OUTDOOR VIEW RECOGNITION BASED ON LANDMARK GROUPING AND LOGISTIC REGRESSION

EDUARDO TODT

Federal University of Parana - UFPR,
R. Francisco Hoffmann dos Santos 100, 81531-980 Curitiba, Brazil
todt@ieee.org

CARME TORRAS
Institut de Robòtica i Informàtica Industrial, CSIC-UPC,
Llorens i Artigas 4-6, 08028 Barcelona, Spain
torras@iri.upc.edu

Vision-based robot localization outdoors has remained more elusive than its indoors counterpart. Drastic illumination changes and the scarceness of suitable landmarks are the main difficulties. This paper attempts to surmount them by deviating from the main trend of using local features. Instead, a global descriptor called *landmark-view* is defined, which aggregates the most visually-salient landmarks present in each scene. Thus, landmark co-occurrence and spatial and saliency relationships between them are added to the single landmark characterization, based on saliency and color distribution. A suitable framework to compare landmark-views is developed, and it is shown how this remarkably enhances the recognition performance, compared against 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. The overall percentage of correct view classification obtained was 80.6%, indicating the adequacy of the approach.

Keywords: visual landmarks, visual saliency, robot navigation, autonomous robot

1 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, as well as indoors without dominant artificial illumination, 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 [7], or on a priori knowledge about the landmarks [3]. Several recent works achieved interesting results using SIFT features to match pairs of images [18, 22, 32], which can be extended to the landmark recognition problem. Since mobile robot navigation tasks require real-time execution, some efforts have been made to reduce the considerable computational cost necessary to evaluate SIFT features for a whole image [17, 28]. Also it was reported that SIFT features fail to consider global context to resolve ambiguities that can occur locally in images, motivating solutions that improve the amount of global information used in the descriptors [10, 24]. More recently SURF features have been introduced, giving comparable results to SIFT but with lower computational cost [4, 40].

Nowadays most image retrieval and image recognition systems are based on interest point detectors that select the features in the images to be characterized by descriptors. Mikolajczyk and Schmid [23] and Schmid, Mohr, and Bauckhage [31] compare several interest point detectors. Finally, Antani, Kasturi and Jain [2] and Li and Allison [21] present excelent surveys of pattern recognition methods and features used for image retrieval and computer vision.

In this context, the present work relies on the concept of *landmark-view*, based on a group description of the most salient landmarks present in each image. The landmark-view combines the compactness of landmark representation with the global information of view-based approaches, here provided by the spatial and saliency relationships between landmarks. The matching scores between landmark-views are evaluated with the use of logistic regression. The definition of landmark-views and their matching using logistic regression is the main contribution of this work.

A suitable framework to compare views is developed, and it is shown how this remarkably enhances the recognition performance.

The remainder of the paper is organized as follows. Section 2 presents the concept of landmark and visual saliency. In this context, landmarks are the visual elements that are used to provide the features for the recognition of places, and they are found or defined based on classical visual saliency criteria, inspired in a biological model of visual opponency. Section 3 describes how to refine the coarse landmark regions found in the previous step, based on visual saliency. Histogram backprojection and mean-shift algorithms are used to expand the spots of salient regions, taken as seeds, to areas that correspond to elements, or part of them, in the scenes. Section 4 presents the landmark characterization, based on chromaticity histograms and relative saliency metrics, and the subsequent matching, based on a quadratic-form distance, that is more robust to small skew in the distributions than simple histogram distances. Section 5 describes the concept of *landmark-view*, simply denoted by *view*. A view is a group of co-occurrent landmarks, more robust to recognition than single landmarks. Also the view concept supports the insertion of spatial distribution and relative saliency metrics in the characterization. Section 6 presents the view matching, generalized from the landmark recognition. In Section 7 a statistical model for view matching is discussed, using logistic regression to evaluate the importance of each metric in the recognition process. Section 8 shows experimental results. Finally, discussion and conclusions are presented in Section 9.

2 Looking for Landmarks based on Visual Saliency

When addressing robot localization, Levitt and Lawton [19] were among the first to refer to research by cognitive psychologists showing that humans and animals record references and use the structure inherent in local and temporal relationships between these references to identify places in the world and to plan and execute paths between locations. This idea continues to be supported by recent research in this field [13, 20, 33], which encourages its use as inspiration for the development of robot localization algorithms.

These references are called *landmarks*, which in robotics are defined as distinctive entities that the robot can recognize whenever they are in their detection range [19, 36]. A *visual landmark* is a stationary

distinctive object or pattern that the robot can recognize with its vision system whenever it is in view.

If we want to recognize visual landmarks, the first task to be done is to locate candidate landmarks in the color images 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 [15]. Human vision and artificial vision have in common the challenge of reducing the amount of sensorial information to be processed in order to analyze a scene image, due to intrinsic limitations in bandwidth, memory, and computational speed. The most accepted models of the primate visual system [12, 39] consider the existence of an attention mechanism responsible for selecting the most relevant visual stimuli for further processing by the available resources, rather than attempting to fully interpret visual scenes in a parallel fashion. The attention mechanism is driven by the *visual saliency* of the scene elements, which refers to the idea that certain parts of a scene are distinctive and that they create some form of significant visual arousal at the early visual stages [16]. This mechanism is essentially data-driven, which is particularly useful in those situations where the semantics of the contents of the image is not known and models of the perceived objects are not available [30].

Light intensity contrast appears to be the primary variable on which humans base visual saliency computation, although other features participate in defining visual saliency at higher processing levels in the visual cortex [12, 25]. Among these are edge or line orientation, color, motion, and stereo. One major observation is that the relevant variable is not the amplitude of visual signals in a particular feature dimension, but the contrast between this amplitude at a given point and at the corresponding surrounding locations [41].

The visual saliency of elements in the images is detected with the color-ratios saliency algorithm [37], which has the interesting characteristic of embedding color constancy within the saliency computation. The color constancy counterbalances the intrinsic variations of illumination outdoors that can affect the color perception and, subsequently, the saliency results. In the following, this algorithm is described shortly.

Therefore, the notion of visual saliency relies on the previous notion of opponency. For example, a red roof is salient in a green landscape, but not if it is surrounded by similarly reddish walls and terraces. Likewise, a vertical pole is salient if it is in the middle of a horizontally stripped fence. Thus, 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 to compute the visual saliency are the opponent colors red-green and blue-yellow, because they are the most stable features of the visual saliency model when the scenes are subject to illumination changes [37]. From the input image, two Gaussian pyramids are constructed, each one corresponding to a color feature in logarithmic space. In the pyramid image structure, 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 at different spatial scales and give the corresponding partial saliency maps. Two sets of partial saliency maps are computed, corresponding to the red-green and blue-yellow color features at several combinations of spatial center-surround scales. In this case, the center images were taken at pyramid levels 2, 3, and 4, and surround images at pyramid levels 5, 6, and 7.

Using several scales, not only for center but also for surround, yields truly multiscale feature extraction, being possible to detect visual salient objects within a wide size range.

The resultant partial maps are combined into a global map, in which salient areas are indicated by large values, whereas non-salient areas have small values. The partial saliency maps cannot simply be added, because salient regions present in only a few maps can be masked by noise or less salient regions present in a larger number of maps. The process of combining the partial saliency maps is structured in three stages. In the first stage, the partial saliency maps are normalized by the maximum saliency value obtained at all center-surround scales. In the second stage, the maps are weighted by their information content. The information content of an image is based on their zero-order entropy. Finally, the partial saliency maps are subject to exponentiation and added to compose the global saliency map. To reduce computational costs, the saliency maps are represented at a scale that corresponds to the second level of the pyramids in the color ratio visual saliency algorithm, yielding 128x128 pixels images. Figure 1 shows some examples of saliency maps resulting from RGB images.

Figure 1. Input RGB images (left) and corresponding saliency maps (right). The brightest areas in saliency maps correspond to the most salient areas in the input images.

3 Refining Landmark Regions

Since the extracted salient regions obtained with the visual saliency algorithm, described in the previous section, are not necessarily bounded by well-defined contours, nor associated to single elements in the scenes, a refinement process is necessary to determine the boundaries of landmark candidates. Figure 2 shows an overview of all the steps taken starting from the input image to the delimitation of landmark areas.

The saliency map obtained with the color ratios algorithm has several salient spots. These spots

typically have diffuse borders, with saliency values decreasing from some local maximum to a background level of noisy saliency. Thus, a segmentation process is necessary to delimit the salient regions.

This is a non-trivial task, because segmentation in outdoor environments usually requires high-level information to validate the results [3]. Here, we assume that there is no high-level information available; thus the segmentation has to be done roughly, based only on the low-level features available. Since the saliency spots can have different peak values, the multilevel thresholding is a well-suited technique to segment the saliency map [27]. In this technique, each saliency spot corresponds to a seed that is expanded until some fraction of the local maximum is attained, constituting an adaptive region-growing segmentation.

As an initial approximation for landmark regions, a minimal rectangular bounding box is computed for each segmented saliency spot. Very small bounding boxes (for instance, in the current implementation the minimum area is set to 64 pixels) are discarded, because the low pixel count does not allow making reliable assumptions about the detected saliency.

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 to a histogram backprojection processing [35], emphasizing where the same colors appear in the whole image. Histogram backprojection identifies where, in some image, are the colors that belong to a target model being looked for. It is based on the ratio histogram, defined as

$$R[i] = \min(h_2[i] / h_1[i], 1), \quad i = 0,..,n-1 \qquad (1)$$

where h_1 is the image histogram, h_2 is the searched target histogram (obtained from the landmark bounding box), and n is the number of histogram bins. This ratio histogram is back-projected onto the image, that is, the image values are replaced by the values of R that they index. The values in the resultant image represent the expectation of the target location.

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 [6]. 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. In our case, this means that bounding boxes defined by the saliency spots are adjusted in size and location to the neighboring areas that have similar color distributions.

To increase the amount of information associated with the bounding boxes, their immediate surrounding region is also analyzed (Figure 2), giving additional context information to the recognition.

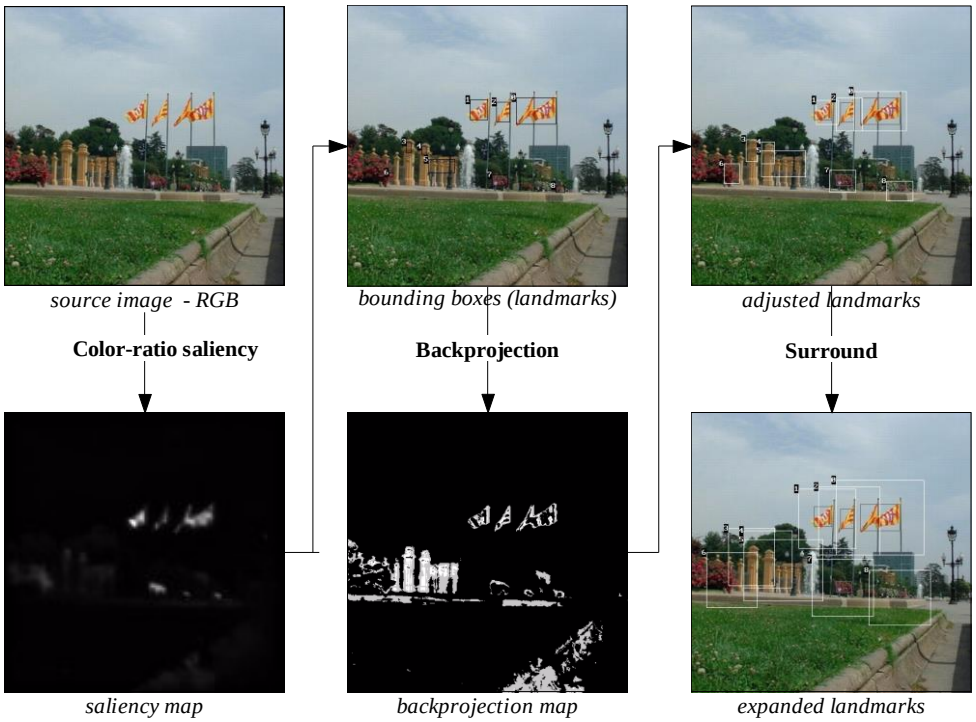

Figure 2. The process of delimiting the landmark regions. From the source image a saliency map is computed, then this map is segmented, generating the seeds of the landmark regions. These seeds are enclosed by bounding boxes, which are fitted to the salient elements in the image using color histogram backprojection and mean-shift algorithms. Finally, the landmark bounding boxes are expanded, encompassing the immediate surrounding regions.

4 Landmark Characterization and Matching

After the determination of the bounding boxes, using the procedure proposed in the previous section, region descriptors are extracted using low-level image features. These descriptors should be appropriate to characterize the bounding boxes as signatures of the landmarks and should make possible the comparison between them.

The descriptors must be invariant to scale, translation and rotation, or at least at limited amounts of these transformations, because the images are taken from several different locations, as the robot moves. Typically the landmarks change their appearance with changes in viewpoint; they can even change completely in shape, color and texture. Considering these constraints, the following region descriptors were implemented:

1. Normalized chromaticity histogram of segmented salient spots.
2. Normalized chromaticity histogram of fitted bounding box (after backprojection and mean shift).
3. Normalized chromaticity histogram of expanded bounding box (including surround area).
4. Mean saliency of fitted bounding box.

The histograms are all normalized to be independent of scale. Texture features are not used because, with the large distance from robot to landmarks, the texture discrimination is not effective, and because texture features are strongly sensitive to the illumination changes present in outdoor environments.

The similarity between the histogram descriptors of two image regions i and j is measured by the distance between their corresponding points h_i and h_j in histogram space [35]. The quadratic form metric

[14] is used:

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

where h_i and h_j are n-dimensional color histograms, and $\mathbf{A}$ is the similarity matrix, whose elements a_{kl} ($0 \le a_{kl} \le 1$) denote similarity between bins k and l.

The histograms are normalized, then $h_0 = h_1 - h_2$ can be defined, resulting in $\sum h_0[i] = 0$, and the quadratic form becomes a distance that can be evaluated by:

$$d^2_{hist}(h_1, h_2) = h_0{}^T \mathbf{A} h_0 . \quad (3)$$

The similarity matrix is defined as follows:

$$a_{ij} = (1 - d_{ij} / d_{\max}) , \quad (4)$$

where a_{ij} is an element of the similarity matrix, d_{ij} is the Euclidean distance (L_2) between colors i and j, and $d_{\max} = \max_{ij}(d_{ij})$. A refined definition of the similarity matrix, adopted in this work, is:

$$a_{ij} = \exp(-\sigma(d_{ij} / d_{\max})^2) , \quad (5)$$

for some positive coefficient σ (in this work, σ=16). The greater the coefficient σ, the more restricted the similarity between color bins. With $\sigma \to \infty$, the matrix becomes a diagonal matrix, and the quadratic-form distance converges to the square of the Euclidean distance.

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. Thus, the quadratic-form metric is more robust to small color shifts due to illumination changes.

Initially, the distances corresponding to the four region descriptors were combined using the root of the sum of the squared distances, resulting in a single value to the distance between landmark pairs.

Using a set of sample images taken in an outdoor environment, 68 landmarks were detected and characterized. The retrieval performance of the system was evaluated taking each time one landmark out of the database and matching it against the other landmarks. This experiment is described in detail in [38]. At this point, the process of detection and refining of landmark areas was validated, and a reasonable *recall* index was achieved (0.697), but the *precision* index was low (0.264), meaning that there were many false positives in the recognition process. This happened because the color and saliency descriptors adopted do not have enough information content to ensure unambiguous recognition of single landmarks. In the following two Sections it is described how the recognition process can be improved with the concept of landmark-view and Section 7 explains how logistic regression was used to define a matching score.

5 Grouping Landmarks and Defining Views

The main idea is that 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.

As the robot moves, other landmark-views are acquired in sequence (Figure 3). It is possible that some observed landmarks in neighboring landmark-views refer to the same objects or environment regions. A path corresponds to a sequence of views, and the relationships between landmark-views can be expressed by means of a graph.

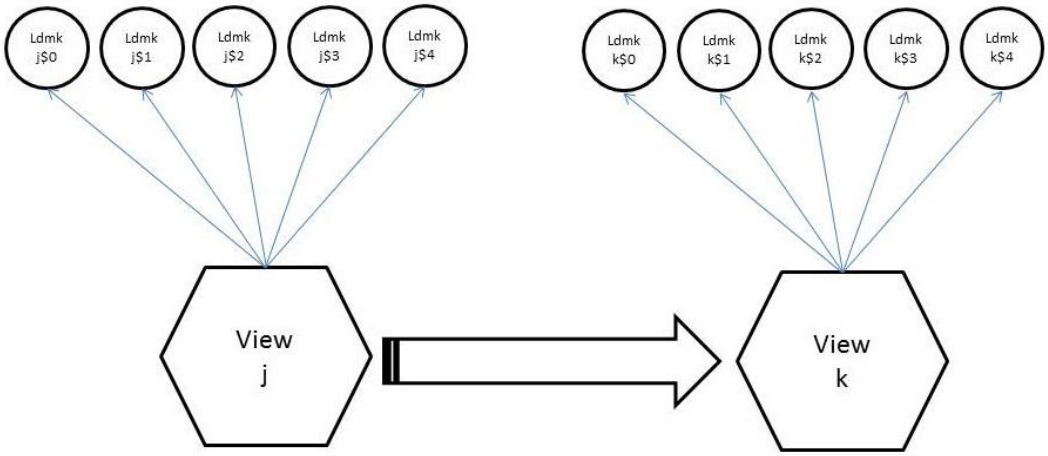

Figure 3. A *landmark-view* is defined as the set of landmarks observed at some location. In this figure two successive views (j and k) are represented, each one with a set of five landmarks. Some of the landmarks could be the same in both views, but not necessarily.

Consequently, the problem of landmark recognition is handled as a component of a higher-level problem, namely landmark-view recognition. In order to be able to recognize a landmark-view, it is necessary to establish a distance metric to match pairs of landmark-views. The next section describes how the similarity between views is evaluated.

6 View Matching

Since landmark-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 is also 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 [5, 6, 26, 29, 38]. 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 that 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 $E\subseteq U\times V$. Disregarding the adjacency constraint, in a bipartite graph a match is any subset of edges of it:

$$M=\{m_i\}\subseteq E \tag{6}$$

The set of unmatched vertices is defined as:

$$S=\{s\mid s\in U,\nexists v:(s,v)\in M\}\bigcup\{s\mid s\in V,\nexists u:(u,s)\in M\} \tag{7}$$

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,

$$\forall(u_i,v_j)\in M,\forall(u_k,v_l)\in M:(i=k)\Leftrightarrow(j=l) \tag{8}$$

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:

$$\text{cost}(M,S)=\sum_{m\in M}c(m)+\sum_{s\in S}c'(s) \tag{9}$$

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.

When edges are weighted with the cost of matching the two linked vertices, 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 (Figure 4). The weight of each edge represents the similarity distance between the two linked landmarks, as defined in Section 4. 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.

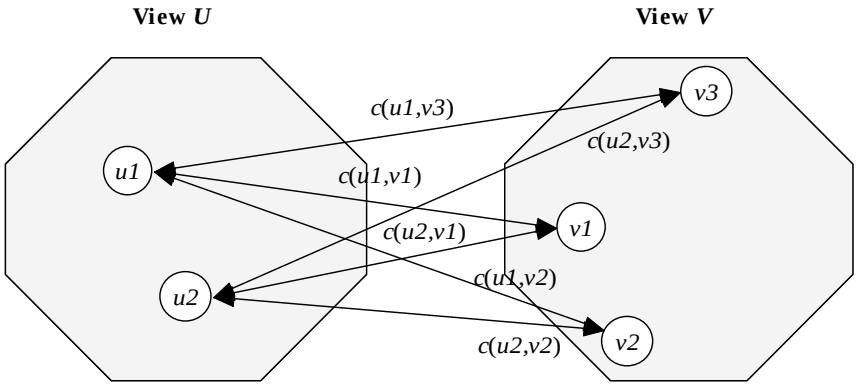

Figure 4. A bipartite graph used to compare two views *U* and *V*. The matching cost $c(u_i, v_i)$ of each landmark pair u_i and v_i is associated with each edge.

Among the several available algorithms to solve the bipartite matching problem [8]**,** the *relaxation algorithm* [5] was adopted because of its broad use, simplicity, and existence of reported successful experiences in the image-matching field. This is an approximate tree-search, graph matching algorithm, with polynomial complexity, based on the A^* algorithm with an incremental heuristic that avoids recalculating its terms at each step [8]. The algorithm consists of the following steps:

1. Initially, the matching restriction is relaxed, allowing any vertex in *V* to be assigned to 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 $v_z \in V$, and 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 search, the resulting computation complexity is $O(N^3)$, but it can be reduced using optimized shortest path search algorithms, for example, to $O(NlogN+M)$ using the Fibonacci heap method, where M denotes the number of edges and N denotes the number of vertexes in the graph [11].

The distance between two landmark-views is computed 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.

As an illustration, we consider the pair of views shown in Figure 5, taken from the image database of the experiment reported later in Section 8. The following matrix *Dist* represents the landmark distances from one view *Q* (left) to another view *D* (right):

$$Dist = \begin{bmatrix} 0.0211 & 0.0560 & 0.1332 & 0.0366 \\ 0.0408 & 0.0497 & 0.1717 & 0.0554 \\ 0.0695 & 0.0326 & 0.0100 & 0.0641 \\ 0.0603 & 0.0940 & 0.2619 & 0.0739 \end{bmatrix},$$

where the rows correspond to landmarks with labels 57, 55, 56, and 54 in the left view and the columns correspond to landmarks 155, 153, 154, and 156 in the right view. The landmarks are arranged in order of decreasing saliency values. Each matrix element corresponds to the distance between the respective landmarks. The sequence of assignments to solve the weighted bipartite matching, and consequently the view matching, is indicated in Figure 6.

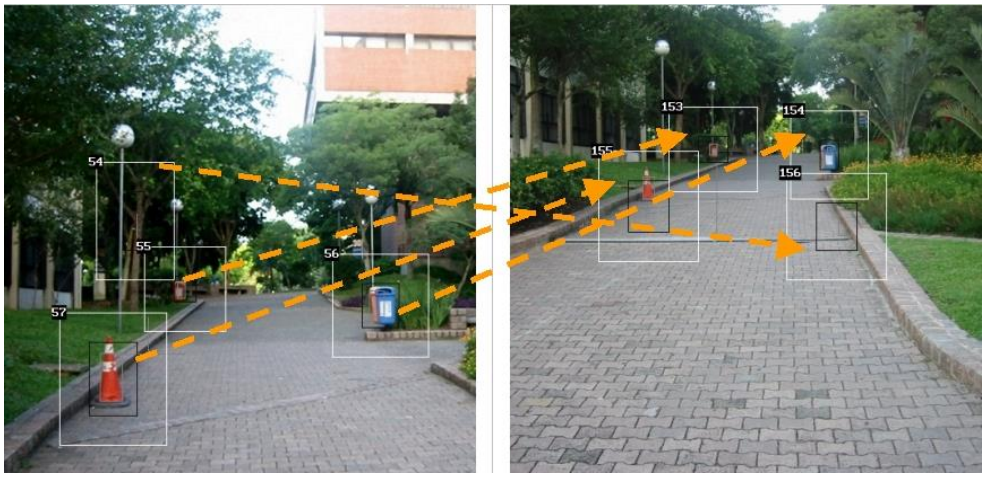

Figure 5. Landmark matching in similar views. The arrows indicate the solution of the weighted bipartite matching. The numbers inside black rectangles are landmark labels.

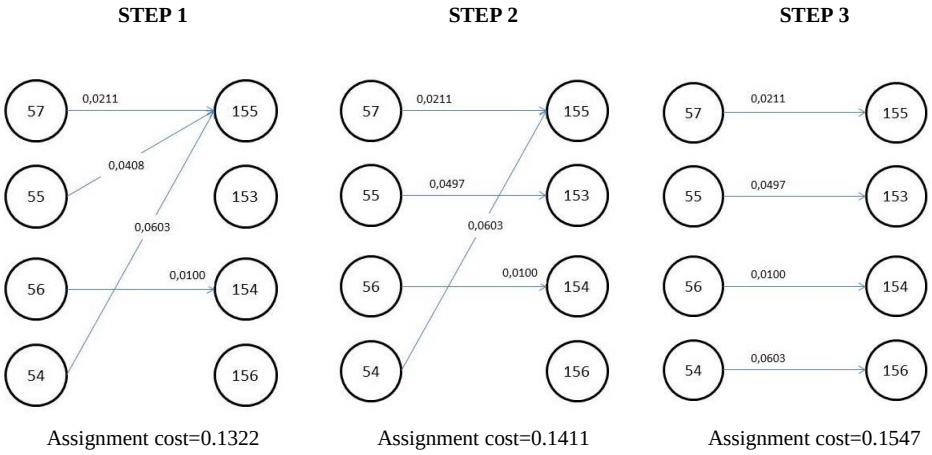

Figure 6. Successive assignment of nodes as the relaxation algorithm evolves, from left to right. In the initial stage, Q-nodes (57, 55, 56, 54) are assigned to D-nodes (155, 153, 154, 156) by minimizing the cost according to matrix *Dist*. At this step, node over-assignments are allowed. In the second step the over-assignment of node 155 is partially solved. In the last stage, the over-assignment of node 155 in view is solved. This is the final assignment corresponding to the solution of the weighted bipartite graph matching. At each step the removed over-assignment is the one that reallocation produces the minimum increment in the total assignment cost.

7 A Statistical Model for View-Matching

In the landmark-view matching algorithm presented in the preceding section, the different distances that can be obtained from each dimension of the landmark descriptors (color histograms, mean saliency, distance to centroid) were just combined with a root mean of squares (Section 4), resulting in a scalar distance.

Here, we propose to use logistic regression [9] to evaluate the significance of each landmark descriptor dimension and to use this information to build a statistical model for the view and landmark recognition process.

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:

$$VM = \frac{e^{A+B_1X_1+B_2X_2+B_3X_3+B_4X_4+B_5X_5+B_6X_6+B_7X_7+B_8X_8}}{1+e^{A+B_1X_1+B_2X_2+B_3X_3+B_4X_4+B_5X_5+B_6X_6+B_7X_7+B_8X_8}} \quad (10)$$

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 [38]. 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 Nagelkerke coefficient represents the proportion of the total variability of the outcome that is accounted for by the model. 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 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.

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	1.57	0.055	1.57	0.056	1.37	0.110	1.33	0.207	0.30	0.790
Fitted Salient Region Color	0.04	0.988	-0.02	0.993	0.81	0.792	0.75	0.819	3.87	0.300
Expanded Salient Region Color	14.0	0.001	14.0	0.001	16.0	0.001	16.0	0.001	18.8	0.001
Saliency			-1.74	0.836	-4.69	0.615	-4.70	0.614	-12.0	0.296
Distance to Centroid					-1.54	0.008	-1.59	0.184	-1.31	0.317
Combined Sum of Squares							0.16	0.958	1.91	0.561
NA1									1.18	0.003
NA2									1.18	0.003
Constant A	-1.5	0.004	-1.4	0.011	-0.8	0.204	-0.8	0.231	-3.5	0.002
% explained (Nagelkerke R2)	43.3		43.3		48.5		48.5		56.7	
% 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	

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 Eq. (10) 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 parameters of the model were computed off-line, using the SPSS package [34], based on a sample set of images, and the match score function was developed ad hoc and implemented in the view recognition system.

8 Experimental Results

To validate the landmark-based view recognition system, a university campus was chosen (Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS, Brazil) as a real outdoor environment. A set of 990 view pairs from 45 different views, with 273 landmarks found by the proposed method, was analyzed (Figure 7). The images were acquired with a standard color CCD camera, producing 512 x 512 pixels RGB images, with 24 bits/pixel, at a rate of 0.5 Hz.

Of the 42 corresponding view pairs, determined by hand, 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%. For comparison, the landmark recognition system described in Section 4 presented good correct landmark classification (69.7% of similar landmarks were correctly classified), but the ability to discriminate non-similar landmarks was poor (just 26.4% of landmarks considered similar to a query landmark were truly similar landmarks).

Using a standard low-performance PC computer (Pentium III 900MHz, 256Mb DRAM, Microsoft Windows XP) each view matching was performed in 0.69 seconds. Reducing the search space for view matching, by taking into account the recent history within a probabilistic approach, would avoid the comparison of the current query view with all the stored views. This can be accomplished with Kalman filtering or using logical connections between views, but this was out of the scope of this work.

Figure 7. Some of the test images taken in the outdoor experiment at PUCRS.

9 Discussion and Conclusions

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 was shown 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- or SURF- based engine, combining the good properties of both techniques. The detection and characterization of landmarks outdoors, based solely on the information from one camera is a very challenging task, due to several factors, such as landmark occlusions, adverse illumination conditions, and possible presence of dynamic elements in the environment. Thus, we consider this work as contributing to part of a more comprehensive solution to autonomous robot navigation outdoors, involving fusion of multiple sensors and techniques, as well as the integration of low-level feature-based recognition with high-level modeling of the world.

Acknowledgment

This work was partially funded by the GARNICS (Gardening with a Cognitive System) project FP7-ICT-247947

References

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

[2] S. Antani, R. Kasturi, and R. Jain, "A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video," *Pattern Recognition*, vol. 35, pp. 945-965, 2002.

[3] 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.

[4] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, pp. 346-359, 2008.

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

[6] 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.

[7] 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 International Conference on Intelligent Robots and Systems (IROS '98), Victoria, Canada, pp. 730-735, 1998.

[8] D. Conte, P. Foggia, C. Sansone, and M. Vento, "Thirty years of graph matching in pattern recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, pp. 265-298, 2004.

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

[10] P. Espinace, D. Langdon, and A. Soto, "Unsupervised identification of useful visual landmarks using multiple segmentations and top-down feedback," *Robotics and Autonomous Systems*, vol. 56, pp. 538-548, 2008.

[11] M. L. Fredman and R. E. Tarjan, "Fibonacci heaps and their uses in improved network optimization algorithms," *Journal of the Acm*, vol. 34, pp. 596-615, 1987.

[12] J. P. Gottlieb, M. Kusunoki, and M. E. Goldberg, "The representation of visual salience in monkey parietal cortex," *Nature*, vol. 39, pp. 481-484, 1998.

[13] S. Gouteux and E. S. Spelke, "Children's use of geometry and landmarks to reorient in an open space," *Cognition*, vol. 81, pp. 119-148, 2001.

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

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

[16] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, pp. 83-105, 2001.

[17] I. Kirigin and S. Singh, "Bearings based robot homing with robust landmark matching and limited horizon view," Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, Technical report CMU-RI-TR-05-02, 2005.

[18] M. Lalonde, D. Byrns, L. Gagnon, N. Teasdale, and D. Laurendeau, "Real-time eye blink detection with GPU-based SIFT tracking," Fourth Canadian Conference on Computer and Robot Vision, 2007 (CRV '07), pp. 481-487, 2007.

[19] T. S. Levitt and D. T. Lawton, "Qualitative navigation for mobile robots," *Artificial Intelligence*, vol. 44, pp. 305-360, 1990.

[20] A. R. Lew, J. G. Bremner, and L. P. Lefkovitch, "The development of relational landmark use in six- to twelve-month old infants in a spatial orientation task," *Child Development*, vol. 71, pp. 1179-1190, 2000.

[21] J. Li and N. M. Allinson, "A comprehensive review of current local features for computer vision," *Neurocomputing*, vol. 71, pp. 1771-1787, 2008.

[22] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91-110, 2004.

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

[24] E. N. Mortensen, H. Deng, and L. Shapiro, "A SIFT descriptor with global context," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005 (CVPR'05), pp. 184-190 vol. 1, 2005.

[25] H.-C. Nothdurft, "Saliency from feature contrast: additivity across dimensions," *Vision Research*, vol.

40, pp. 1183-1201, 2000.
[26] M. Peura, "Attribute trees as adaptive object models in image analysis," *Neural Networks Research Centre*. Espoo: Helsinki University of Technology, 2001.
[27] G. X. Ritter and J. N. Wilson, *Handbook of Computer Vision Algorithms in Image Algebra*. Boca Raton, Ney York, London, Tokyo: CRC Press, 1996, page 128 (Multilevel Thresholding).
[28] P. Sala, R. Sim, A. Shokoufandeh, and S. Dickinson, "Landmark Selection for Vision-Based Navigation," *IEEE Transactions on Robotics and Automation*, vol. 22, pp. 334-349, 2006.
[29] A. Sanfeliu, R. Alquézar, J. Andrade-Cetto, J. Climent, F. Serratosa, and J. Vergés, "Graph-based representations and techniques for image processing and image analysis," *Pattern Recognition*, vol. 35, pp. 639-650, 2002.
[30] S. Santini and R. Jain, "Gabor space and the development of preattentive similarity," International Conference on Pattern Recognition, Vienna, Austria, pp., 1996.
[31] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *International Journal of Computer Vision*, vol. 37, pp. 151-172, 2000.
[32] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *International Journal of Robotics Research*, vol. 21, pp. 735-758, 2002.
[33] M. Spetch, D. M. Kelly, and D. P. Lechelt, "Encoding of spatial information in images of an outdoor scene by pigeons and humans," *Animal Learning & Behavior*, vol. 26, pp. 85-102, 1998.
[34] SPSS, "Statistical Package for the Social Sciences," 10.0.5 ed. Chicago, Illinois: SPSS Inc., 2000.
[35] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, pp. 11-32, 1991.
[36] C. J. Taylor and D. J. Kriegman, "Vision-based motion planning and exploration algorithms for mobile robots," *IEEE Transactions on Robotics and Automation*, vol. 14, pp. 417-426, 1998.
[37] E. Todt and C. Torras, "Detecting salient cues through illumination-invariant color ratios," *Robotics and Autonomous Systems*, vol. 48, pp. 111-130, 2004.
[38] E. Todt and C. Torras, "Color-contrast landmark detection and encoding in outdoor images," The 11th International Conference on Computer Analysis of Images and Patterns (CAIP'05), Versailles, France, pp. 612-619, 2005.
[39] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97-136, 1980.
[40] C. Valgren and A. J. Lilienthal, "SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments," *Robotics and Autonomous Systems*, vol. 58, pp. 149-156, 2010.
[41] R. VanRullen, "Visual saliency and spike timing in the ventral visual pathway," *Journal of Physiology - Paris*, 2002.