Pattern Recognition Letters 30 (2009) 1464–1476

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A novel approach for salient image regions detection and description

R. Vázquez-Martín [a,*], R. Marfil [a], P. Núñez [b], A. Bandera [a], F. Sandoval [a]

[a] Grupo ISIS, Dpto. Tecnología Electrónica, E.T.S.I. Telecomunicación, Universidad de Málaga, Campus de Teatinos s/n, 29071 Málaga, Spain
[b] Department Tecnología de los Computadores y las Comunicaciones, Universidad de Extremadura, Plaza de Caldereros s/n, 10071 Cáceres, Spain

ARTICLE INFO

Article history:
Received 2 October 2008
Received in revised form 18 June 2009
Available online 8 August 2009
Communicated by H.H.S. Ip

Keywords:
Visual landmarks
Irregular pyramids
Salient image regions detection
Kernel-based description

ABSTRACT

This paper proposes a new algorithm for visual landmarks detection and description. The detection is achieved using a hierarchical grouping mechanism, which combines a color contrast measure defined between regions with internal region descriptors and with attributes of the shared boundary. This detector reliably finds the same salient regions under different viewing conditions. Then, geometrically and photometrically normalized regions are characterized by a kernel-based descriptor. This descriptor is rotation-invariant and robust against noise. Several tests are conducted in order to compare the proposed approach with other similar approaches. Experimental results prove that the performance of our proposal is high in terms of computational consuming and visual landmark detection and description abilities.

1. Introduction

Reliable navigation is an essential component of an autonomous mobile robot. In order to perform this task correctly, the robot typically needs to represent the information perceived by external sensors into a navigation map. One popular choice is to build this map with distinguished natural landmarks that the robot can acquire from the environment without human supervision. Recognizable landmarks are essential since they will be used as reference marks to identify locations in the world.

In the past, a variety of approaches for feature-based mobile robot localization and navigation has been developed. These approaches mainly differ in the method employed to represent the belief of the mobile robot about its current pose or to find and track a safe path to a goal. Furthermore, they can be differentiated according to the type of sensor information that they use. Thus, a significant number of approaches use range sensors to detect these distinguished landmarks. However, although these approaches can robustly address the landmark detection problem in indoor environments, they become less robust in outdoors, since they have to deal with highly unstructured and dynamic environments. In these cases, a slight change in the robot pose can provoke a large change in the obtained range scan (Lingemann et al., 2004). An alternative to the active ranging devices are vision systems. These systems are passive and of high resolution, and they provide a huge amount of information (color, texture or shape) which allow disambiguating landmarks for subsequent data association purposes. On the other hand, to perform the data association, visual-based systems need to compare image patches obtained by the robot's cameras with patches stored in a map. This detection and matching process has a high computational complexity, consuming a big amount of computational resources. In fact, in order to develop a practical visual-based navigation system, this issue constitutes the main hurdle to overcome (Davison and Murray, 2002). Lighting, dynamic backgrounds and view-invariant matching are issues which must be also addressed (Siagian and Itti, 2007).

In this paper, we describe a vision-based approach for natural landmark detection and description. The detector assumes that there is a set of regions in most images that can be detected with high repeatability since they posses some distinguishing and stable properties. In our approach, such salient regions will be extracted using a hierarchical algorithm, which presents two stages: firstly, it segments the input image into blobs of homogeneous color and then, it merges these blobs using a similarity criterion. Basically, this criterion complements a contrast measure defined between regions with internal region descriptors and with attributes of the shared boundary. Both grouping processes are performed over a modified version of the Bounded Irregular Pyramid (BIP) (Marfil et al., 2004, 2007a). The data structure of the BIP is a mixture of the regular and irregular data structures (see Section 3), and it has been previously employed to track non-rigid objects (Marfil et al., 2007b) or to segment color images (Marfil et al., 2006, 2007a). However, experimental results have shown that, although computationally efficient, the BIP-based approaches are excessively affected by the shift-variance problem (Marfil et al., 2006). In this paper, we propose to modify the decimation scheme

* Corresponding author. Tel.: +34 650562348; fax: +34 952131447.
E-mail address: rvmartin@uma.es (R. Vázquez-Martín).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.08.003

of the BIP. This scheme is referred as uBIP and it allows a high degree of mixture between the regular and irregular parts of the BIP data structure, reducing the shift-variance problem without an increase of the computational cost. The main differences between the BIP and the uBIP is described in Section 3. Perceived regions of data-dependent shape will serve as natural landmarks. Experimental results show that our approach could be used to build sparse maps where landmarks are perceptually distinguished and, besides, they usually have an underlying semantic significance. On the other hand, to characterize these visual landmarks, we have chosen as feature space its color probability density function (pdf), which must be estimated from the region data. To reduce the computational cost, n-bin histograms are employed. Besides, in order to take into account the spatial information and not only the spectral one, geometrically and photometrically normalized visual landmarks are characterized by spatially masking them with an isotropic kernel. The similarity between kernel-based representations of a detected landmark and a reference one will be measured using the metric derived from the Bhattacharyya coefficient (Comaniciu et al., 2003), which will have the meaning of a correlation score. Finally, it must be noted that the hierarchical grouping mechanism employed by the visual landmarks detector uses depth information. In our tests, we use a stereoscopic vision to directly acquire this information from the perceived images.

The paper is organized as follows: after discussing related work to the feature-based vision systems in Section 2, the proposed approach for the acquisition and description of visual landmarks is described in Section 3. Section 4 deals with some obtained experimental results. In this Section, the results of a comparative study of the proposed method with other methods are given. Finally, the paper concludes along with discussions and future work in Section 5.

2. Related work

2.1. Local image features detectors and descriptors

Feature-based vision systems for mobile robot localization and navigation identify each scenario or environment pose with a set of landmarks and their spatial distribution. These landmarks must own some invariant and stable property in order to be detected with high repeatability in images taken from arbitrary viewpoints. One of the main advantages of these approaches is that they transform images in a more compact form before attempting to compare them to the ones presented on a map or to store them in a map built simultaneously. This allows to increase the efficiency and robustness of the localization process. Then, the matching between an input image and a map is posed as a search in the correspondence space established between the associated sets of landmarks. If both sets of landmarks are robustly matched, then these approaches will provide a high localization resolution.

The majority of feature-based vision systems use local interest points as landmarks. These points define regions within the image which are distinctive from the rest of the image (Asmar et al., 2006). The development of algorithms which use a set of local distinguished items can be traced back to the work of Moravec (1977) and Harris (1992). For instance, the 3D vision system DROID uses the visual motion of image corner features for scene reconstruction (Harris, 1992). It is able to determine the camera motion and landmark positions from the locations of the tracked image features. Although it is sensitive to the scale of the image, the Harris–Stephens corner detector (Harris and Stephens, 1988) has been used to detect visual landmarks (Kim et al., 2005). It must be noted that this is not a problem for the case of no scale change between views, as in ceiling images (Jeong and Lee, 2005). Landmarks can be also detected using this corner detector as applied by Shi and Tomasi (1994) to relatively large pixel patches (15×15 rather than the usual 5×5 for corner detection) (Davison and Murray, 2002; Kim and Chung, 2005). With respect to the landmark description, all these approaches describe them using their associated image patches. Then, matching can be achieved using normalized sum-of-squared-differences (NSSD) for the best match to the stored landmark patch (Davison and Murray, 2002; Kim et al., 2005). The Difference of Gaussians has been applied by Lowe (1999) in order to achieve scale invariance. The scale-invariant feature transform (SIFT) have been widely applied for visual landmark detection and description (Se et al., 2002; Elinas et al., 2006). Se et al. (2002) use a trinocular stereo system to determine 3D estimates for landmark locations. Landmarks are used only when they appear in all three images with consistent disparities, resulting in very few outliers. This work has also addressed the problem of place recognition, in which a robot can be switched on and recognize its location anywhere within a large map (Se et al., 2005). Despite of its excellent properties, the SIFT detector tends to extract visually meaningless features on the blob-like parts of images (Ahn et al., 2006). Recently, other similar visual landmark detectors have been employed that overcome this problem (e.g. the multi-scale Harris detector (Lin et al., 2005; Ahn et al., 2006) or the Harris-Laplace detector (Jensfelt et al., 2006; Wang et al., 2006). Lin et al. (2005) describe landmarks using the Zernike moments. However, this descriptor suffers from large computational burden and low discriminating capability. Therefore, it is more usual that these approaches employ SIFT (Ahn et al., 2006; Wang et al., 2006), PCA-SIFT (Ke and Sukthankar, 2004) or rotation-variant SIFT (Jensfelt et al., 2006) to describe the obtained visual landmarks. Other scale and rotation invariant interest point detector and descriptor is SURF (speeded up robust features) (Bay et al., 2006). The interest points detection is based on the Hessian matrix and their description uses a distribution of Haar-Wavelet responses within the interest points neighborhood, but it relies on integral images to reduce the computation time.

The advantage of systems based on local interest point descriptors is that no model of landmarks has to be specified to the vision system a priori. Besides, these approaches generate dense occupancy maps, but comprising of landmarks with no underlying semantic significance. The disadvantage of such systems is scalability. Thus, these systems are usually implemented in environments where the number of detected landmarks is relatively small (Asmar et al., 2006). Moving the robots to a larger environment requires the management and recognition of a much larger number of landmarks. An excessively huge number of landmarks can provoke that the reliability and repeatability of visual features can not always be guaranteed, appearing outliers in feature matching which can lead to unreliable data association (Ahn et al., 2006). This problem has been addressed by grouping local interest points together and using these groups as landmarks (Ahn et al., 2006) or by imposing a fixed number of landmarks (e.g., the iterative SIFT (Tamimi et al., 2006)). Other solution has been suggested by model-based visual landmark detectors. These detectors are employed to build sparse maps using landmarks that have an underlying semantic significance. Thus, image edges (Folkesson et al., 2005) or planar quadrangles (Hayet et al., 2003; Vázquez-Martín et al., 2005) can be employed to match images. Environment-specific features like walls or doors are used by Horswill (1993). To deal with outdoor environments, Asmar et al. (2006) propose a tree detection approach.

2.2. Computational visual attention approaches for salient regions detection

In biological vision systems, the attention mechanism is the responsible of selecting the relevant information from the sensed field of view so that the complete scene can be analyzed using a sequence of rapid eye saccades (Aziz and Mertsching, 2007). In the recent years, efforts have been made to imitate such attention behavior in artificial vision systems, because it allows to optimize the computational resources as they can be focused on the processing of a set of selected regions only. Probably one of the most influential theoretical models of visual attention is the spotlight metaphor (Eriksen and Yenh, 1985), by which many concrete computational models have been inspired (Koch and Ullman, 1985; Milanese, 1993; Itti, 2002). These approaches are related with the feature integration theory, a biologically plausible theory proposed to explain human visual search strategies (Treisman and Gelade, 1980). According to this model, these methods are organized into two main stages. First, in a preattentive task-independent stage, a number of parallel channels compute image features. The extracted features are integrated into a single saliency map which codes the saliency of each image region. The most salient regions are selected from this map. Second, in an attentive task-dependent stage, the spotlight is moved to each salient region to analyze it in a sequential process. Analyzed regions are included in an inhibition map to avoid movement of the spotlight to an already visited region. Thus, while the second stage must be redefined for different systems, the preattentive stage is general for any application. Although these models have good performance in static environments, they cannot in principle handle dynamic environments due to their impossibility to take into account the motion and the occlusions of the objects in the scene. In order to solve this problem, an attention control mechanism must integrate depth and motion information to be able to track moving objects. Thus, Maki et al. (2000) propose an attention mechanism which incorporates depth and motion as features for the computation of saliency.

The previously described methods deploy attention at the level of space locations (space-based models of visual attention). The models of space-based attention scan the scene by shifting attention from one location to the next to limit the processing to a variable size of space in the visual field. Therefore, they have some intrinsic disadvantages. In a normal scene, objects may overlap or share some common properties. Then, attention may need to work in several discontinuous spatial regions at the same time. If different visual features, which constitute the same object, come from the same region of space, an attention shift will be not required (Sun and Fisher, 2003). On the contrary, other approaches deploy attention at the level of objects. Object-based models of visual attention provide a more efficient visual search than space-based attention. Besides, it is less likely to select an empty location. In the last few years, these models of visual attention have received an increasing interest in computational neuroscience and in computer vision. Object-based attention theories are based on the assumption that attention must be directed to an object or group of objects, instead of a generic region of the space (Orabona et al., 2007). Therefore, these models will reflect the fact that the perception abilities must be optimized to interact with objects and not just with disembodied spatial locations. Thus, visual systems will segment complex scenes into objects which can be subsequently used for recognition and action.

Finally, space-based and object-based approaches are not mutually exclusive, and several researchers have proposed attentional models that integrate both approaches. Thus, in the Sun and Fisher's proposal (Sun and Fisher, 2003), the model of visual attention combines object- and feature-based theories. In its current form, this model is able to replicate human viewing behaviour. However, it needs that input images will be manually segmented. That is, it uses information that is not available in a preattentive stage, before objects are recognized (Orabona et al., 2007).

Computational visual attention systems determine globally which regions in the image discriminate instead of locally detecting predefined properties like corners. Hence, they can be useful to detect good landmark candidates. Thus, Ouerhani and Hügli (2005) propose an approach which takes advantage of the saliency-based model of attention to automatically learn configurations of salient visual landmarks. The selected landmarks are organized into a topological map that is used for self-localization. On the other hand, Newman and Ho (2005) use a saliency measure based on entropy to define important locations primarily for the loop closing detection in the simultaneous localization and mapping (SLAM) problem. Other main advantage of attention systems is that they can additionally integrate previous knowledge about the landmark-based map into the computations. This enables a better re-detection of landmarks when presuming to revisit a known location (Frintrop et al., 2006).

3. Visual landmark detection and description

The proposed approach for visual landmark detection and description consists of three stages. Firstly, a *hierarchical grouping algorithm* is applied to perform a domain-independent segmentation of the image pixels into regions (Marfil et al., 2007c). Then, the set of regions which satisfies certain rules are selected as landmarks and geometrically and photometrically normalized (Obdrzálek and Matas, 2006) (visual landmarks detection and normalization stage). These rules do not depend on the environment or application, and they impose to the obtained landmarks properties like high contrast with respect to its surrounding background. The shape of these salient regions is adapted to real items of the scene. Therefore, continuous geometric changes of the viewpoint preserve their internal topology, i.e. pixels from a single connected region are transformed to a new single connected region (see Fig. 1). Finally, landmarks are characterized by a rotation-invariant kernel-based descriptor (Comaniciu et al., 2003), which is adopted to represent its internal color distribution.

3.1. Hierarchical grouping algorithm

The hierarchical grouping algorithm performs the segmentation of the input image using two consecutive stages. The pre-segmentation stage employs a color distance to group the image pixels into a set of blobs whose spatial distribution is physically representative of the image content. Then, the perceptual grouping stage groups the set of homogeneous blobs into a smaller set of regions taking into account not only the internal visual coherence of the obtained regions but also the external relationships among them (Marfil et al., 2007c). To accomplish this grouping process, where the fine details are clustered into more coarse entities, the contents of the input image can be described using multiple representations with decreasing resolution. Pyramids are hierarchical structures which have been widely used to represent the perceptual organization of the image by a tree of regions, ordered by inclusion (Marfil et al., 2006). In this hierarchy, each level is a graph which is at least defined by a set of nodes, which represent regions, connected by a set of arcs, which represent region adjacency relationships.

The efficiency of a pyramid to represent the information is strongly influenced by two features: the graph selected to encode the information within each pyramid level and the decimation scheme used to build one graph from the graph below (Marfil et al., 2006). The choice of a graph encoding determines the information that may be encoded explicitly at each level of the pyramid.

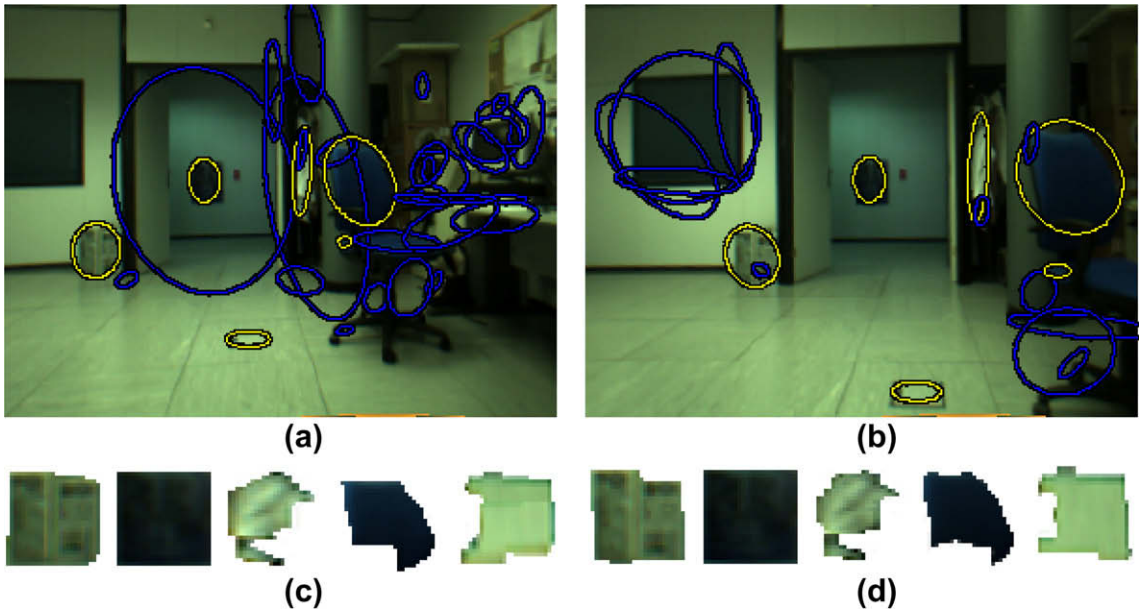

Fig. 1. (a and b) Regions generated by the proposed detector on two images taken from different viewpoints. Representing ellipses have been chosen to have the same first and second moments as the originally arbitrarily shaped region (matched regions are marked on yellow); and (c) and (d) normalized versions of five matched regions at images (a) and (b) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

Thus, it roughly corresponds to setting the horizontal properties of the pyramid. On the other hand, the reduction or decimation scheme used to build the pyramid determines the dynamic of the pyramid (height, preservation of details...). It corresponds to the vertical properties of the pyramid. Depending on these two features, pyramids have been classified as regular and irregular ones. Regular approaches have a rigid structure where the decimation process is fixed. This rigid structure allows to build and process them with a low computational cost. However, this inflexibility can also provoke three main problems: non-connectivity of the obtained regions, impossibility to represent elongated objects and shift-variance (Marfil et al., 2006). Irregular pyramids solve these problems using a structure which dynamically adapts to the image layout. However, they require a computational time which is usually higher than the one required by regular pyramids.

In order to combine the advantages of regular and irregular pyramids, the Bounded Irregular Pyramid (BIP) was proposed by Marfil et al. (2004). The BIP arose as a mixture of regular and irregular structures whose goal is to obtain accurate results at a low computational cost. The regular decimation is applied in the homogeneous parts of the image, meanwhile the heterogeneous parts are decimated using a classical irregular process (Marfil et al., 2004; Marfil et al., 2006). Basically, the BIP is a graph hierarchy where each level l is a graph $G_l = (N_l, E_l)$ consisting of a set of nodes, N_l, linked by a set of intra-level edges E_l. Each graph G_l has a regular part which built from G_{l-1} using a $2 \times 2/4$ regular decimation procedure and an irregular part which is built from G_{l-1} using an irregular decimation process (Marfil et al., 2006; Marfil et al., 2007a). Therefore, there are two types of nodes: nodes belonging to the $2 \times 2/4$ regular part (regular nodes) and nodes belonging to the irregular part (irregular nodes). Each level of the BIP is computed in four steps (see Marfil et al. (2007a) for further details):

- *$2 \times 2/4$* regular decimation process: if four regular adjacent nodes of level l have similar color, a new regular node is created at $l + 1$.
- *Regular parent search and intra-level twining:* once the regular structure is generated, there are some regular nodes without parent. Each of these nodes n_r looks for regular neighbour nodes which are similar to them according to a given feature. If this searching is successfully achieved, then the regular node n_r will be linked to the parent of its most similar neighbour or, if this neighbour is not linked to a parent node, both nodes will be linked to a new irregular node.
- *Virtual parent search and virtual node linking:* each irregular node n_i looks for irregular neighbour nodes which are similar to them according to a given feature. If this searching is successfully achieved, then n_i will be linked to the parent of its most similar neighbour or, if this neighbour is not linked to a parent node, both nodes will be linked to a new irregular node at the level above.
- *Intra-level* edge generation in G_{l+1} : the edges of G_{l+1} are computed taking into account the neighborhood of nodes in G_l.

The BIP approximates or even outperforms previously proposed hierarchical segmentation schemes, yet can be computed much faster (Marfil et al., 2006). However, it is highly affected by the shift variance problem, i.e. it provides an image segmentation which varies when this image is shifted slightly. In this paper, we propose to modify the structure of the BIP in order to improve the mixture of the regular and irregular decimation processes, avoiding the shift-variance problem. This new pyramid will be referred as uBIP, as it uses a union-find algorithm to merge the nodes resulting of the regular and irregular decimation processes. In the uBIP, it is allowed that a node of the structure (regular or irregular) can be linked with any type of node from its same level. Next subsections briefly describe the pre-segmentation and perceptual grouping stages of the proposed segmentation approach.

3.1.1. Pre-segmentation stage

As it is described in (Marfil et al., 2007c), the pixels of the input image can be considered as the nodes of the graph G_0. Then, the pre-segmentation stage divides the image into regions of uniform color using the uBIP. Contrary to the BIP, this decimation algorithm only runs two consecutive steps to obtain the set of nodes N_{l+1}. The first process generates the set of regular nodes of G_{l+1} from the regular nodes at G_l, meanwhile the second one determines the set of irregular nodes at level $l + 1$. In this proposal, this second process

conducts an union-find decimation algorithm which is simultaneously conducted over the whole set of regular and irregular nodes of G_l which do not present a parent in the level $l+1$.

Let $G_l = (N_l, E_l)$ be a graph where N_l stands for the set of regular and irregular nodes and E_l for the set of intra-level arcs. Let ε_l^{xy} be equal to 1 if $(\boldsymbol{x},\boldsymbol{y}) \in E_l$ and equal to 0 otherwise. Let $\xi_{\boldsymbol{x}}$ be the neighborhood of the node $\boldsymbol{x}$ defined as $\{\boldsymbol{y} \in N_l : \varepsilon_l^{xy}\}$. It can be noted that a given node $\boldsymbol{x}$ is not a member of its neighborhood, which can be composed by regular and irregular nodes. Each node $\boldsymbol{x}$ has associated a $v_{\boldsymbol{x}}$ value. Besides, each regular node has associated a boolean value $h_{\boldsymbol{x}}$: the homogeneity (Marfil et al., 2007a). At the base level of the hierarchy, G_0, all nodes are regular, and they have $h_{\boldsymbol{x}}$ equal to 1. Only regular nodes which have $h_{\boldsymbol{x}}$ equal to 1 are considered to be part of the regular structure. Regular nodes with an homogeneity value equal to 0 are not considered for further processing. The proposed decimation process transforms the graph G_l in G_{l+1} such that the reduction factor is greater to 1. In our case, we focus on dividing the image into a set of homogeneous blobs. This aim is achieved using the pairwise comparison of neighboring nodes (Haxhimusa et al., 2003). Then, a pairwise comparison function, $g(v_{\boldsymbol{x}_1}, v_{\boldsymbol{x}_2})$ is defined. This function is true if the $v_{\boldsymbol{x}_1}$ and $v_{\boldsymbol{x}_2}$ values associated to the $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ nodes are similar according to some criteria and false otherwise. The decimation process consists of the following steps:

(1) Regular decimation process. The $h_{\boldsymbol{x}}$ value of a regular node $\boldsymbol{x}$ at level $l+1$ is set to 1 if the four regular nodes immediately underneath $\{\boldsymbol{y}_i\}$ are similar according to some criteria and their $h_{\{\boldsymbol{y}_i\}}$ values are equal to 1. That is, $h_{\boldsymbol{x}}$ is set to 1 if

$$\left\{ \bigcap_{\forall \boldsymbol{y}_j \boldsymbol{y}_k \in \{\boldsymbol{y}_i\}} g(v_{\boldsymbol{y}_j}, v_{\boldsymbol{y}_k}) \right\} \cap \left\{ \bigcap_{\boldsymbol{y}_j \in \{\boldsymbol{y}_i\}} h_{\boldsymbol{y}_j} \right\} \tag{1}$$

Besides, at this step, inter-level arcs among regular nodes at levels l and $l+1$ are established. If $\boldsymbol{x}$ is an homogeneous regular node at level $l+1$ ($h_{\boldsymbol{x}}$==1), then the set of four nodes immediately underneath $\{\boldsymbol{y}_i\}$ are linked to $\boldsymbol{x}$.

(2) Irregular decimation process. Each irregular or regular node $\boldsymbol{x} \in N_l$ without parent at level $l+1$ chooses the closest neighbor $\boldsymbol{y}$ according to the $v_{\boldsymbol{x}}$ value. Besides, this node $\boldsymbol{y}$ must be similar to $\boldsymbol{x}$. That is, the node $\boldsymbol{y}$ must satisfy

$$\left\{ \|v_{\boldsymbol{x}} - v_{\boldsymbol{y}}\| = \min\left(\|v_{\boldsymbol{x}} - v_{\boldsymbol{z}}\| : \boldsymbol{z} \in \xi_{\boldsymbol{x}}\right) \right\} \cap \left\{ g(v_{\boldsymbol{x}}, v_{\boldsymbol{y}}) \right\} \tag{2}$$

If this condition is not satisfied by any node, then a new node $\boldsymbol{x}'$ is generated at level $l+1$. This node will be the parent node of $\boldsymbol{x}$. Besides, it will constitute a root node and its receptive field at base level will be an homogeneous set of pixels according to the specific criteria. On the other hand, if $\boldsymbol{y}$ exists and it has a parent $\boldsymbol{z}$ at level $l+1$, then $\boldsymbol{x}$ is also linked to $\boldsymbol{z}$. If $\boldsymbol{y}$ exists but it does not have a parent at level $l+1$, a new irregular node $\boldsymbol{z}'$ is generated at level $l+1$. In this case, the nodes $\boldsymbol{x}$ and $\boldsymbol{y}$ are linked to $\boldsymbol{z}'$.This process is sequentially performed and, when it finishes, each node of G_l is linked to its parent node in G_{l+1}. That is, a partition of N_l is defined. It must be noted that this process constitutes an implementation of the union-find strategy. The union-find uses tree structures to represent sets. A find operation looks for the parent of a node at level l. If two nodes at level l are similar, then a union operation will be performed by setting one of the two nodes to be the parent of both ones at level $l+1$.

(3) Definition of intra-level arcs. The set of edges E_{l+1} is obtained by defining the neighborhood relationships between the nodes N_{l+1}. Two nodes at level $l+1$ are neighbors if their reduction windows are connected at level l.

The structure hierarchy stops growing when it is no longer possible to link together any more nodes because they are not similar. The set of nodes which are not linked to any node at upper levels define a partition of the input image (see Marfil et al. (2007c) for further details).

Fig. 2 shows an example of the described decimation process. Regular nodes are drawn as rectangles meanwhile irregular nodes are drawn as circles. The $v_{\boldsymbol{x}}$ values are represented by the grey level of the cells. Fig. 2a shows the regular part of the data structure after being built. The base level of the structure is composed by the 8×8 image pixels. The 4-to-1 regular decimation procedure generates a 4×4 level. Regular nodes with $h_{\boldsymbol{x}}$ equal to 0 are not depicted on the figure. Inter-level arcs join regular nodes between levels 0 and 1. Fig. 2b and c present the results of applying the irregular decimation process. The closest neighbor to node $\boldsymbol{x}_1^{(0)}$ is a regular node which has a parent at level 1, $\boldsymbol{x}_1^{(1)}$. Then, $\boldsymbol{x}_1^{(0)}$ is also linked to it. The closest neighbor of other nodes, like $\boldsymbol{x}_2^{(0)}$, does not have a parent at level 1. In this case, $\boldsymbol{x}_2^{(0)}$ and its closest neighbor, $\boldsymbol{x}_3^{(0)}$, generate a new irregular node at level 1, $\boldsymbol{x}_2^{(1)}$. Inter-level arcs link both nodes at level 0, $\boldsymbol{x}_2^{(0)}$ and $\boldsymbol{x}_3^{(0)}$, with the new irregular node at level 1, $\boldsymbol{x}_2^{(1)}$. Fig. 2c shows the set of irregular nodes and inter-level arcs generated after applying the irregular decimation process. It can be noted that several nodes of level 0 which were not involved in the regular decimation step, have been now linked to regular nodes at level 1. Fig. 2d shows the generation of level 2 from level 1. In this case, the regular decimation process does not generated any new regular node at level 2. The irregular decimation scheme merges $\boldsymbol{x}_1^{(1)}$ and $\boldsymbol{x}_6^{(1)}$, generating the irregular node $\boldsymbol{x}_1^{(2)}$. Then, the rest of regular nodes at level 1 and the irregular node $\boldsymbol{x}_4^{(1)}$ are sequentially joined to $\boldsymbol{x}_1^{(2)}$. The irregular nodes $\boldsymbol{x}_2^{(1)}$ and $\boldsymbol{x}_3^{(1)}$ generate the irregular node $\boldsymbol{x}_2^{(2)}$ at level 2. Finally, the irregular node $\boldsymbol{x}_5^{(1)}$ is joined to a node at level 2 because its $v_{\boldsymbol{x}}$ value is not similar to values of its neighbors according to the function $g(\cdot,\cdot)$. This node becomes the root of an uniform image region at base level. Irregular nodes $\boldsymbol{x}_1^{(2)}$ and $\boldsymbol{x}_2^{(2)}$ present different $v_{\boldsymbol{x}}$ values and they are not merged. Then, they are the roots of two uniform regions at base level. The input image is then divided into three regions which are uniform according to the criterion defined by $g(\cdot,\cdot)$.

This new decimation process avoids the shift-variance problem associated to the BIP. To demonstrate this issue, we have compared the proposed modification with the original BIP and with the main irregular structures present in the literature in a color-based segmentation framework. Obtained results are shown at Section 4.

3.1.2. Perceptual grouping stage

After the local similarity pre-segmentation stage, grouping blobs aims at simplifying the content of the obtained partition in order to provide a final image segmentation. Two constraints are taken into account for an efficient grouping process: first, although all groupings are tested, only the best groupings are locally retained; and second, all the groupings must be spread on the image so that no part of the image is advantaged. For managing this grouping, the uBIP structure is used: the roots of the pre-segmented blobs are considered as irregular nodes which constitute the first level of the perceptual grouping multiresolution output. Subsequent higher levels can be built using the segmentation scheme proposed by Marfil et al. (2007c). However, if the distance between two nodes in the pre-segmentation stage is based on a color criterion, in order to achieve this second grouping process, a more complex distance must be defined. This distance has three main components: the colour contrast between image blobs, the edges of the original image computed using the Canny detector, and the disparity of the image blobs, provided by the stereoscopic vision system. In order to speed up the process, a global contrast measure is used instead of a local one. It avoids to work at pixel

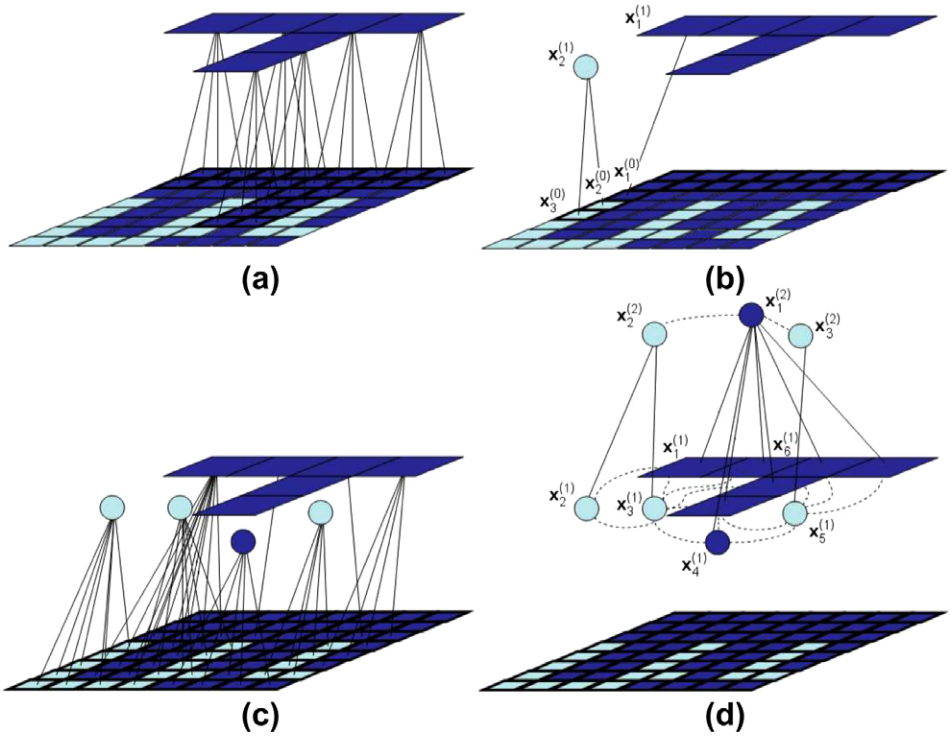

Fig. 2. Hierarchy generation: (a) regular nodes generated at level 1 and inter-level arcs among children and parents, (b) individual examples of the irregular decimation process, (c) nodes generated at level 1 and inter-level arcs among children and parents after the irregular decimation process, and (d) intra-level arcs at level 1 and generated nodes at level 2 (see text for details).

resolution, increasing the computational speed. This contrast measure is complemented with internal regions properties and with attributes of the boundary shared by both regions. To perform correctly, the nodes of the uBIP which are associated to the perceptual grouping multiresolution output store statistics about the CIELab colour values of the roots generated by the pre-segmentation stage which are linked to them and about their mean disparity. Then, the distance between two nodes n_i and n_j, $\Upsilon(n_i, n_j)$, is defined as

$$\Upsilon(n_i,n_j)=\sqrt{w_1\cdot\left(\frac{d(n_i,n_j)\cdot\min(b_i,b_j)}{\alpha\cdot c_{ij}+\beta(b_{ij}-c_{ij})}\right)^2+w_2\cdot(disp(n_i)-disp(n_j))^2} \tag{3}$$

where $d(n_i, n_j)$ is the color distance between n_i and n_j and $disp(x)$ is the mean disparity associated to the base image region represented by node x. b_i is the perimeter of n_i, b_{ij} is the number of pixels in the common boundary between n_i and n_j and c_{ij} is the set of pixels in the common boundary which corresponds to pixels of the boundary detected by the Canny detector. α and β are two constant values used to control the influence of the Canny edges in the grouping process. In the same way, w_1 and w_2 are two constant values which weight the terms associated with the color and the disparity.

The grouping process is iterated until the number of nodes remains constant between two successive levels. Fig. 3c shows the set of regions associated to the image in Fig. 3a. It can be noted that the obtained regions do not always correspond to the set of natural objects presented in the image, but they provide an image segmentation which is more coherent with the human-based image decomposition than the one provided by a colour-based image decomposition (Fig. 3b).

3.2. Visual landmarks detection and normalization stage

The proposed grouping approach provides a partition of the input image into a set of regions. Among these regions, the approach selects those which satisfy certain conditions. Thus, selected regions cannot be located at the image border in order to avoid errors due to partial occlusions. They must also exhibit a relatively high color contrast with respect to its neighbor regions. Finally, large regions will be discarded because they could be more likely associated to non-planar surfaces. Specifically, the conditions imposed to be a landmark are the following:

- The area of a landmark must be less than a percentage U_s of the total area of the image.
- The bounding box of a selected region must not be located in an image border. The bounding box of a region is the minimum box which enclosed it.
- The contrast between the color of a landmark and its surrounding regions will be higher than a threshold value U_c.

Fig. 3c shows the set of selected landmarks associated to the image in Fig. 3a. Threshold values U_s and U_c have been fixed to 25% and 100%, respectively. It must be noted that threshold values employed at this stage are not very restrictive, i.e. similar results have been obtained using U_s values ranging from 20% to 30 % and U_c values ranging from 75 to 125.

Once the set of visual landmarks have been chosen, the region will be geometrically normalized. Firstly, the centroid of the image region $\mathscr{C}$ is estimated as

$$\mu=\frac{1}{|\mathscr{C}|}\int_{\mathscr{C}}\boldsymbol{x}d\mathscr{C} \tag{4}$$

where $|\mathscr{C}|$ is the area of the image region. Then, the matrix of second-order central algebraic moments (covariance matrix) of the region is calculated as

$$\Sigma=\frac{1}{|\mathscr{C}|}\int_{\mathscr{C}}(\boldsymbol{x}-\mu)(\boldsymbol{x}-\mu)^T d\mathscr{C}. \tag{5}$$

Once the covariance matrix is computed, the region is normalized so that the covariance matrix of the transformed region equals to

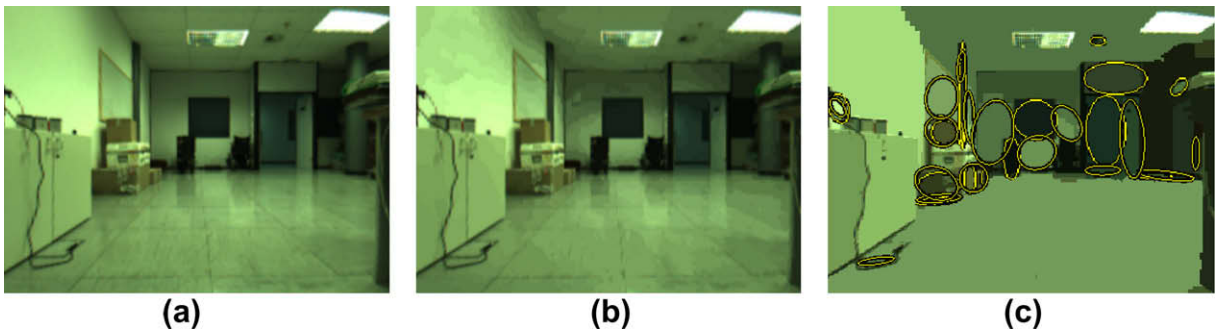

Fig. 3. (a) Original image (640 × 480 pixels size), (b) obtained regions after the pre-segmentation stage (16523 image regions), and (c) obtained regions after the perceptual grouping. Ellipses enclose the finally selected visual landmarks.

the identity matrix (Obdrzálek and Matas, 2006). This is achieved by transforming every region pixel by the inverse of the covariance matrix. Fig. 1c–d show the normalized versions of the image regions at Fig. 1a and b. Assuming local planarity of the detected region, this geometric normalization, together with the position of the centroid of the region, provides a rotation-variant image measurement. Therefore, if we also assume that the geometric changes induced by the camera motion can be described by an affine transformation, we will need to represent the image region by a rotational invariant descriptor to achieve a view-point invariant description. This descriptor will be presented at Section 3.3.

Finally, the image region is photometrically normalized. In this case, it is assumed that the combined effect of different scene illumination and capture system settings can be modeled by affine transformations of individual color channels. Then, the values of individual color channels are transformed to have zero mean and unit variance, allowing to represent a patch invariantly to photometric changes Obdrzálek and Matas, 2006.

3.3. Visual landmark description

Vision can be useful to avoid data association failures allowing landmarks to be characterized by a robust descriptor, i.e. a descriptor which will be invariant to illumination and viewpoint changes. As it was aforementioned, in our case we only need to describe the normalized image region by a rotational invariant descriptor to achieve view-point invariant description.

In this work, we assume that color distribution can provide an efficient feature for region description (Nummiaro et al., 2003). Besides, colour histograms can be easily quantized into a small number of bins to satisfy the low-computational cost imposed by real-time processing. In order to take into account the spatial information, which could be also useful, the regions can be masked with a kernel in the spatial domain Comaniciu et al., 2003. Thus, the appearance of the region is described by a set of scalar features, $\{s_i\}_{i=1\ldots N}$ which will be obtained from the image region defined by the visual landmark, $\mathscr{C}$. The value s_n of the nth bin is defined by

$$s_n = \frac{1}{\eta} \sum_{(x,y)_i \in \mathscr{C}} \mathscr{N}((x,y)_i)\delta(\gamma[I((x,y)_i] - n) \tag{6}$$

where $\mathscr{N}(\cdot)$ defines a Gaussian-based kernel function which assigns higher weights to the pixels near the centroid than pixels at the borders of the landmark, and η is a normalization constant ($\eta = \sum \mathscr{N}((x,y)_i)$) (Comaniciu et al., 2003). δ is the Kronecker delta function and γ is a quantization function, which associates with each observed pixel value $I((x,y)_i)$ a particular bin index. Finally, it must be commented that in our implementation, the CIELab colour space has been chosen at the hierarchical grouping algorithm and then also to characterize the colour of the landmark. We have also chosen to quantize the histogram in 16 bins, resulting in a landmark descriptor of 16 × 16 × 16 scalar values.

In order to compute the similarity between kernels (regions), Comaniciu et al. (2003) propose a metric derived from the Bhattacharyya coefficient. The distance between the discrete distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ associated to two visual landmarks is defined as:

$$d(\boldsymbol{p},\boldsymbol{q}) = \sqrt{1 - \rho[\hat{\boldsymbol{p}},\hat{\boldsymbol{q}}]} \tag{7}$$

where

$$\rho[\hat{\boldsymbol{p}},\hat{\boldsymbol{q}}] = \sum_{i=1}^{m} \sqrt{\hat{p}_i \cdot \hat{q}_i} \tag{8}$$

being $\hat{p}_i$ and $\hat{q}_i$ the corresponding bins of the discrete representations $\boldsymbol{p}$ and $\boldsymbol{q}$, respectively.

4. Experimental results

Different tests have been performed to evaluate the performance of the proposed modification of the BIP approach (Section 4.2), the ability of the proposed detector to extract salient regions (Section 4.3), the stability of the detected regions and the capacity of the descriptor to correctly characterize the detected regions. Different toolboxes and protocols have been also used to conduct these tests. Finally, we also provide in this Section an estimation of the parameters that the approach employs (Section 4.1) and the environment mapping framework where the proposed approach is currently applied with some preliminar qualitative results (Section 4.5).

4.1. Estimation of parameters

The proposed method requires choosing values for a set of parameters. These parameters are:

- The color threshold employed at the pre-segmentation stage. In all tests, this value has been fixed to a low value. This provides a over-segmentation of the input image, which groups the image pixels into a reduced set of blobs. If this threshold value is increased, then the compression factor will be also increased and the detection algorithm will run much faster. However, obtained blobs could contain image pixels of different colors. In all tests, a threshold value of 1.0 have been used.
- The threshold value T_{perc}, which determines the maximum distance between two nodes that are considered similar at the perception-based stage.
- The set of constant values (α, β, w_1 and w_2) employed at Eq. (3). In all our tests, α and β have been set to 0.1 and 1.0, respectively, and w_1 and w_2 have been set to 0.5 and 1.0, respectively.

- The set of parameters employed by the landmarks detection stage. The values of these parameters have been discussed in Section 3.2.

In order to choose a threshold value T_{perc} which can remain unaltered for the experiments show in Sections 4 and 4.4, different values were tested and the best value was chosen. In our tests, the best choice for this threshold was $T_{perc} = 20.0$.

4.2. Evaluation of the uBIP

With the aim of quantitatively evaluating the uBIP approach, we have compared it to other similar algorithms into a color-based segmentation framework. Three empirical methods have been selected: the shift variance (SV) proposed by Prewer and Kitchen (2001) and the F and Q functions (Marfil et al., 2006). Shift variance means that the image simplification produced by pyramid-based approaches varies when the base of the pyramid is shifted slightly. This is an undesirable effect, so the SV can be taken as a measurement of an algorithm quality. This method compares the segmentation results provided by a given algorithm on slightly shifted versions of the same image. To do that, we have taken a 128×128 pixels window in the center of the original image. We have compared the segmentation of this subimage with each segmented image obtained by shifting the window a maximum shift of 11 pixels to the right and 11 pixels down. Thus, there is a total of 120 images to compare with the original one. In order to perform each comparison between a segmented shifted image $\mathscr{X}_j$ and the segmented original one $\mathscr{O}$, the root mean square color difference (RMSD) is employed. The RMSD is defined as (Marfil et al., 2006).

$$\mathrm{RMSD}_{\mathscr{O}\mathscr{X}_j} = \sum_{i=1}^{N} \sqrt{(L^o_{i,j} - L^x_{i,j})^2 + (a^o_{i,j} - a^x_{i,j})^2 + (b^o_{i,j} - b^x_{i,j})^2} \tag{9}$$

where N is the number of pixels of the input image. L represents lightness values and a, b chrominance values corresponding to the original $o_{i,j}$ and shifted $x_{i,j}$ segmentation images expressed in the CIELab color space. Then, the SV is expressed as

$$SV = \frac{1}{120} \sum_{j=1}^{120} \mathrm{RMSD}_{\mathscr{O}\mathscr{X}_j} \tag{10}$$

It must be noted that the smaller the value of this parameter, the better the segmentation result should be.

The F and Q functions measure uniformity and homogeneity in the segmented regions, and other properties are required as simplicity, without too many small holes. Finally, these functions also consider that adjacent regions must present significantly different values for uniform characteristics. Specifically, the F function is computed as:

$$F(I) = \frac{1}{1000(N \cdot M)} \sqrt{R} \sum_{i=1}^{R} \frac{e_i^2}{\sqrt{A_i}} \tag{11}$$

I being the segmented image, NxM the image size and R the number of segmented regions. A_i and e_i are the area of the region i and its average color error, respectively. On the other hand, the Q function penalizes in a more rigid way the existence of small regions. It is defined by

$$Q(I) = \frac{1}{1000(N \cdot M)\sqrt{R}\sum_{i=1}^{R}\left[\frac{e_i^2}{1+\log A_i} + \left(\frac{R(A_i)}{A_i}\right)^2\right]} \tag{12}$$

$R(A_i)$ being the number of segmented regions with area equal to A_i.

For comparison purposes, five irregular pyramids have been employed: the BIP (Marfil et al., 2007a), the localized pyramid (LP) (Huart and Bertolino, 2005), the segmentation algorithm proposed by Lallich et al. (2003) (MP), the hierarchy of image partitions (HP) (Haxhimusa and Kropatsch, 2004) and the combinatorial pyramid (CoP) (Brun and Kropatsch, 2003). Six features have been evaluated: the number of obtained levels, which indicates the complexity of the obtained structure, the number of segmented regions, the F and Q functions, the shift-variance (SV) measure and the execution time. A set of 50 images from Waterloo and Coil 100 databases has been used (Marfil et al., 2006). The algorithms run in a 3 GHz Pentium IV PC. Table 1 shows the quantitative obtained results. It can be appreciated that the uBIP exhibits a significantly reduced shift-variance value. Besides, the F and Q values have been also improved. Regarding with the computational time, although it has been slightly increased with respect to the original BIP, it is still at least ten times less than in the rest of irregular pyramids.

Table 1
Quantitative segmentation results: hierarchy height, number of obtained regions, execution time, and F, Q and shift-variance values.

	Height	Regions	F	Q	SV	Time (sec)
LP	25.5	73.7	743.2	1011.5	30.2	2.75
MP	32.9	107.6	650.1	818.5	29.3	3.42
HP	11.4	76.1	670.3	955.1	28.4	4.23
CoP	74.2	91.2	630.7	870.2	30.5	2.85
BIP	8.7	83.6	720.2	1090.1	44.1	0.20
uBIP	9.3	60.5	700.1	950.3	24.3	0.23

4.3. Evaluating the performance of the proposed salient region detector

To evaluate the ability of the proposed detector to extract salient regions, we compared the discriminant saliency maps obtained from a collection of natural images to the eye fixation locations recorded from human subjects, in a free-viewing task. Specifically, we have employed the human fixation database from Bruce and Tsotsos (2006). This data set was obtained from eye tracking experiments performed while subjects observed 120 different color images (see Bruce and Tsotsos (2006) for further details). The colour contrast measure is employed as the feature to define our saliency map and, to measure the performance of the approach, obtained saliency maps are first quantized into a binary image: pixels with larger saliency values than a threshold are classified as fixated while the rest of the pixels in that image are classified as non-fixated (Tatler et al., 2005; Gao et al., 2008). Human fixations are then used as ground truth and, by varying the threshold, a receiver operator characteristic (ROC) curve can be drawn. The area under the curve indicates how well the saliency map predicts actual human eye fixations. Fig. 5 shows the ROC curve obtained for the proposed approach.

The quantitative performance of the proposed approach is shown in Table 2. In this table we also summarized the results obtained using the algorithms of Itti and Koch (2000), obtained using the Matlab saliency toolbox Walther and Koch, 2006[1], Bruce and Tsotsos (2006) and Gao et al. (2008). Besides, as an absolute benchmark, the 'inter-subject' ROC area is also included (Gao et al., 2008; Harel et al., 2007). It can be noted that obtained results are similar to the other detectors.

4.4. A comparative study with other local feature detectors and descriptors

In order to compare our method to other local feature detectors and descriptors, images, Matlab code to carry out the performance

[1] http://www.saliencytoolbox.net/index.html.

Fig. 4. (a) Reference image and detected landmarks, (b–f) images matched against the reference image. The colour of the ellipses determines if the associated region has been matched to a reference landmark (displayed in yellow) or not (displayed in blue) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

tests, and binaries of the approaches have been downloaded[2]. The database is composed by eight different image sets that represent five changes in imaging conditions (viewpoint changes, scaling, image blur, jpeg compression and illumination changes). Image sets can be grouped into two different scene types. Thus, one scene type contains homogeneous regions which present distinctive boundaries (structured scenes), and the other contains repeated textures of different forms (textured scenes). As our approach is based on structure cues in images, it is reasonable that it exhibits a superior performance on structured scenes. Fig. 6 shows an example from each image set. It must be noted that the set of parameters employed by the proposed approach has not been modified to deal with the different image sets (see Section 4.1).

For the detectors, we use the repeatability score, as described by Mikolajczyk et al. (2006). The objective of this test is to measure how many of the detected regions are found in images under different transformations, relative to the lowest total number of regions detected (where only the part of the image that is visible in both images is taken into account). In all cases, the ground truth is provided by mapping the regions detected on the images in a set to the image of highest quality of this set (reference image) using homographies. The measure of repeatability is the relative amount of overlap between regions detected in the reference image and in the other image. This region is projected onto the reference image using the homography relating the images. It must be noted that

[2] http://www.robots.ox.ac.uk/~vgg/research/affine.

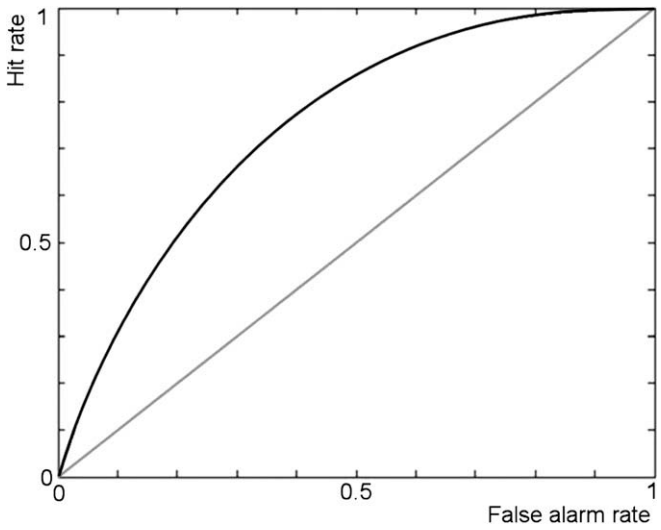

Fig. 5. The ROC curve provided by the proposed approach.

Table 2
ROC areas for different saliency models with respect to all human fixations.

Saliency model	ROC area
Gao et al. (2008)	0.7694
Itti and Koch (2000)	0.7277
Bruce and Tsotsos (2006)	0.7547
Proposed	0.7593
Inter-subject	0.8766

the output for our detector is a set of arbitrarily shaped regions. However, for the purpose of the comparisons using the Matlab code mentioned above, the output region of all detectors are represented by an ellipse. These ellipses have the same first and second moments as the detected regions.

The proposed detector is compared to the difference of Gaussian (DoG) (Lowe, 2004), the Hessian-affine detector (Mikolajczyk and Schmid, 2002), the maximally stable extremal region detector (MSER) (Matas et al., 2002), the intensity extrema-based region detector (IBR) (Tuytelaars and Van Gool, 2004) and the Fast-Hessian (Bay et al., 2006). For all experiments, the default parameters

Table 3
Number of detected regions and computation times for different detectors for GRAF image (see Fig. 6).

Detector	Number of regions	Run time (s)
DoG	1520	0.39
Hessian-affine	1649	2.43
Fast-Hessian	1418	0.12
MSER	533	0.56
IBR	679	9.77
Proposed	147	0.32

given by the authors are used for each detector. It must be noted that disparity values are not available, so the parameter value w_2 has been set to 0.0 in these cases. From Table 3, it can be noted that the detectors generate very different numbers of regions, although this also depends on the image type. Thus, some of them provide good results to structured scenes (e.g. the proposed approach and the MSER) and others to more textured scenes (e.g. Hessian-affine). Table 3 shows that computation times are also very different. They have been measured on a Pentium 4.2 GHz Linux PC, for the GRAF image shown in Fig. 6, which is 800×640 pixels.

The repeatability for four sets of images are illustrated in Fig. 7. Similar results are obtained for the rest of sequences. These results show that the proposed detector ranks similar to the rest of approaches when it deals with structured images. In these images, only few regions are detected and the thresholds can be set very sharply, resulting in very stable regions. On the contrary, the scores associated to textured images are significantly bad when compared to the point-based detectors (see Fig. 7, the WALL set).

Finally, the kernel-based descriptor is evaluated using the recall-precision criterion for image pairs, i.e. the number of correct and false matches between two images (Mikolajczyk and Schmid, 2005). Recall is defined as the number of correctly matched regions with respect to the number of corresponding regions between two images of the same scene. The precision is defined as the number of correct matches with respect to the total number of matches. The results are represented with recall versus 1-precision. Fig. 8 shows the results for three sets of images. Regions have been detected using the proposed approach. Two regions are matched if the distance between their descriptors is below a threshold U. The value of this threshold is varied to obtain the curves (see Mikolajczyk and Schmid (2005) for further details). Compared descriptors are the SURF-128 (Bay et al., 2006), SIFT (Lowe, 1999) and the cross correlation (evaluated for a path of 11×11 pixels

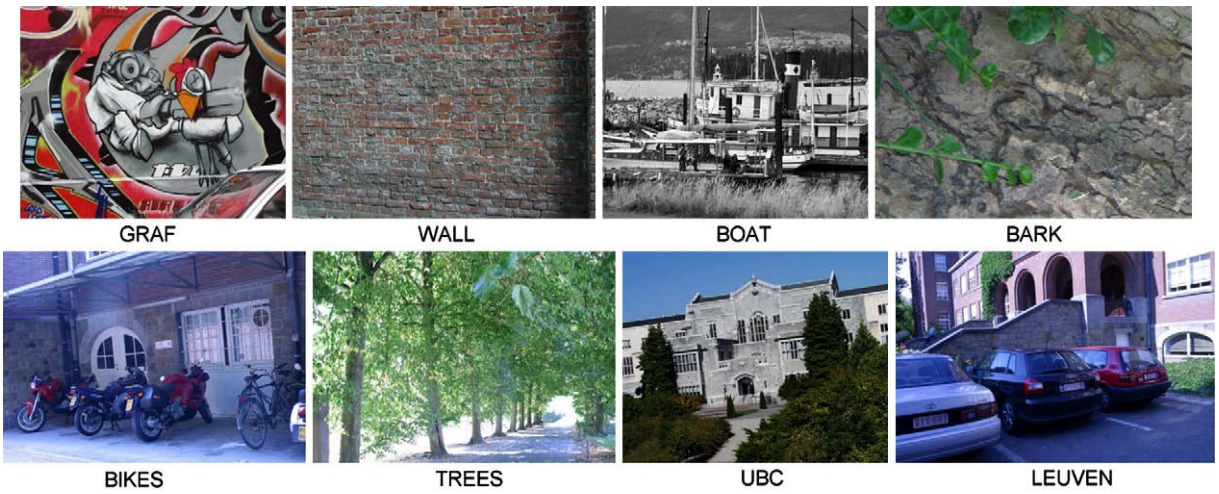

Fig. 6. Image examples of the eight sets used for comparison purposes.

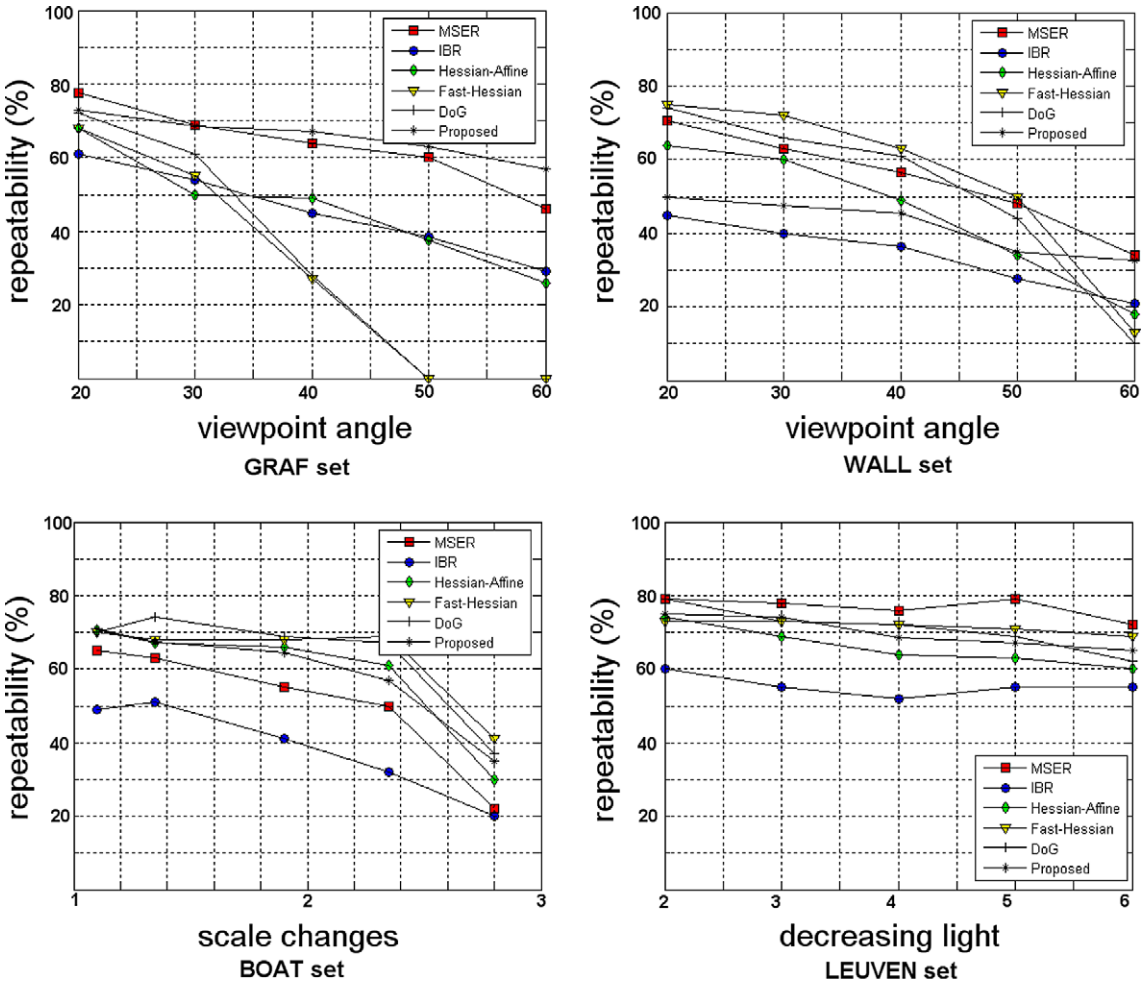

Fig. 7. Repeatability scores for GRAF, WALL, LEUVEN and BOAT sequences (see Fig. 6).

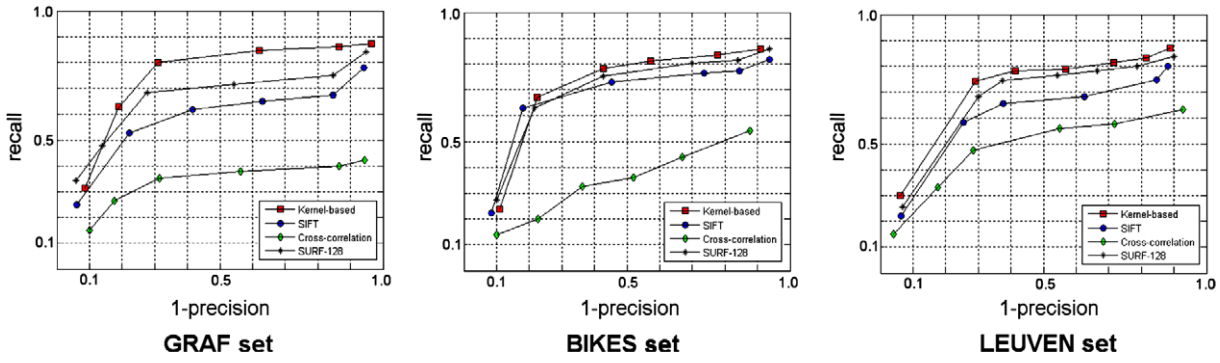

Fig. 8. Recall vs. 1-precision curves for GRAF, BIKES and LEUVEN sequences (see Fig. 6).

centered at the centroid of the detected region), which are three of the most employed visual landmark descriptors (see Section 2).

From the results, it can be noted that the kernel-based descriptor performs better than the rest of descriptors. The number of regions is significantly low, and this implies that regions are usually not overlapped. Besides, although the textured scenes contain similar motifs, the regions capture distinctive image variations. For these reasons, distribution-based descriptors like the kernel-based one or the SIFT, exhibit a good performance. On the other hand, the size of the kernel-based descriptor is significantly larger than the rest of descriptors. This implies more computational time and storage resources, which are compensated by its good performance, specially when dealing with real acquired images.

4.5. Testing the approach in a environment mapping framework

To test the validity of the landmark detector, data was collected with an ActiveMedia Pioneer 2AT robot mounted with an

stereoscopic camera. The robot was driven through different environments while capturing real-life stereo images. The stereo head is the STH-MDCS from Videre Design, a compact, low-power colour digital stereo head with an IEEE 1394 digital interface. It consists of two 1.3 megapixel, progressive scan CMOS imagers mounted in a rigid body, and a 1394 peripheral interface module, joined in an integral unit. The camera was mounted at the front and top of the vehicle at a constant orientation, looking forward. Images were restricted to 640×480 or 320×240 pixels.

To qualitatively check the viewpoint invariance of our detector, we take images of an scene starting from head on (reference pose) and gradually increasing the viewing angle and/or the distance to the reference pose. Results of one of these experiments are shown in Fig. 4, where each visual landmark is represented by an ellipse. It must be noted that those image regions which are so far from the robot have been discarded as the system cannot provide a good estimation of their disparity values. For each image, visual landmarks are extracted and matched to the landmarks found in the zero degrees reference image (Fig. 4a). A nearest neighbor-based matching strategy has been used, i.e. two regions $\boldsymbol{A}$ and $\boldsymbol{B}$ are matched if the descriptor D_B is the nearest neighbor to D_A and if the distance between them is below a threshold U. With this approach, a descriptor has only one match. The color of the ellipses represented in Fig. 4b–f determines if the associated region has been matched to a reference landmark (displayed in yellow) or not (displayed in blue). Experimental results show that the system can deal with changes in viewpoint up to 50 or 60 degrees and with scale changes of 2–2.5. It can be also noted that the number of matches found slightly decreases with increasing scale change.

5. Conclusions and future work

We have presented a visual landmark detection scheme whose performance is similar to the current state-of-the art algorithms, both in speed and accuracy. To obtain these landmarks, a hierarchical segmentation approach has been developed. Thus, the contents of the image are described using multiple representations with decreasing resolution. Pyramid segmentation algorithms exhibit interesting properties when compared to segmentation algorithms based on a single representation: local operations can adapt the pyramid hierarchy to the topology of the image, allowing the detection of global regions of interest and representing them at low resolution levels. From the segmented regions, a set of significant regions are selected as landmarks. These landmarks have been characterized by a kernel-based descriptor, whose performance is comparable or even better than for other similar approaches. The main disadvantage of this descriptor is its high size. In order to reduce it, instead of using the kernel-based histograms, future work will be focused on testing the application of principal components analysis (PCA) to this descriptor. This technique has been employed by Ke and Sukthankar (2004) to the normalized gradient patches provided by the SIFT detector. Future work will also include the development and evaluation of an EKF-based algorithm for robot localization and environment mapping using the extracted landmarks for scene recognition and loop-closing.

Acknowledgement

This work has been partially granted by the Spanish Ministerio de Ciencia e Innovación (MCINN) and FEDER funds, and by the Junta de Andalucía, under Projects nos. TIN2008-06196 and P07-TIC-03106, respectively.

References

Ahn, S., Choi, M., Choi, J., Chung, W., 2006. Data association using visual object recognition for EKF-SLAM in home environment. In: Proc. of the IEEE/RSJ International Conf. Intelligence Robots Systems, pp. 2588–2594.

Asmar, D., Zelek, J., Abdallah, S., 2006. Tree trunks as landmarks for outdoor vision SLAM. In: Proc. Conf. on Computer Vision and Pattern Recognition Workshop, pp. 196–203.

Aziz, M.Z., Mertsching, B., 2007. Color saliency and inhibition using static and dynamic scenes in region based visual attention. In: Paletta, L., Rome, E. (Eds.), WAPCV 2007, LNAI 4840. Springer, Heidelberg, pp. 234–250.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. SURF: Speeded Up Robust Features. Computer Vision – ECCV 2006, LNCS. 3951/2006, pp. 407–417.

Bruce, N., Tsotsos, J., 2006. Saliency based on information maximization. In: Weiss, Y., Schölkopf, B., Platt, J. (Eds.), Advances in Neural Information Processing Systems, vol. 18. MIT Press, Cambridge, MA, pp. 155–162.

Brun, L., Kropatsch, W., 2003. Construction of combinatorial pyramids. In: Proc. of Graph-based Representation in Pattern Recognition, pp. 1–12.

Comaniciu, D., Ramesh, V., Meer, P., 2003. Kernel-based object tracking. IEEE Trans. Pattern Anal. Machine Intell. 25 (5), 564–577.

Davison, A., Murray, D., 2002. Simultaneous localization and map-building using active vision. Pattern Anal. Machine Intell. 24 (7), 865–880.

Elinas, P., Sim, R., Little, J., 2006. σSLAM: Stereo vision SLAM using the rao-blackwellised particle filter and a novel mixture proposal distribution. In: Proc. IEEE Int. Conf. on Robotics and Automation, pp. 1564–1570.

Eriksen, C.W., Yenh, Y.Y., 1985. Allocation of attention in the visual field. J. Exp. Psychol.: Hum. Percept. Perform. 11 (5), 583–597.

Folkesson, J., Jensfelt, P., Christensen, H., 2005. Graphical SLAM using vision and the measurement subspace. In: IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 325–330.

Frintrop, S., Jensfelt, P., Christensen, H., 2006. Attentional landmark selection for visual SLAM. In: IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 2582–2587.

Gao, D., Mahadevan, V., Vasconcelos, N., 2008. On the plausibility of the discriminant center-surround hypothesis for visual saliency. J. Vision 8 (7), 1–18, 13.

Harel, J., Koch, C., Perona, P., 2007. Graph-based visual saliency. In: Schölkopf, B., Platt, J., Hoffman, T. (Eds.), Advances in Neural Information Processing Systems, vol. 19. MIT Press, Cambridge, MA, pp. 545–552.

Harris, C., Stephens, M.J., 1988. A combined corner and edge detector. In: Alvey Vision Conference, pp. 147–152.

Harris, C., 1992. Geometry from visual motion. In: Blake, A., Yuille, A. (Eds.), Active Vision. MIT Press, Cambridge.

Haxhimusa, Y., Glantz, R., Kropatsch, W.G., 2003. Constructing stochastic pyramids by MIDES – maximal independent directed edge set. In: Hancock, E., Vento, M. (Eds.), 4th IAPR-RC15 Workshop on GbR in Pattern Recognition, LNCS, vol. 2726. Springer Verlag, pp. 35–46.

Haxhimusa, Y., Kropatsch, W., 2004. Segmentation graph hierarchies. In: Proc. of IAPR Int. Workshop on Syntactical and Structural Pattern Recognition and Statistical Pattern Recognition, pp. 343–351.

Hayet, J.B., Lerasle, F., Devy, M., 2003. Visual landmarks detection and recognition for mobile robot navigation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 313–318.

Horswill, I., 1993. Polly: A vision-based artificial agent. In: Proc. Nat. Conf. Artif. Intelligence, pp. 824–829.

Huart, J., Bertolino, P., 2005. Similarity-based and perception-based image segmentation. In: Proc. IEEE Int. Conf. on Image Processing 3(3), pp. 1148–1151.

Itti, L., Koch, C., 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res. 40, 1489–1506.

Itti, L., 2002. Real-time high-performance attention focusing in outdoors color video streams. In: Proc. SPIE Human Vision and Electronic Imaging (HVEI'02), pp. 235–243.

Jensfelt, P., Kragic, D., Folkesson, J., Björkman, 2006. A framework for vision based bearing only 3D SLAM. In: Proc. IEEE Int. Conf. on Robotics and Automation, pp. 1944–1950.

Jeong, W., Lee, K., 2005. CV-SLAM: A new ceiling vision-based SLAM technique. In: IEEE/RSJ Int. Conf. Intelligence Robots Systems, pp. 3070–3075.

Ke, Y., Sukthankar, R., 2004. PCA-SIFT: A more distinctive representation for local image descriptors. In: Proc. Computer Vision Pattern Recognition, vol. 2, pp. 506–513.

Kim, J., Chung, M., 2005. Absolute stereo SFM without stereo correspondence for vision based SLAM. In: Proc. IEEE Int. Conf. on Robotics and Automation, pp. 3360–3365.

Kim, G., Kim, J., Hong, K., 2005. Vision-based simultaneous localization and mapping with two cameras. In: IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 1671–1676.

Koch, C., Ullman, S., 1985. Shifts selective visual attention: Towards the underlying neural circuitry. Human Neurobiol. 4, 219–227.

Lallich, S., Muhlenbach, F., Jolion, J., 2003. A test to control a region growing process within a hierarchical graph. Pattern Recognition 36, 2201–2211.

Lin, Z., Kim, S., Kweon, I., 2005. Robust invariant features for object recognition and mobile robot navigation. In: Proc. of IAPR Conf. Machine Vision Applications.

Lingemann, K., Surmann, H., Nüchter, A., Hertzberg, J., 2004. Indoor and outdoor localization for fast mobile robots. In: Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robot and Systems, pp. 2185–2190.

Lowe, D., 1999. Object recognition from local scale invariant features. In: Proc. 7th Int. Conf. on Computer Vision, pp. 1150–1157.
Lowe, D., 2004. Distinctive image features from scale-invariant keypoints, cascade filtering approach. Int. J. Comput. Vision 60, 91–110.
Maki, A., Nordlund, P., Eklundh, J.O., 2000. Attentional scene segmentation: Integrating depth and motion. Computer Vision and Image Understanding 78 (3), 351–373.
Marfil, R., Rodríguez, J.A., Bandera, A., Sandoval, F., 2004. Bounded irregular pyramid: A new structure for colour image segmentation. Pattern Recognition 37 (3), 623–626.
Marfil, R., Molina-Tanco, L., Bandera, A., Rodríguez, J.A., Sandoval, F., 2006. Pyramid segmentation algorithms revisited. Pattern Recognition 39 (8), 1430–1451.
Marfil, R., Molina-Tanco, L., Bandera, A., Sandoval, F., 2007a. The construction of bounded irregular pyramids with a union-find decimation process. Lecture Notes Comput. Sci. 4538, 307–318.
Marfil, R., Molina-Tanco, L., Rodríguez, J.A., Sandoval, F., 2007b. Real-time object tracking using bounded irregular pyramids. Pattern Recognition Lett. 28 (9), 985–1001.
Marfil, R., Bandera, A., Sandoval, F., 2007c. Perception-based image segmentation using the bounded irregular pyramid. In: Proc. of the 29th German Society of Pattern Recognition Symposium (DAGM'07), pp. 244–254.
Matas, J., Chum, O., Urban, M., Pajdla, T., 2002. Robust wide-baseline stereo from maximally stable extremal regions. In: Proc. British Machine Vision Conference, pp. 384–393.
Mikolajczyk, K., Schmid, C., 2002. An affine invariant interest point detector. In: Proc. 7th European Conference on Computer Vision, pp. 128–142.
Mikolajczyk, K., Schmid, C., 2005. A performance evaluation of local descriptors. Pattern Anal. Mach. Intell. 27, 1615–1630.
Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L., 2006. A comparison of affine region detectors. Int. J. Comput. Vision 65, 43–72.
Milanese, R., 1993. Detecting Salient Regions in An Image: From Biological Evidence to Computer Implementation. PhD Thesis, University of Geneva.
Moravec, H.P., 1977. Towards automatic visual obstacle avoidance. In: Proc. 5th Int. Joint Conf. on Artificial Intelligence, p. 584.
Newman, P., Ho, K., 2005. SLAM-loop closing with visually salient features. In: Proc. IEEE Int. Conf. Robotics and Automation, pp. 644–651.
Nummiaro, K., Koller-Meier, E., Roth, D., Van Gool, L., 2003. Color-based object tracking in multi-camera environments. In: Proc. of the 25th German Society of Pattern Recognition Symposium (DAGM03), pp. 591–599.
Obdrzálek, S., Matas, J., 2006. Object recognition using local affine frames on maximally stable extremal regions. In: Toward Category-Level Object Recognition, LNCS, vol. 4170, pp. 83–104.
Orabona, F., Metta, G., Sandini, G., 2007. A proto-object based visual attention model. In: Paletta, L., Rome, E. (Eds.) WAPCV 2007, LNAI 4840. Springer, Heidelberg, pp. 198–215.
Ouerhani, N., Hügli, H., 2005. Robot self-localization using visual attention. In: Proc. IEEE Int. Symposium Computational Intelligence in Robotics and Automation, pp. 309–314.
Prewer, D., Kitchen, L.J., 2001. Soft image segmentation by weighted linked pyramid. Pattern Recognition Lett. 22 (2), 123–132.
Se, S., Lowe, D., Little, J., 2002. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. Int. J. Robotics Res. 21 (8), 735–758.
Se, S., Lowe, D., Little, J., 2005. Vision-based global localization and mapping for mobile robots. IEEE Trans. Robotics 21 (3), 364–375.
Shi, J., Tomasi, C., 1994. Good features to track. In: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 593–600.
Siagian, C., Itti, L., 2007. Rapid biologically-inspired scene classification using features shared with visual attention. IEEE Trans. Pattern Anal. Mach. Intell. 29 (2), 300–312.
Sun, Y., Fisher, R.B., 2003. Object-based visual attention for computer vision. Artif. Intell. 146 (1), 77–123.
Tamimi, H., Andreasson, H., Treptow, A., Duckett, T., Zell, A., 2006. Localization of mobile robots with omnidirectional vision using Particle Filter and iterative SIFT. Robot. Auton. Syst. 54, 758–765.
Tatler, B.W., Baddeley, R.J., Gilchrist, I.D.M., 2005. Visual correlates of fixation selection: Effects of scale and time. Vision Res. 45, 643–659.
Treisman, A.M., Gelade, G., 1980. A feature integration theory of attention. Cognitive Psychol. 12 (1), 97–136.
Tuytelaars, T., Van Gool, L., 2004. Matching widely separated views based on affine invariant regions. Int. J. Comput. Vision 59 (1), 61–85.
Vázquez-Martín, R., del Toro, J., Bandera, A., Sandoval, F., 2005. Data- and model-driven attention mechanism for autonomous visual landmark acquisition. In: Proc. IEEE Int. Conf. Robotics and Automation, pp. 3372–3377.
Walther, D., Koch, C., 2006. Modeling attention to salient proto-objects. Neural Networks 19, 1395–1407.
Wang, J., Zha, H., Cipolla, R., 2006. Coarse-to-fine vision-based localization by indexing scale-invariant features. IEEE Trans. Systems Man Cybernet. – Part B 36 (2), 413–422.