Description Length optimality criterion to fit our data. COIL-100 object database was tested with and without occlusions, obtaining very promising results. Also, experiments with complex and cluttered scenes demonstrate that this technique is well suited to object detection and localization tasks in natural environments. Finally, a large experiment with the MNIST digit database was performed in order to validate the underlying assumption that increasing the high-order dependencies of our factored distributions does in fact lead to improved performance. This experiment also demonstrated that different feature representations (other than invariant jets) can be readily used in our k-tuple ICA-factorization framework.

References

[1] S. Belongie, J. Malik and J. Puzicha. Shape Matching and Object Recognition Using Shape Contexts. PAMI 24(24), pp. 509-522, April 2002.

[2] M. Bressan, D. Guillamet and J. Vitria. Using a local ICA Representation of High Dimensional Data for Object Recognition and Classification. In CVPR'01, pp. 1004-1009.

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

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

[5] 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.

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

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

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

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

[10] Y. LeCun. The MNIST DataBase of Handwritten digits. http://yann.lecun.com/exdb/mnist/index.html.

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

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

[13] 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.

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

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

[16] 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.

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

[18] 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.