2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

Outdoor Landmark-view Recognition Based on Bipartite-graph Matching and Logistic Regression

Eduardo Todt, Carme Torras

Abstract— **This paper describes the extraction of visual landmarks from outdoor images for mobile robot applications. The concept of group of landmarks, called landmark-view, is introduced, aggregating the most relevant landmarks present in each scene. The relevance of the landmarks is determined by their relative visual saliency. Thus, landmark co-occurrence and spatial and saliency relationships between them are added to the single landmark descriptors, which are based on saliency and color distribution in chromaticity space. A suitable framework to compare landmark-views is developed, and it is shown how this remarkably enhances the recognition performance, compared against the single landmark recognition. A view-matching model is constructed using logistic regression. Experimentation using 45 views, acquired outdoors, containing 273 landmarks, yielded good recognition results. Of the 42 corresponding view pairs, 30 were recognized correctly, resulting in 71.4% of correct classification of similar views. Of the 948 non-corresponding view pairs, 768 were recognized correctly, resulting in 81.0% of correct classification in non-similar views. The overall percentage of correct view classification obtained was 80.6%, indicating the convenience of the approach.**

I. Introduction

THE extraction of reliable visual landmarks for mobile robot localization in unknown outdoor unstructured environments is still an open research problem. One of the key factors that makes the detection and recognition of visual landmarks in outdoor environments a challenging task is that acquired visual information is strongly dependent on lighting geometry (direction and intensity of light source) and illuminant color (spectral power distribution), which change with sun position and atmospheric conditions.

Most feature extraction approaches are not adequate for this type of environments, since they rely on either structured information from non-deformable objects [5], or on a priori knowledge about the landmarks [2]. There are recent works about using SIFT features to match pairs of images [11, 10] with interesting results. Since mobile robot navigation tasks require real-time execution, some efforts have been made to reduce the considerable computational effort necessary to evaluate SIFT features for a whole image [9].

In the present work we propose the concept of *landmark-view*, aggregating the most salient landmarks present in each image. A suitable framework to compare views is developed, and it is shown how this remarkably enhances the recognition performance.

Manuscript received September 15, 2006. This work was supported in part by the EU PACO-PLUS project FP6-2004-IST-4-27657 (Spain) and by PUCRS (Brazil).

E. T. is with the Faculty of Engineering, PUCRS, Porto Alegre, Brazil (e-mail: todt@ ieee.org).

C. T. is with the Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain (e-mail: torras@iri.upc.edu).

II. Looking for Landmarks in the Scenes

The first step to recognize visual landmarks is to locate the candidate landmarks in the color images (512 x 512 pixels, 24 bits/pixel) acquired by the mobile robot. The candidate landmarks are image regions selected according to their visual saliency, inspired on a biological model of visual attention [8]. The *color-ratios saliency* algorithm [15] embeds color constancy within the saliency computation, counterbalancing the intrinsic variations of illumination outdoors, which could affect the color perception and, subsequently, the saliency results. In the following, this saliency algorithm is described shortly.

A region in an image is considered salient if it ranks high in a given feature and its surround ranks high in the opposite feature. Here, the features considered are the opponent colors (red-green and blue-yellow), because they are the most stable features of the original visual saliency model (color, intensity and orientation) when the scenes are subject to illumination changes. From the input image, two Gaussian pyramids, corresponding to each feature in logarithmic space, are constructed. In each pyramid, a pixel at a fine scale corresponds to a center region, whereas the respective pixel at a coarser scale corresponds to its surround. The ratios between features at different pyramid levels correspond to the computation of the center-surround saliencies and give the corresponding partial saliency maps. The pyramid levels used to compute each partial saliency map define the spatial scale of the salient elements detected.

Two sets of partial saliency maps are constructed, corresponding to the color features at several spatial center-surround scales.

Finally, the partial saliency maps are combined, taking into account their information content, to compose a global saliency map.

III. Delimiting Landmark Regions

Since the extracted salient regions are not necessarily bounded by well-defined contours, nor associated to single elements in the scenes, a refinement step is necessary in the process of determining the boundaries of landmark candidates.

As an initial approximation (Fig. 1), a minimal rectangular bounding box is computed for each segmented saliency spot. Very small bounding boxes (in the current implementation, the minimum size is fixed to 64 pixels) are discarded, because the low pixel count does not allow making reliable assumptions about the detected saliency. The bounding boxes correspond to the first representation of the landmarks detected with the visual saliency method. Due to the sensitivity of saliency to the surrounding information and shadowing, the spatial distribution of saliency can change significantly in images taken from the same scene under different conditions. The objective of the next two processing steps is to adjust the bounding box size and position, getting a better fitting to the detected salient elements.

In the next step, for each bounding box a chromaticity histogram is computed and the image is submitted a histogram backprojection processing [13], emphasizing where the same colors appear in the whole image.

After this, the size and position of all bounding boxes are adjusted, taking into account the color spatial distribution obtained with backprojection. This is achieved using the continuously adaptive mean shift algorithm [4]. This is a non-parametric technique that climbs the gradient of a probability distribution to find the nearest dominant mode, with the capability to adapt the window size. To increase the amount of information associated with the bounding boxes, their immediate surrounding region is also analyzed (Fig. 1), giving additional context information to the recognition.

IV. Landmark Characterization and Matching

After the determination of the bounding boxes, the following region descriptors are extracted:

1. Normalized chromaticity histogram of salient region inside bounding box.
2. Normalized chromaticity histogram of fitted bounding box.
3. Normalized chromaticity histogram of expanded bounding box.
4. Mean saliency of fitted bounding box.

The similarity between the histogram descriptors (1,2 and 3 above) of two image regions i and j is measured by the distance between their corresponding points h_i and h_j in histogram space [13]. The quadratic form metric [7] is used:

$$d^2_{hist}(h_i, h_j) = (h_i - h_j)^T \mathbf{A} \, (h_i - h_j) \qquad (1)$$

where h_i and h_j are n-dimensional color histograms, and $\mathbf{A}$ is the similarity matrix, whose elements a_{kl} denote similarity between bins k and l. This metric was selected because it allows for similarity matching between different colors, while other histogram metrics, like histogram intersection, just evaluate exact color matching.

The distances corresponding to the four descriptors are combined using the root of the sum of the squared distances, resulting in a single distance value between two landmarks.

V. Grouping Landmarks and Defining Views

The results of experiments about single landmark recognition, using the algorithms described in the preceding sections [16], indicate that the color and saliency descriptors defined don't have enough information content to ensure unambiguous recognition of single landmarks. In the following it is described how the recognition process can be improved with the concept of landmark-view.

The landmarks detected in the same scene are grouped, constituting landmark-views, and these views are compared with other views to recognize places already visited by the mobile robot, instead of comparing single landmarks. The grouping of landmarks combines the individual recognition evidences of the single landmarks detected in each observation, and adds the information on the relationship between landmarks.

A landmark-view is defined as the set of landmarks observed in one image captured by the robot in a specific spatial location and orientation. Thus, at each observation, instead of just trying to recognize isolated landmarks, their mutual spatial and saliency relationships are also taken into account, adding context information to the landmark recognition task. Consequently, the problem of landmark recognition is handled as a component of a higher-level problem, namely view recognition. In order to be able to recognize views, it is necessary to establish a distance metric to compare pairs of views.

VI. View Matching

Since views are defined as sets of landmarks, their similarity can be assessed by finding the optimal matching between the respective landmark sets. The idea is that corresponding views (images taken from similar robot location and orientation) should match better than non-corresponding views. The relative visual saliency of each landmark is used to select the most relevant landmarks and also is used as a feature in the matching process.

A powerful tool to model objects and relationships between them are graphs. They have been widely used in the fields of image analysis and image processing [4, 16]. In the following it is explained how a graph-matching algorithm can be applied to the view recognition problem [1]. A graph $G = (V, E)$ consists of a set of vertices $V = \{v_i\}$ and edges $E = \{e_i\}$. The edges are connections between vertices. Vertex v_j is adjacent to v_i if there is an edge $e = (v_i, v_j)$ between them. Two edges are adjacent if they have a common vertex. A matching is generally defined as a subset of the edges of a given graph such no two edges are adjacent. A particular case of matching is defined between two distinct vertex sets $U = \{u_i\}$ and $V = \{v_j\}$, thus assuming a bipartite graph $G = (U, V, E)$, where $E \subseteq U \times V$.

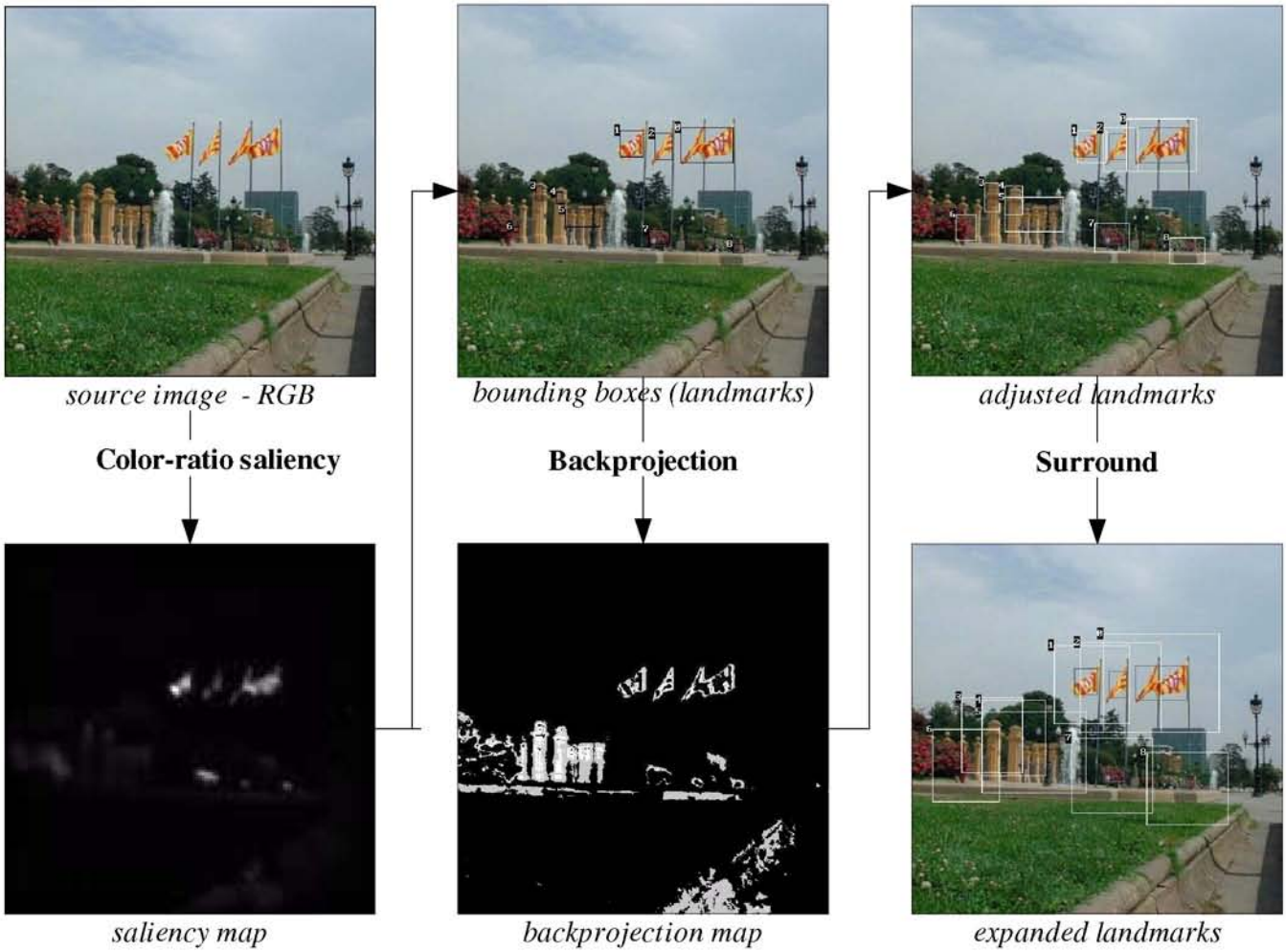

Fig. 1. The process of delimiting the landmark regions. From the source image a saliency map is computed, then this map is segmented, generating the seeds of the landmark regions. These seeds are enclosed by bounding boxes, which are fitted to the salient elements in the image using color histogram backprojection and mean-shift algorithms. Finally, the landmark bounding boxes are expanded, encompassing the immediate surrounding regions.

Disregarding the adjacency constraint, in a bipartite graph a match is any subset of edges of it:

$$M = \{m_i\} \subseteq E \tag{2}$$

The set of unmatched vertices is defined as:

$$S = \{s \mid s \in U, \text{ó} v : (s,v) \in M\} \cup \{s \mid s \in V, \text{ó} u : (u,s) \in M\} \tag{3}$$

There are several ways to match the vertices of U to those of V. A matching is maximal if the number of matched vertices is maximum. In the classical problem of bipartite matching, the objective is to find a maximal one-to-one matching. In a one-to-one matching,

$$\forall (u_i, v_j) \in M, \forall (u_k, v_l) \in M : (i = k) \Leftrightarrow (j = l) \tag{4}$$

The bipartite matching problem can involve the minimization of a cost function, taking into account the cost of the matching and penalizing for the unmatched vertices:

$$\text{cost}(M,S) = \sum_{m \in M} c(m) + \sum_{s \in S} c'(s) \tag{5}$$

where $c(m)$ with $m=(u,v)$ is the cost of matching u to v, and $c'(s)$ is the cost of leaving a vertex s unmatched.

The cost of matching two vertices can be defined through a metric distance between attributes of the respective vertices. Weights of the vertices, denoted by w, can be considered in addition to attributes, meaning the strength, activity, probability or significance of each vertex. When the costs and/or weights of the matching are considered, the problem is called *weighted bipartite matching* [1].

In a bipartite graph, the matching is done between two separate vertex sets, which have no internal structure. Both bipartite matching and weighted bipartite matching can be reduced to the more general maximum flow problem, which can be solved in polynomial time. The set U of vertices corresponds to the set of landmarks in one view, and the set V corresponds to the set of landmarks in the other view. The weight of each edge represents the similarity distance between the individual landmarks. The solution of the weighted bipartite matching defined by U and V gives the best matching between the landmarks and thus provides a measure of view similarity.

Among the several available algorithms to solve the

bipartite matching problem, the relaxation algorithm [3] was adopted because of its broad use, simplicity, and existence of successful reported experiences in the image-matching field. The algorithm consists of the following steps:

1. Initially, the matching restriction is relaxed, allowing any vertex in V to be assigned more than one vertex in U. Each vertex u_i in U is assigned to the vertex in V with the minimum matching cost among all edges.
2. The algorithm then iteratively selects an overassigned vertex v_k in V, obtains the shortest path from vertex v_k to all other unassigned vertices in V, considering each matching *cost* $c(u_i,v_j)$ reduced by the minimum matching cost from u_i to any $v_z \in V$, updates the assignments using the shortest path found, until there are no more overassigned vertices in V. The algorithm reaches optimality by executing a maximum of N iterations.

With a naive implementation of shortest path, the resulting computation complexity is $O(N^3)$, but it can be reduced using optimized shortest path search algorithms, for example, to $O(N^{2(1+logN)})$ using the Fibonacci heap method [1].

Thus, we compute the distance between two landmark-views according to the following steps:

1. In each view the k-most salient landmarks are selected.
2. A k x k matrix with the quadratic-form distances between all pairs of landmarks, one taken from each view, is computed. Note that, in addition to the four descriptors listed in the preceding section, the distance of each individual landmark to the centroid of the set of landmarks is considered as an additional descriptor.
3. The k landmarks of the two views are paired using the weighted bipartite matching algorithm, based on the quadratic-form distances between the landmarks.
4. The minimum assignment cost resulting from the weighted bipartite matching is taken as the distance between the two views.

The view with the lowest distance to a newly acquired view is considered the matching view. If no view has a distance to the query view below some threshold, then it is assumed that the query view is a new view in the system.

VII. A Statistical Model for View-Matching

In the landmark-view matching algorithm presented in the preceding section, the distances obtained from each descriptor (color histograms, mean saliency, distance to centroid) of the landmarks were just combined with a root mean of squares.

Here, we propose to use logistic regression [6] to evaluate the significance of each landmark descriptor and to build a statistical model for view and landmark recognition.

Logistic regression analysis evaluates the significance of each variable in a multivariable model whose output is a single binary variable. This variable has the semantics of a binary classifier based on the values of the input variables. We define a binary variable, named *view match* and denoted *VM*, which takes the value 0 when two landmark-views match, and 1 otherwise.

The input variables considered are the following:

- X_1:Salient region chromaticity histogram.
- X_2:Fitted salient region chromaticity histogram.
- X_3:Expanded salient region chromaticity histogram.
- X_4:Landmark saliency.
- X_5:Landmark distance to the centroid of the set of landmarks in the view.
- X_6:Combined sum of squares of the previous features.
- X_7, X_8:Non-assigned nodes in the weighted bipartite view matching.

The resulting model has the form:

$$VM = \frac{e^{A+B_1X_1+B_2X_2+B_3X_3+B_4X_4+B_5X_5+B_6X_6+B_7X_7+B_8X_8}}{1+e^{A+B_1X_1+B_2X_2+B_3X_3+B_4X_4+B_5X_5+B_6X_6+B_7X_7+B_8X_8}} \quad (6)$$

where A is a constant term, and B_i are the beta coefficients (see Table 1), outputs of the logistic regression carried out with a training set of data. Xi are the input variables. The training set of data consisted of a sample of outdoor images with 68 landmarks and 78 cases of possible view pairs [14].Table 1 presents the logistic regression results using these sample images. The regression was carried out in five steps, each one constituting a new model aggregating a new group of variables.

In the first step, just the *color* descriptors of the landmarks (salient region, fitted salient region and expanded salient region chromaticity histograms) were used. These variables explained 43.3% of the model variance (Nagelkerke R2 0.433). The model was able to classify correctly 82.4% of the matching view pairs and 75.4% of the non-matching view pairs (overall correct classification 76.9%). The significant color variable was the expanded salient region color ($p<0.05$).

It turned out that the *saliency* variable does not contribute to the model quality. Its introduction in step 2 did not improve the variance explained by the model, neither the

Table 1 Logistic regression of the "view match" variable

Independent variables	Step 1		Step 2		Step 3		Step 4		Step 5	
	beta	sig.	beta	sig.	beta	sig.	beta	sig.	beta	sig.
Salient Region Color	1.566	0.055	1.566	0.056	1.366	0.110	1.333	0.207	0.297	0.790
Fitted Salient Region Color	0.042	0.988	-0.024	0.993	0.809	0.792	0.748	0.819	3.867	0.300
Expanded Salient Region Color	13.981	0.001	13.693	0.001	15.952	0.001	15.937	0.001	18.775	0.001
Saliency			-1.738	0.836	-4.689	0.615	-4.700	0.614	-11.965	0.296
Distance to Centroid					-1.537	0.008	-1.593	0.184	-1.310	0.317
Combined Sum of Squares							0.159	0.958	1.905	0.561
NA1									1.177	0.003
NA2									1.177	0.003
Constant A	-1.491	0.004	-1.441	0.011	-0.787	0.204	-0.800	0.231	-3.488	0.002
% explained (Nagelkerke R2)	0.433		0.433		0.485		0.485		0.567	
% correct classification same view	82.4		82.4		76.5		76.5		82.4	
% correct classif. on different view	75.4		75.4		82.0		82.0		83.6	
Overall % correct classification	76.9		76.9		80.8		80.8		83.3	

classification scores. However, it is important to consider that the saliency was used to select the landmarks to be taken into account in the view-comparison process, thus it has an important indirect contribution to the classification result.

In step 3, the variable *distance to landmark centroid* was introduced. It improved the variance explained by the model and the classification scores. These variables together explained 48.5% of the model variance. The model was able to classify correctly 76.5% of the matching view pairs and 82.0% of the non-matching view pairs (overall correct classification 80.8%). The significant variables were the expanded salient region color ($p<0.01$) and distance to centroid ($p<0.1$).

In step 4, a *root mean square* of the previous features was considered. The model already included the variables involved in the computation of this variable, and so there were no changes in the model prediction performance.

In the last step, the variables *NA1* and *NA2*, corresponding to the cost of *non-assigned nodes* in the bipartite graph matching of the landmarks in the two views, were introduced. They improved considerably the variance explained by the model and the classification scores. These variables explained 56.7% of the model variance. The model was able to classify correctly 82.4% of the matching view pairs and 83.6% of the non-matching view pairs. The significant variables were the expanded salient region color ($p<0.01$), and the non-assigned nodes *NA1* and *NA2* ($p<0.05$). The overall correct prediction of matching was 83.3%. The *NA1* and *NA2* variables have the same significance, because they have the same semantics, i.e., the count of non-matched landmarks in each view. Since *NA1* and *NA2* carry implicit a direction of matching, in the regression analysis each pair of views was considered two times, inverting the query and database roles.

It is important to observe that the effect of introducing the variables in the regression model is not necessarily cumulative, regarding the significance of variables. The significance of a variable could be affected with the introduction of a new variable in the model, because the significance is computed in the context of that model.

The constant term *A* in equation (6) appears as significant because it is related to the part of the model that is not explained by the variables. All regression models were statistically significant, with $p<0.01$ in all steps. It can be observed that the most significant variables in the complete model were the expanded salient region color and the non-assigned nodes, which constitute a combination of color and spatial information. The initial model parameters were computed off-line, using the SPSS package [12], based on a sample image set, and the match score function was developed *ad hoc* and embedded in the view recognition system.

VIII. Experimental Results

To validate the landmark-based view recognition system, a university campus was chosen (PUCRS, Brazil) as a real outdoor environment. A set of 990 view pairs from 45 different views, with 273 landmarks, was analyzed (Fig. 2). An example of weighted bipartite matching for two similar views is shown in Fig. 3.

Fig. 2 Some images taken in the outdoor experiment.

Of the 42 corresponding view pairs, 30 were recognized correctly, resulting in 71.4% of correct classification of similar views. Of the 948 non-corresponding view pairs, 768 were recognized correctly, resulting in 81.0% of correct classification of non-similar views. The overall percentage of correct view classification was 80.6%.

Using a standard low-performance PC computer (Pentium III 900MHz, 256Mb DRAM, Microsoft Windows XP) the view matching was performed in 0.69 seconds.

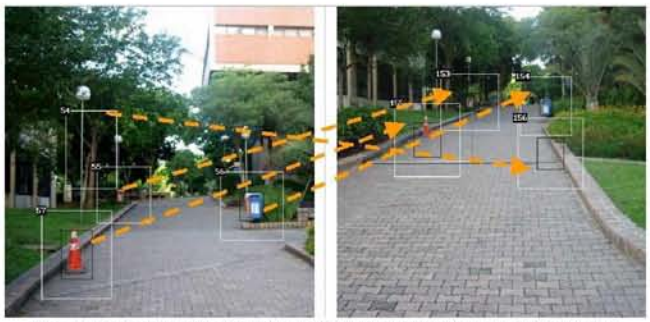

Fig. 3 Landmark matching at similar views. The arrows indicate the solution of the weighted bipartite matching

IX. Discussion

A noticeable increase of performance in correct classification was observed with the introduction of landmark-views in the landmark recognition process. The results were good, even in a real outdoors experiment subject to illumination effects, like highlights, shadows, and illumination changes present in this experimental sample.

This work contributes to the robot localization field by proposing a new procedure for visual saliency detection and characterization of candidate landmarks in scenes, as well as an application of logistic regression analysis to determine a suitable matching model. A binary function to compare a query view with each view in a database of previous views and to decide about the similarity between them was developed with the aid of logistic regression. Very good view discrimination ability was observed, with scores of correct classification that validate the concept of landmark-view, and the proposed view recognition procedure.

Logistic regression proved to be a powerful tool to build the matching model. Without it, on a trial-and-error basis, it was extremely difficult to compose the available information to decide the matching of views. The resulting model is simple and allows for the future incorporation of reinforcement mechanisms, through the continuous tuning of the model parameters as a background task.

The use of view descriptors aggregating co-occurrence and spatial relationships of landmarks significantly improved the recognition process, preserving the simplicity and low quantity of stored information.

Some lines of future research are envisaged. The first one is to reduce the search space for view matching by taking into account the recent history within a probabilistic approach. And the second, as mentioned above, is to endow views with a reinforcement strategy that would tune the descriptors each time a view is recognized. Finally, it could be interesting to use our saliency-based approach together with a SIFT-based engine, combining the good properties of both techniques.

Acknowledgment

The authors would like to thank Enric Celaya and Pablo Jimenez for productive discussions about visual saliency and robot localization, and Alessandra Bianchi for hints about SPSS.

References

[1] R. K. Ahuja, T. L. Magnanti, and J. Orlin, *Network flows*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[2] J. Batlle, A. Casals, J. Freixenet, and J. Martí, "A review on strategies for recognizing natural objects in colour images of outdoor scenes," *Image and Vision Computing*, vol. 18, pp. 515-530, 2000.

[3] S. Berretti, A. Bimbo, and E. Vicario, "Efficient matching and indexing of graph models in content-based retrieval," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 1089-1105, 2001.

[4] G. R. Bradski, "Computer vision face tracking for use in a perceptual user interface," Fourth IEEE Workshop on Applications of Computer Vision, pp. 214-219, 1998.

[5] W. Burgard, A. Derr, D. Fox, and A. B. Cremers, "Integrating global position estimation and position tracking for mobile robots: the dynamic Markov localization approach," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, Victoria, Canada, pp. 730-735, 1998.

[6] D. R. Cox and E. J. Snell, *Analysis of binary data*, 2nd. edition ed. London: Chapman & Hall, 1989.

[7] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 729-736, 1995.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254-1259, 1998.

[9] I. Kirigin and S. Singh, "Bearings based robot homing with robust landmark matching and limited horizon view," Robotics Institute, Carnegie Mellon Univ., Pittsburgh, Tech. Rep. CMU-RI-TR-05-02, 2005.

[10] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1615-1630, 2005.

[11] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *Int. Journal of Robotics Research*, vol. 21, pp. 735-758, 2002.

[12] SPSS, "Statistical Package for the Social Sciences," v.10.0.5, Chicago, Illinois: SPSS Inc., 2000.

[13] M. J. Swain and D. H. Ballard, "Color indexing," *Int .Journal of Computer Vision*, vol. 7, pp. 11-32, 1991.

[14] E. Todt, "Visual landmark detection for navigation in outdoor environments," *Institut d'Organització i Control de Sistemes Industrials*. Barcelona: Universitat Politècnica de Catalunya, 2005, pp. 200.

[15] E. Todt and C. Torras, "Detecting salient cues through illumination-invariant color ratios," *Robotics and Autonomous Systems*, vol. 48, pp. 111-130, 2004.

[16] E. Todt and C. Torras, "Color-contrast landmark detection and encoding in outdoor images," The 11th Int. Conf. on Computer Analysis of Images and Patterns, Versailles, France, pp. 612-619, 2005.