Recognizing Indoor Images with Unsupervised Segmentation and Graph Matching

Miguel Angel Lozano and Francisco Escolano

Robot Vision Group
Departamento de Ciencia de la Computación e Inteligencia Artificial
Universidad de Alicante, Spain
{malozano, sco}@dccia.ua.es
http://rvg.dccia.ua.es

Abstract. In this paper we address the problem of recognizing scenes by performing unsupervised segmentation followed by matching the resulting adjacency region graph. Our segmentation method is an adaptive extension of the Asymetric Clustering Model, a distributional clustering method based on the EM algorithm, whereas our matching proposal consists of embodying the Graduated Assignement cost function in a Comb Algorithm modified to perform constrained optimization in a discrete space. We present both segmentation and matching results that support our initial claim indicating that such an strategy provides both class discrimination and individual-within-a-class discrimination in indoor images which usually exhibit a high degree of perceptual ambiguity.

1 Introduction

Scene recognition is a key element in mobile robotics tasks like self-localization or exploration. Current approaches can be broadly classified as holistic [1] [2] [3] [4] and region-based[5] [6] [7]. Holistic methods exploit texture,color, and shape statistics without identifying objects previously. For instance, Torralba and Sinha propose a low dimensional representation that encodes statistics of Gabor filters' outputs and it is suitable for distinguishing views associated to specific parts of the environment [2]. On the other hand, region-based approaches allow the access to objects properties at the cost of incrementing the computational load due to the need of segmenting the images. One recent example is Carson et al's Blobworld framework [7], which relies on grouping pixels with similar features in regions and then use their characteristic statistics for identifying images with similar regions. We conjecture that inside indoor environments, which tend to be very ambiguous, the integration of segmentation and structural matching provides not only good recognition results at the class level, that is, distinguishing between views of corridor-A and of room-123, but at the individual level, allowing us to discriminate which part of corridor-A are we visiting. As this requires an extra cost and thus, the main contribution of this work is to provide effective and efficient segmentation and matching algorithms to that purpose.

Our segmentation module, described in Sect.2, relies on the Asymetric Clustering Model (ACM) proposed by Hoffman and Puzicha [8] [9], a distributional strategy that outperforms the classical K-means approach. We extend this model by making it adaptive, that is, able of identifying the optimal number of texture+color classes, and adaptivity is facilitated by the EM nature of the approach [10]. On the other hand, our proposal to region-matching is to embody the quadratic cost function proposed by Gold and Rangarajan in their Graduated Assignment approach [11], in the Comb Algorithm, a random search method proposed by Li [12], and adapt it to ensure matching constraints. In Sect.3 we present several recognition results that support our initial claim about class and individual performance. Our conclusions and future work issues are summarized in Sect.4.

2 Unsupervised Segmentation

2.1 EM Algorithm for Asymetric Clustering

Given N image blocks $x_1, \ldots, x_N$, each one having associated M possible features $y_1, \ldots, y_M$, the Asymetric Clustering Model (ACM) maximizes the log-likelihood

$$L(I,q) = -\sum_{i=1}^{N}\sum_{\alpha=1}^{K} I_{i\alpha} KL(p_{j|i}, q_{j|\alpha}) \ , \tag{1}$$

where: $p_{j|i}$ encodes the individual histogram, that is, the empirical probability of observing each feature y_j given x_i; $q_{j|\alpha}$ is the prototypical histogram associated to one of the K classes c_α; $KL(p_{j|i}, q_{j|\alpha})$ is the symmetric Kullback-Leibler divergence between the individual and the prototypical histograms; and $I_{i\alpha} \in \{0,1\}$ are class-membership variables.

As $p_{j|i}$ are fixed, one must find both the most likely prototypical histograms $q_{j|\alpha}$ and membership variables $I_{i\alpha}$. Prototypical histograms are built on all individual histograms assigned to each class, but such an assignment depends on the membership variables. Following the EM-approach proposed by Hoffman and Puzicha, in which the class-memberships are hidden or unobserved variables, we start by providing good initial estimations of both the prototypes and the memberships, feeding with them an iterative process in which we alternate the estimation of expected memberships with the re-estimation of the prototypes.

Initialization. Initial prototypes are selected by a greedy procedure: First prototype is assumed to be a block selected randomly, and the following ones are the most distant blocks from any of the yet selected prototypes. Given these initial prototypes $\hat{q}^0_{j|\alpha}$, initial memberships $\hat{I}^0_{i\alpha}$ are selected as follows:

$$\hat{I}^0_{i\alpha} = \begin{cases} 1 & \text{if } \alpha = arg\min_\beta KL(p_{j|i}, \hat{q}^0_{j|\beta}) \\ 0 & \text{otherwise} \end{cases} .$$

E-step. Consists of estimating the expected membership variables $\hat{I}_{i\alpha} \in [0,1]$ given the current estimation of the prototypical histogram $q_{j|\alpha}$:

$$\hat{I}_{i\alpha}^{t+1} = \frac{\hat{\rho}_{\alpha}^{t} \exp\{-KL(p_{j|i}, \hat{q}_{i|\alpha})/T\}}{\sum_{\beta=1}^{K} \hat{\rho}_{\beta}^{t} \exp\{-KL(p_{j|i}, \hat{q}_{i|\beta})/T\}} \text{ , being } \hat{\rho}_{\alpha}^{t} = \frac{1}{N} \sum_{i=1}^{N} \hat{I}_{i\alpha}^{t} \ , \qquad (2)$$

that is, the probability of assigning any block x_i to class c_α at iteration t, and T the temperature, a control parameter which is reduced at each iteration (we are using the deterministic annealing version of the E-step, because it is less prone to local maxima than the un-annealed one).

M-step. Given the expected membership variables $\hat{I}_{i\alpha}^{t+1}$, the prototypical histograms are re-estimated as follows:

$$\hat{q}_{j|\alpha}^{t+1} = \sum_{i=1}^{N} \pi_{i\gamma} p_{j|i} \text{ , where } \pi_{i\alpha} = \frac{\hat{I}_{i\alpha}^{t}}{\sum_{k=1}^{N} \hat{I}_{k\alpha}^{t}} \ , \qquad (3)$$

that is, the prototype consists of the linear combination of all individuals $p_{j|i}$. The weights of such a combination are the ratios $\pi_{i\alpha}$ between the membership of each individual to c_α and the sum of all memberships to the same class. This is consistent with a distributional-clustering strategy.

Adaptation. Assuming that the iterative process is divided in epochs, our adaptation mechanism consists of starting by a high number of classes K_{max} and then reducing such a number, if proceeds, at the end of each epoch. At that moment we select the two closest prototypes $\hat{q}_{j|\alpha}$ and $\hat{q}_{j|\beta}$ as candidates to be fused, and we compute h_α the heterogeneity of c_α

$$h_{\alpha} = \sum_{i=1}^{N} KL(p_{j|i}, q_{j|\alpha}) \pi_{i\alpha} \ , \qquad (4)$$

obtaining h_β in the same way. Then, we compute the fused prototype $\hat{q}_{j|\gamma}$ by applying Equation 3 and considering that $I_{i\gamma} = I_{i\alpha} + I_{i\beta}$, that is

$$\hat{q}_{j|\gamma} = \sum_{i=1}^{N} \pi_{i\gamma} p_{j|i} \ . \qquad (5)$$

Finally, we fuse c_α and c_β whenever $h_\gamma < (h_\alpha + h_\beta)\mu$, where $\mu \in [0,1]$ is a merge factor addressed to facilitate class fusion. After such a decision a new epoch begins. A minimal number of iterations per epoch are needed to reach a stable partial solution before trying two fuse two other classes.

2.2 Segmentation Results

Considering indoor images of 320×240 pixels, the feature extraction step consists on recovering texture and color statistics at blocks of size 32×32, that is, of radius 16 pixels. These blocks are taken each 8 pixels, that is, there is a partial overlap of 25%, providing $N = 37 \times 27 = 999$ blocks per image. Texture features rely on 8 Gabor filters with 4 orientations ($0, 45, 90$, and 135 degrees) and 2 scales ($\sigma = 1.0$ and $\sigma = 2.0$, corresponding to 7×7 and 13×13 windows respectively). Filter-output frequencies associated to each filter are registered in histograms of 16 equally spaced bins. Thus, there are $8 \times 16 = 128$ texture features per block, which are completed with 16 more features provided by the histogram associated to the first HSB color component (hue or chromaticity) inside the block. Consequently, the overall number of features is 144.

The unsupervised clustering algorithm proceeds through 10 epochs of 10 iterations each (100 iterations). Temperature range is fixed to $[1.0 \ldots 0.05]$, that is, each iteration t, T value is $0.095 + e^{-t} + 0.05$. On the other hand, the merge factor μ is set to 0.8, and to $K_{max} = 10$. In Fig. 1 we compare the segmentation results obtained with and without adaptation. After clustering we proceed to group neighboring blocks belonging to the same class in homogeneous regions. Small regions (those with less than $\nu = 20$ blocks) are removed and absorbed by the more similar region in its neighborhood).

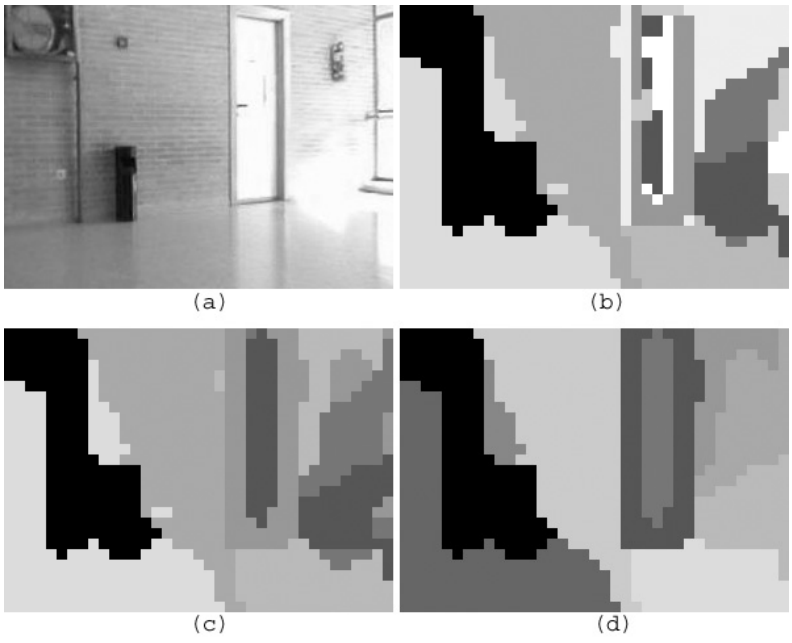

Fig. 1. *Segmentation results.* (a) Input indoor image.(b) Non-adaptive segmentation. (c) Adaptive segmentation. (d) After removing spurious blocks in (c).

3 Graph Matching

3.1 Stochastic Search for Assignement

Given an input segmented image we build an undirected data graph $G_D = (V_D, E_D)$ with one vertex $a \in V_D$ per region and one edge $(a, b) \in E_D$ per pair of adjacent regions. Similarly, we consider a stored graph $G_S = (V_S, E_S)$ with vertexes $i \in V_S$ and edges $(i, j) \in E_D$. Then, the adjacency matrices D and S of both graphs are defined by

$$D_{ab} = \begin{cases} 1 & \text{if } (a,b) \in E_D \\ 0 & \text{otherwise} \end{cases} , \text{ and } S_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E_S \\ 0 & \text{otherwise .} \end{cases}$$

A feasible solution to the graph matching problem between G_D and G_S is encoded by a matrix M of size $|V_D| \times |V_S|$ with binary variables

$$M_{ai} = \begin{cases} 1 & \text{if } a \in V_D \text{ matches } i \in V_S \\ 0 & \text{otherwise} \end{cases} ,$$

satisfying the constraints defined respectively over the rows and columns of M

$$\sum_{i=1}^{|V_S|} M_{ai} \leq 1, \forall a \text{ and } \sum_{a=1}^{|V_D|} M_{ai} \leq 1, \forall i \, . \tag{6}$$

Cost Function. Gold and Rangarajan formulated the problem in terms of finding the feasible solution M that maximizes the following cost function,

$$F(M) = \sum_{a=1}^{|V_D|} \sum_{i=1}^{|V_S|} \sum_{b=1}^{|V_D|} \sum_{j=1}^{|V_S|} M_{ai} M_{bj} C_{aibj} \, , \tag{7}$$

where $C_{aibj} = D_{ab} S_{ij}$, that is, when $a \in V_D$ matches $i \in V_S$, and also $b \in V_D$ matches $j \in V_S$, it is desirable that edges $(a, i) \in E_D$ and $(b, j) \in E_S$ exist, that is, that $M_{ai} = M_{bj} = 1$. However, this cost only encodes structural compatibility between both graphs. In order to enforce the preference of matching vertexes (regions) with compatible features (texture and color) we redefine C_{aibj} as

$$C_{aibj} = D_{ab} S_{ij} \exp\{-KL(q_a, q_i)\} \, , \tag{8}$$

where $q_a = q_{j|\alpha(a)}$ and $q_i = q_{j|\alpha(i)}$ are respectively the prototypical histograms of the classes of vertexes a and i.

Constrained Maximization. Gold and Rangarajan's deterministic annealing algorithm proceeds by estimating the averaged matching variables at each temperature T, while enforcing the satisfaction of matching constraints for the rows and columns of the candidate solution. Our preliminary matching experiments

with this method showed that it assigns each vertex with another one with similar structure but this does not usually ensures that the mapping is globally consistent. This is why we replaced annealing phase by a global strategy, a modified Comb (Common structure of the best local maxima) algorithm, originally applied to labeling problems in MRF models, which explores the set of extended feasible solutions. An extended feasible solution is a matching matrix $\hat{M}$ with one more row and one more column, corresponding to slack variables (which are very useful to deal with noisy nodes), whose rows and columns add up to the unit, that is, a permutation matrix of binary variables.

The Comb algorithm maintains a population $P = \{\hat{M}^{(1)}, \ldots, \hat{M}^{(L)}\}$ with the L (experimentally set to 10 individuals) best local maxima found so far. Such a population is initialized according to an uniform distribution over the space of feasible solutions. Each iteration begins by selecting, also randomly, a pair of local maxima $\hat{M}^{(a)}$ and $\hat{M}^{(b)}$. As this method relies on the assumption that local maxima share some matching variables with the global maxima, it derives a new candidate to local maximum $\hat{M}^{(0)}$ by combining the latter pair. Such a combination consists of (i) retaining common variables, (ii) randomly generating new values for components with different variables and (iii) ensuring that the result is still a permutation matrix. This provides the starting point of a hill-climbing process which consists of randomly changing the value at a component while ensuring that the resulting matrix satisfies the matching constraints and then testing whether it provides a better solution. If so, a new hill-climbing step begins. Otherwise, if after $A = 10$ attempts it is not possible to improve the current matrix a new local maximum $\hat{M}^*$ is declared. Such a local maximum updates P as follows: If

$$F(\hat{M}^*) > \hat{M}^{worst} \text{ where } \hat{M}^{worst} = \arg\min_{\hat{M} \in P} F(\hat{M}) \,, \tag{9}$$

then the worst local maximum so far $\hat{M}^{worst}$ is replaced by $\hat{M}^*$. Otherwise the population does not change. Such an updating rule ensures that the quality of the individuals in P is improved, expecting that such an improvement eventually reaches the global maximum. Thus, if we detect that the quality of P can not be improved we assume that the algorithm has found the global maximum (the best local maximum so far). The algorithm also terminates when the latter termination condition is not satisfied after $I = 1000$ iterations.

3.2 Recognition Results

In order to test the adequacy of our approach in scene recognition, we have build two subjective classes of images, each one registering different viewpoints of two different places (natural landmarks) in our lab (see Fig. 2): class-A (images A1,A2,A3, and A4) and class-B (B1 and B2). Sample matching results between images of the same class (A1 and A2) and of different classes (A1 and B1) are showed in Fig. 3. The main question addressed here is whether these classes are not subjective but real classes. In Table 1 we show the best costs for each

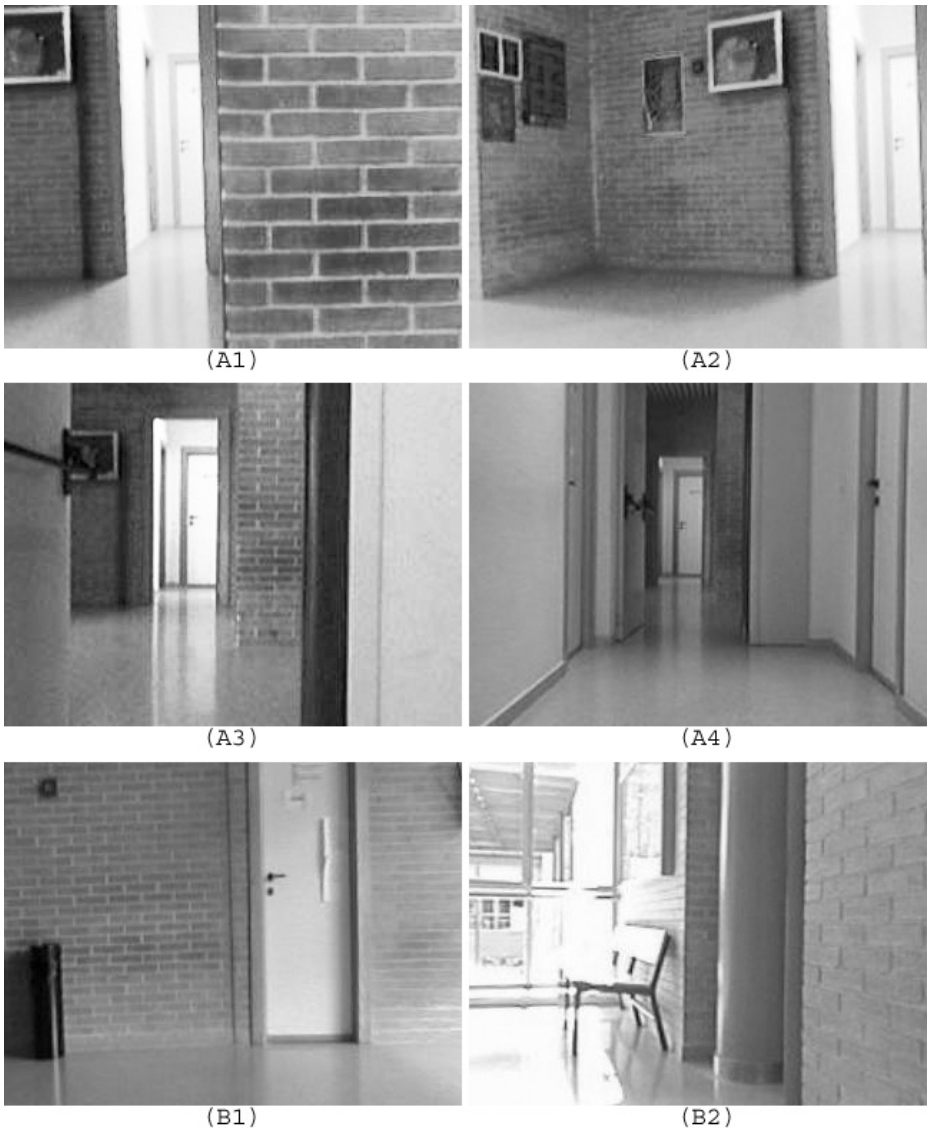

Fig. 2. Experimental set. Images A1 (first reference), A2 (15-degrees-rotation), A3 (2-meters-backwards), A4 (4-meters-backwards), B1 (second reference) and B2 (90-degrees-rotation

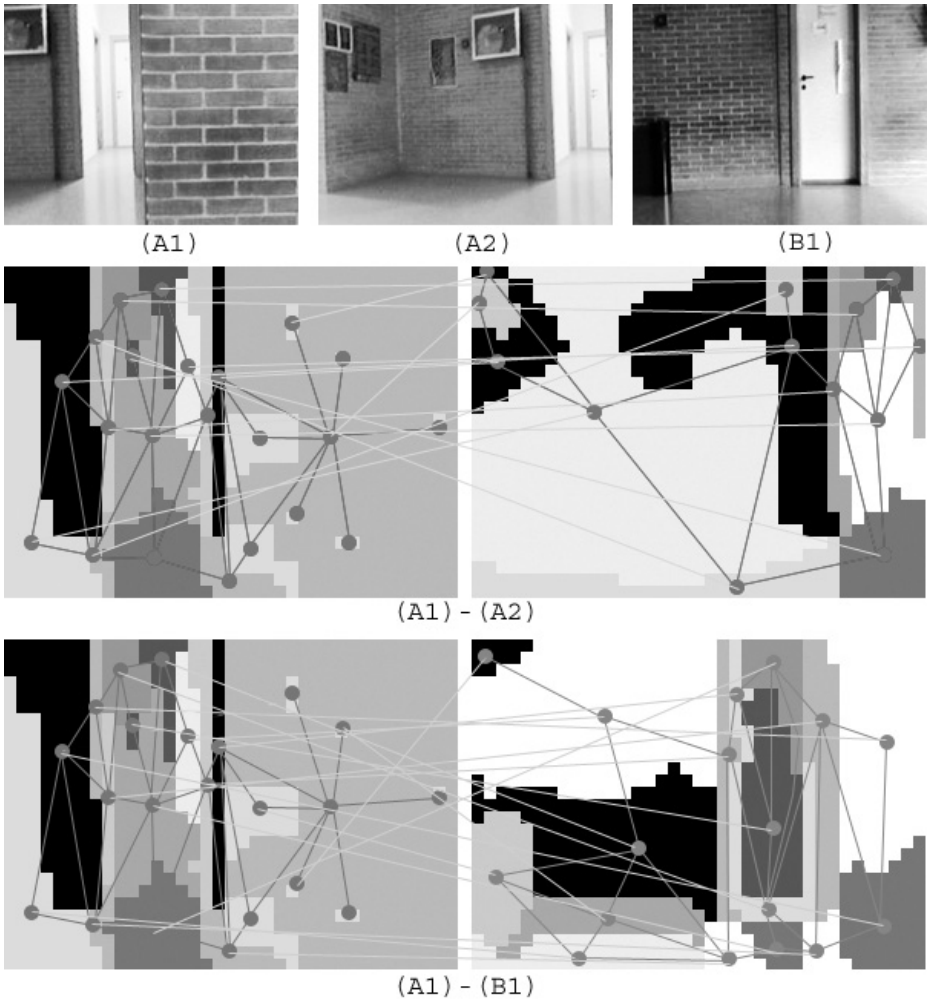

Fig. 3. Matching results. Top: Images A1 , A2, and B1; Middle: Matching between A1 and A2; Bottom: Matching between A1 and B1.

matching. Self matchings appear in boldface, matching between class-A images are emphasized, whereas matchings between class-B images, and between class-A and class-B images appear in normal text. We also show the a sort list of the three preferred matchings for each image (row in the matrix). Each image not only prefers itself, as expected, but its second and third choices are images in the same subjective class, when possible). In the case of B1 and B2, their third choice is A4, the more distant image from the first reference A1. Furthermore, we also show the degrees of ambiguity of the first and second matchings (ratios between the best costs of the second and the first matching, and between the best costs of the third and the second ones, respectively) and these degrees tend to be low at least for the winner matching.

Such a good performance is due to the co-occurrence of structural and appearance information between viewpoints of the same landmark. However, when we relax such a constraint and evaluate each matching only on behalf of structural compatibility, classes A and B are no longer distinct. In Table 2 we see that in many cases a given image does not prefer itself or even an image of the same subjective class. Furthermore, the analysis of the degrees of ambiguity reveals that the highest ambiguity in the latter case (0.62) is even lower that the current lowest ambiguity, reaching even 1.0 in the case of B2.

The averaged segmentation time was of 7.1 secs. in an ATHLON-XP-1700 processor, whereas the averaged processing time of the graph-matching step was of 4.8 secs, given an averaged size of 22 nodes per graph.

Table 1. Cost matrix when fusing structure and appearance.

	A1	A2	A3	A4	B1	B2	Sorted Preferences	Ambiguities
A1	**1.55**	*0.96*	*0.42*	*0.15*	0.03	0.01	A1, A2, A3	0.62, 0.44
A2	*1.07*	**2.92**	*0.49*	*0.20*	0.02	0.01	A2, A1, A3	0.37, 0.46
A3	*0.36*	*0.49*	**2.66**	*0.23*	0.02	0.03	A3, A2, A1	0.18, 0.73
A4	*0.20*	*0.23*	*0.29*	**1.47**	0.03	0.19	A4, A3, A2	0.20, 0.79
B1	0.03	0.03	0.03	0.04	**3.06**	0.68	B1, B2, A4	0.22, 0.06
B2	0.01	0.01	0.03	0.19	0.68	**2.00**	B2, B1, A4	0.34, 0.28

4 Conclusions

There are two main contribution in this paper: the adaptation and integration of state-of-the-art algorithms for unsupervised clustering and graph matching to the context of scene recognition, and the finding that this framework provides promising results for addressing the appearance-based localization problem in indoor environments. Future work includes the automatic inference of visual landmarks in the environment as well as the development of incremental localization algorithms.

Table 2. Cost matrix when using only structural information.

	A1	A2	A3	A4	B1	B2	Sorted Preferences	Ambiguities
A1	**1.73**	*1.69*	*1.66*	*1.30*	1.76	2.00	B2, B1, A1	0.88, 0.98
A2	*1.85*	**2.92**	*1.33*	*1.69*	1.24	1.00	A2, A1, A3	0.63, 0.72
A3	*1.50*	*1.16*	**2.66**	*0.83*	1.66	2.00	A3, B2, B1	0.75, 0.83
A4	*1.30*	*1.24*	*1.83*	**1.80**	1.29	2.00	B2, A3, A4	0.92, 0.98
B1	1.65	1.54	1.66	1.65	**3.41**	2.00	B1, B2, A3	0.59, 0.83
B2	2.00	1.00	2.00	2.00	2.00	**2.00**	B1, B2, A3	1.00, 1.00

References

1. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In Proc. IEEE Int. Workshop on Content-based Access of Image adn Video Databases (1998)
2. Torralba, A., Sinha, P.: Recognizing Indoor Scenes. AI Memo 2001-015, CBCL Memo 202, Cambridge, MA (2001)
3. Vailaya, A., Figueiredo, M.A.T., Jain, A.K., Zhang, H.-J.: Image Classification for Content-Based Indexing. IEEE Trans. on Image Processing, Vol. 10, No. 1 (2001) 117-130.
4. Kroese, B.J.A., Vlassis, N., Bunschoten, R., Motomura, Y.: A Probabilistic Model for Appearance-based Robot Localization. Image and Vision Computing, Vol. 19, No. 6 (2001) 381-391.
5. Lipson, P., Grimson, E., Sihna,P.: Configuration Based Scene Classification in Image Indexing. In Proc. IEEE CS Conference on Computer Vision and Pattern Recognition. Puerto Rico (1997) 1007-1013
6. Huet, B., Hancock, E.R.: Relational Object Recognition from Large Structural Libraries. Pattern Recognition (2002)
7. Carson, C., Belongie, S., Greenspan, H., Malik, J. : Blobworld: Image segmentation using Expectation-Maximization and its application to image querying. IEEE Trans. on Pattern Analysis and Machine Intelligence (2002)
8. Hofmann, T., Puzicha, J.: Statistical Models for Co-occurrence Data. MIT AI-Memo 1625 Cambridge, MA (1998)
9. Puzicha, J.: Histogram Clustering for Unsupervised Segmentation and Image Retrieval. Pattern Recognition Letters, 20, (1999) 899-909.
10. Figueiredo, M.A.T, Leitao, J.M.N, Jain, A.K.: On Fitting Mixture Models. In: Hancock, E.R., Pelillo, M. (eds.): Energy Minimization Methods in Computer Vision and Pattern Recognition. Lecture Notes in Computer Science, Vol. 1654. Springer-Verlag, Berlin Heidelberg New York (1999) 54-69.
11. Gold, S., Rangarajan, A.: A Graduated Assignement Algorithm for Graph Matching. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 18, No. 4 (1996) 377-388.
12. Li, S.Z.: Toward Global Solution to MAP Image Estimation: Using Common Structure of Local Solutions. In: Pelillo, M., Hancock, E.R.(eds.): Energy Minimization Methods in Computer Vision and Pattern Recognition. Lecture Notes in Computer Science, Vol. 1223. Springer-Verlag, Berlin Heidelberg New York (1997) 361-374.

SUMMARY INFORMATION
IBERAMIA-2002 Summary Submission

Title: Recognizing Indoor Images with Unsupervised Segmentation and Graph Matching

---

Authors: Miguel Angel Lozano, Francisco Escolano

Miguel Angel Lozano
Dto. Ciencia de la Computacion e Intelignencia Artificial
Unversidad de Alicante
Ap. Correos, 99 E03330 Alicante Spain
Tel: +34 965 90 39 00 Fax: +34 965 90 39 02
Email: malozano@dccia.ua.es

Francisco Escolano
Dto. Ciencia de la Computacion e Intelignencia Artificial
Unversidad de Alicante
Ap. Correos, 99 E03330 Alicante Spain
Tel: +34 965 90 38 97 Fax: +34 965 90 39 02
Email: sco@dccia.ua.es

Abstract: In this paper we address the problem of recognizing scenes by performing unsupervised segmentation followed by matching the resulting adjacency region graph. Our segmentation method is an adaptive extension of the Asymetric Clustering Model, a distributional clustering method based on the EM algorithm, whereas our matching proposal consists of embodying the Graduated Assignement cost function in a Comb Algorithm modified to perform constrained optimization in a discrete space. We present both segmentation and matching results that support our initial claim indicating that such an strategy provides both class discrimination and individual-within-a-class discrimination in indoor images which usually exhibit a high degree of perceptual ambiguity.

Keywords: Scene Recognition, Visual Localization, Unsupervised Segmentation, Graph Matching

---

Topics: Robotics, Perception, Artificial Vision

Section: Paper Track
