Affine image region detection and description

R. Vázquez-Martín, R. Marfil and A. Bandera

***Abstract*—This paper describes a novel approach for affine invariant region detection and description. At the detection stage, a hierarchical clustering mechanism is employed to group image pixels into regions. This process is based on the Bounded Irregular Pyramid (BIP) and takes into account a colour contrast measure, internal region descriptors and attributes of their shared boundaries. High-contrasted regions are selected as salient regions. On the other hand, geometrically and photometrically normalized regions are represented by a kernel-based descriptor. The lenght descriptor is reduced by applying Principal Component Analysis (PCA). The protocol proposed by Mikolajczyk et al. [17], [18] has been conducted to compare the proposed approach with other similar methods. Experimental results prove that the performance of our proposal is high in terms of computational consuming and distinguished region detection and description abilities.**

***Index Terms*—salient regions, feature detection, affine invariant regions, feature description.**

I. Introduction

IMAGE matching is defined in artificial vision as the process of bringing two images into agreement so that corresponding items in the two images correspond to the same real, physical region of the scene. The similarity may be applied to global features derived from the original images. However, this is not the most robust solution when images are taken from different viewpoints. In this work, the image matching problem is accomplished from a feature-based strategy, where images are analyzed first in order to extract some distinguished features. Detected features or image regions are then characterized by a descriptor which will be subsequently employed to solve the matching problem.

In this paper, we propose a novel approach for affine, distinguished image regions detection and description. The core of the detector is a hierarchical algorithm for perceptual grouping of the image pixels. Image segmentation is not a robust process, and obtained results can change depending on the illumination conditions or the viewpoint. However, there are a set of image regions whose properties allow them to be robustly detected in despite of these changes. The aim of the proposed approach is to find these salient regions. On the other hand, to solve the correspondence problem stated when different views of the same scene are compared, detected image regions can be characterized using information obtained from the image. In our proposal, a weighted histogram is employed. This descriptor provides satisfactory results, specially when used in real acquired images. However, to reduce the lenght of the obtained descriptor, PCA is applied.

R. Vázquez-Martín, R. Marfil and A. Bandera are with Department of Electronic Technology, University of Málaga, Campus de Teatinos 29071-Málaga, Spain.
E-mail: rvmartin@uma.es

A. Related work

Given a set of images taken from different viewpoints, the process of finding the projections on each image of real 3D surface patches can be useful for a large number of applications, such as object recognition, robot localization or wide baseline matching for stereo pairs. Among other issues, this process must deal with the problem that image regions associated to the projections change covariantly with the class of transformation induced by the viewpoint change. When the viewpoint change can be approximated by an affine transformation, approaches which solve this problem are called affine region detectors [18].

The detection of regions which change covariantly with affine transformations was described in detail by Mikolajczyk et al. [18]. In this work, the authors provide a review of affine covariant region detectors, and compare their performance on a set of test images under varying imaging conditions. The requirement for these detectors is that they must provide regions whose shapes depend on the underlying image features, so that they correspond to the projections of the same 3D surface patch on the different images. Although the boundaries of these covariant regions do not have to be associated to changes in image features such as colour or texture, some of the approaches described in [18] look for these abrupt changes. Thus, the intensity extrema-based region detector (IBR) [19] starts from intensity extrema and studies a intensity-based function along rays emanating from this extrema to define a region of arbitrary shape. The region is delineated by the image points defined over these rays where the intensity suddenly increases or decreases. A *maximally stable extremal region* (MSER) [15] is a connected component of an appropriately thresholded image where all internal pixels have either higher or lower intensity than all the pixels on its outer boundary. Among these extremal regions, the *'maximally stable'* ones are those corresponding to thresholds were the relative area change as a function of relative change of threshold is at a local minimum.

To match the projections of a real 3D surface on a set of images taken from different viewpoints does not only require to find distinguished image features, but also to solve the correspondence problem established among these sets of features. This issue can be addressed by characterizing the distinctive regions in terms of certain patterns. For this reason, the computation of feature descriptors is done as a separate step from that of feature detection. For local interest point (corner) detection, features are usually described using their associated image patches. Then, Normalized Sum-of-Squared-Differences (NSSD) is employed to find the best matchings. On the other hand, descriptor for scale invariant features are computed at the distinctive points with the associated

scale. Gaussian derivatives computed at the characteristic scale over image patches can be employed to achieve invariance to image rotation. However, among the large number of proposed techniques, the distribution-based descriptors are probably the most used ones. Thus, inside the scale invariant feature transform (SIFT), Lowe [11] proposed to compute a histogram of local oriented gradients around the interest point and scores the bins in a 128-dimensional vector. Mikolajczyk and Schmid [17] proposed a variant of SIFT, called gradient location and orientation histogram (GLOH), which has proved to be more distinctive but also computationally more expensive. The SURF approach includes a region descriptor which uses a distribution of Haar-wavelet responses within the interest points neighbourhood [1].

B. Overview of the proposal

This paper describes a novel approach for affine region detection which extends the idea of looking for abrupt changes. However, instead of changes in intensity or colour (edges), our approach looks for image boundaries which delimitate high-contrasted regions of data-dependent shape. To detect these boundaries, we use a hierarchical clustering scheme which presents two stages: firstly, it groups neighbour image pixels into blobs of homogeneous colour and then, it merges these blobs using a more complex similarity criterion. Basically, this criterion complements a contrast measure defined between regions with image edges detected using the Canny detector, with internal region descriptors and with attributes of their shared boundaries. Finally, it must be noted that the hierarchical clustering algorithm represents the input image at different levels with decreasing resolution. This hierarchy constitutes a scale-space representation where salient regions could be detected at different scales. On the other hand, to describe the detected regions, we have chosen as feature space the colour 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 salient regions are characterized by spatially masking them with an isotropic kernel. Finally, we have applied PCA to the obtained kernel-based histograms to reduce its large lenght.

C. Summary and comparison with other approaches

Unlike affine region detectors based on interest point detectors such as the Harris–Affine or the Hessian–Affine techniques [18], our proposal provides complementary image information. Thus, it is more closely related to those region detectors based on image intensity analysis, such as the MSER and IBR approaches. The main difference with those approaches is that it searches for high contrasted regions of uniform properties using a hierarchical structure instead of intensity extrema in a 2D image as do the MSER and IBR detectors. Using this segmentation strategy, it is possible to work in scale-space, improving repeatability for significant scale changes.

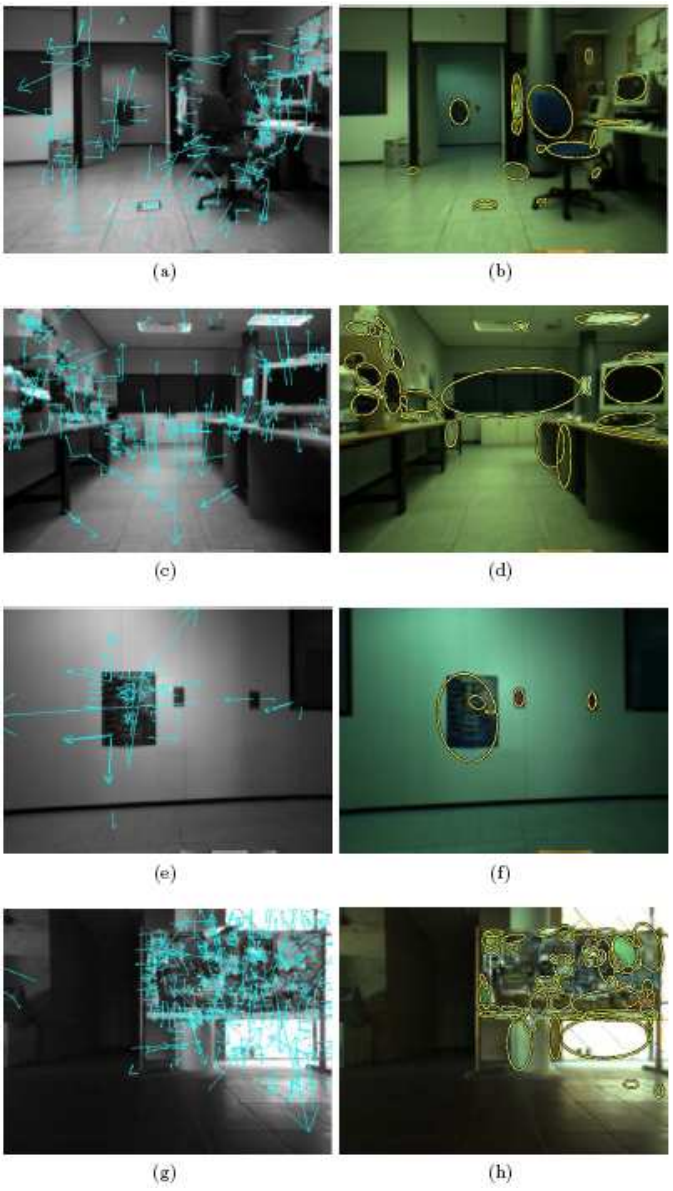

Fig. 1. a-c-e-g) SIFT features and b-d-f-h) affine covariant regions detected by the proposed approach in different environments and situations.

On the other hand, Fig. 1 illustrates the main difference with respect to the popular Difference–of–Gaussians (DoG) detector, the scale invariant feature detector used by SIFT. Typically, the DoG provides an immense number of keypoints, as can be seen in Figs. 1a-c-e-g. On the contrary, the proposed approach detects a far smaller set of regions. This is due to the grouping process inherent to any region–based feature detector. In this case, this grouping merges different image blobs in accordance with their similarity in colour and the shared boundary. These regions have a greater underlying semantic significance than the keypoints detected using a scale invariant detector. Using the proposed approach the image is described by a more organized set of features that allows a reliable matching since comparatively little information is needed to describe an scene.

The rest of the paper is organized as follows: The proposed

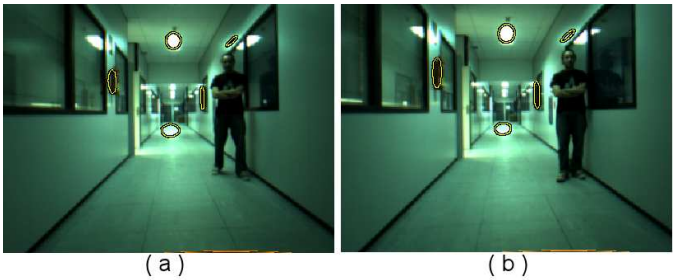

Fig. 2. a-b) Regions generated by the proposed detector on two images taken from a mobile robot. Representing ellipses have been chosen to have the same first and second moments as the originally arbitrarily shaped region (see Fig. 11 for more examples).

approach for the acquisition and description of salient regions is described in Sections II and III. Experimental results are provided in Section IV, where the segmentation algorithm is evaluated and it is also provided an example of the application of the approach in an environment mapping framework. The results of a comparative study of the proposed approach with other similar approaches are given in Section V. Finally, the paper concludes along with discussions in Section VI.

II. Hierarchical clustering approach for salient regions detection

The proposed approach employs a hierarchical graph-based clustering algorithm to detect the high-contrasted regions of the input image. In this hierarchy, the input image defines the base level, which is arranged as a graph where each pixel is a node and neighbourhood relationships are encoded as arcs (intra-level arcs). Upper hierarchy levels are encoded as undirected graphs where the nodes are generated by grouping a set of nodes of the level below and the arcs encode their adjacency relationships. If intra-level arcs represent the neighbourhood of each node at the same level, another set of arcs establish a dependence relationship between each node of level l+1 and a set of nodes at level l. These relationships may be extended by transitivity down to the base level. The set of pixels linked to a node is named its receptive field. The receptive field defines the embedding of this node on the original image. This hierarchy defines an irregular pyramid [5], [8], 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. In order to speed up the hierarchical clustering process, the employed irregular pyramid combines the classical irregular simple graph with a regular structure. This regular decimation process is only applied in the homogeneous parts of the image. Then, each graph G_l has a regular part which built from G_{l-1} using a 2x2/4 regular decimation procedure and an irregular part which is built from G_{l-1} using an union-find decimation process. This also implies that there are two types of nodes in our structure: nodes belonging to the 2x2/4 regular part (regular nodes) and nodes belonging to the irregular part (irregular nodes). Experimental results demonstrate that the shape of the obtained salient regions is adapted to real items of the scene, being no affected by the regular tessellation (see Fig. 2).

Each level of the proposed pyramid is computed in three steps:

- 2x2/4 regular decimation process: if four regular adjacent nodes of level l have similar colour, a new regular node is created at l+1.
- Irregular node generation process: any regular or irregular node of level l which is not linked to a node at l+1 is included in a union-find grouping process [15], [14]. This union-find process only generates irregular nodes at level l+1.
- Intra-level edge generation in G_l+1: the edges of G_l+1 are computed taking into account the neighbourhood of nodes in G_l.

In order to speed up the building process, the pyramid can be initialized with a first image partition. This initial partition divides the input image in a set of homogeneous regions, constituting an over-segmentation of the input image. Typically, this pre-segmentation process only generates the first pyramid level, and the rest of levels are built using a more complex grouping criterion. Our proposal accomplishes the pre-segmentation and the subsequent clustering process into an irregular pyramid as two consecutive stages. The first stage employs a colour distance to group the image pixels into a set of blobs whose spatial distribution is physically representative of the image content. It must be noted that the hierarchy automatically stops growing when it is no longer possible to link together any more nodes because they are not similar in colour. The set of nodes which are not linked to any node at upper levels define a partition of the input image (see Fig. 3). Then, the second stage clusters 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. 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 pyramid structure is used: the roots of the pre-segmented blobs are considered as irregular nodes which constitute the first level of the grouping multiresolution output. However, if the distance between two nodes in the pre-segmentation stage is based on a colour criterion, in order to achieve this second grouping process, a more complex distance must be defined. This distance has two main components: the colour contrast between image blobs and the edges of the original image computed using the Canny detector. Then, the distance between two nodes n_i and n_j, $\Upsilon(n_i, n_j)$, is defined as

$$\Upsilon(n_i, n_j) = \frac{d(n_i, n_j) \cdot min(b_i, b_j)}{\alpha \cdot c_{ij} + \beta(b_{ij} - c_{ij})} \quad (1)$$

where $d(n_i, n_j)$ is the colour distance between n_i and n_j, 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. We set these parameters to 0.1 and 1.0 respectively.

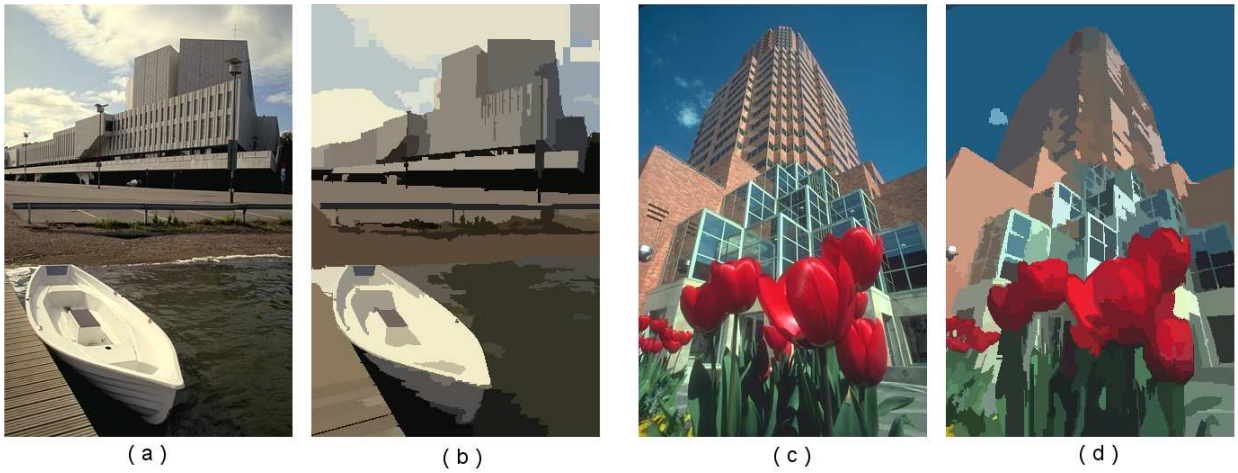

Fig. 3. Presegmentation stage: a-c) Original images; and b-d) colorized pre-segmentation images.

When the whole irregular pyramid is built, there is a set of nodes which are not linked to any parent node at upper levels. The union of the receptive fields of these nodes will generate a partition of the input image. It must be noted that these root nodes can be located at the different pyramid levels, i.e. they are selected at different scales. Among these root nodes, the set of keynodes will be chosen. Two constraints are required to be a keynode: its receptive field must not be in contact with the image border, and its colour must be different from the colour of its neighbours. This second condition is always satisfied by a root node, but imposing a minimum value to this contrast measure the algorithm looks for region locations and scales that can be repeatably associated under differing views of the same object. That is, among the image regions associated to the root nodes, the 'maximally stable' ones are those corresponding to nodes whose colour is very different to the colour of their neighbours.

Once the set of visual features has been chosen, each region is normalized geometrically using the covariance matrix. The aim is that the covariance matrix of the transformed region will be equal to the identity matrix. This is achieved by transforming every region pixel by the inverse of the covariance matrix of the original region [18]. Assuming local planarity of the detected region, this geometric normalization, together with the position of the image centroid, provides a rotation-variant measurement of the image. Therefore, if one also assumes that the geometric changes induced by the camera's motion can be described by an affine transformation, one will need to represent the image region by a rotationally invariant descriptor to achieve a viewpoint invariant description. This descriptor will be presented in Section III.

Finally, the image region is normalized photometrically. 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 colour channels. Then, the values of individual colour channels are transformed to have zero mean and unit variance, allowing a patch to be represented invariantly to photometric changes.

III. PCA KERNEL-BASED DESCRIPTOR

Colour histograms have been traditionally employed to provide an efficient image region descriptor, encoding the inner colour distribution of the corresponding set of pixels. Besides, colour histograms can be easily quantized into a small number of bins to satisfy the low-computational cost imposed by real-time processing. On the contrary, colour histograms do not take into account the spatial information. To avoid this problem, the regions can be masked with a kernel in the spatial domain [4].

Specifically, in our implementation, the CIELab colour space has been chosen at the hierarchical grouping algorithm and then also to characterize the colour of the salient regions. Histogram have been quantized in 16 bins, resulting in a descriptor of $16 \times 16 \times 16$ scalar values (4096 values). The descriptor length is then significantly larger than the one of other distribution-based descriptors like SIFT (128 values). This implies more computational time and storage resources. In order to reduce it, we have applied PCA to the kernel-based histograms.

PCA is an approach for dimensionality reduction that determines the directions along which the variability of the data is maximal. For instance, this technique has been applied by Ke and Sukthankar [9] to the normalized gradient patches provided by the SIFT detector or by Mikolajczyk and Schmid [17] to obtain the final GLOH descriptor. PCA is conducted by extracting the eigenvectors of the total scatter matrix of the database S_T, defined as

$$S_T = \sum_{i=1}^{N} (G_i - \bar{G})(G_i - \bar{G})^T \tag{2}$$

where $\bar{G}$ is the mean value of the database of N descriptors G_i.

Eigenvectors W_i and associated eigenvalues λ_i are calculated by solving

$$S_T W_i = \lambda_i W_i \qquad \forall i \in \{1, ..., d\} \tag{3}$$

The transformation matrix is then defined as $\mathcal{W} = \{W_1, W_2, ...W_K\}$, where K is the minimal number of eigen-

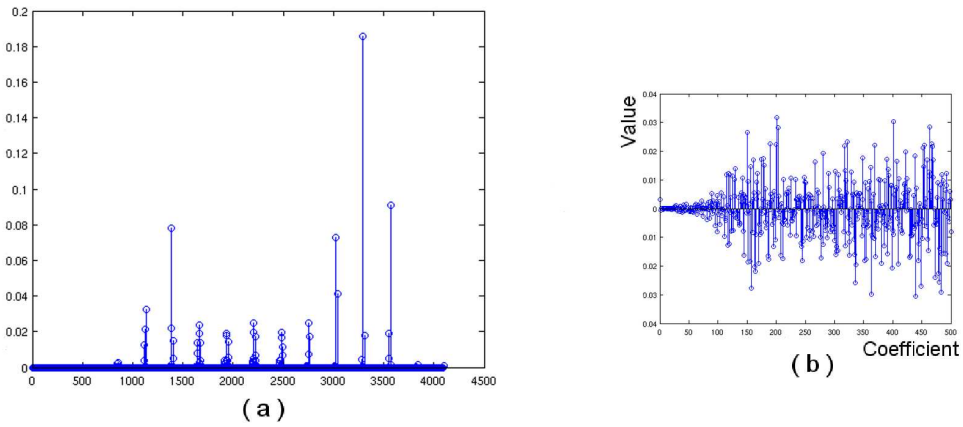

Fig. 4. a) Kernel-based descriptor; and b) PCA projections of the kernel-based descriptor in a).

vectors used to obtain a satisfying representation of the data. Thus, $\mathcal{W}$ is an orthogonal transformation that diagonalises the covariance matrix S_T. In our case, a set of 3100 training samples was used to extract the set of eigenvectors. Then, the compressed feature vector associated to a kernel-based descriptor will be obtained by projecting it onto this set. This projection onto a latent space not only reduces dimensionality but also decorrelates the data. When considering a small database of high dimensionality, this decorrelation can be useful for further encodings, due to the sparsity of data in high-dimensional space.

Fig. 4b shows the feature vectors associated to the kernel-based descriptors depicted in Fig. 4a. In this case, a set of 500 eigenvectors has been chosen to ensure that the projection of the data onto this reduced set covers at least 90% of the data's spread, $\sum_{i=1}^{K} \lambda_i / \sum_i \lambda_i > 0.9$.

Finally, it must be noted that the Euclidean distance between two descriptors can be used to determine whether the two salient regions correspond to the same patch in different images.

IV. Experimental results

A. Evaluation of the proposed segmentation algorithm

The proposed segmentation algorithm has been quantitatively evaluated and compared with other similar algorithms. Three empirical measures have been employed: the Shift-Variance (SV) and the F and Q functions [14]. Shift variance means that the image segmentation produced by pyramid-based algorithms varies when the base of the pyramid is shifted slightly. This is an undesirable effect, so that the SV can be taken as a measure of the quality of a segmentation algorithm. The F and Q functions are measures of the uniformity or homogeneity within the segmented regions together with simplicity in the sense of a relative lack of small holes in the segmentation. Finally, these functions also take into consideration that adjacent regions must present significantly different values of their uniform characteristics.

The SV method compares the segmentation results provided by a given algorithm for slightly shifted versions of the same image. In our case, a window of 128 $\times$ 128 pixels in the centre of the original image has been taken. The segmentation of this subimage is compared with each segmented image obtained by shifting the window a maximum of 11 pixels to the right and 11 pixels down. Thus, there are a total of 120 images to compare with the original one. In order to perfom each comparison between a segmented shifted image I_i and the segmented original image I_{or}, the root mean square colour difference ($\mathrm{RMSD}_{I_{or},I_i}$) has been employed [13]. Then, the SV is expressed as

$$\mathrm{SV} = \frac{1}{120} \sum_{j=1}^{120} \mathrm{RMSD}_{I_{or},I_j} \tag{4}$$

The smaller the value of this parameter, the better is the segmentation results.

On the other hand, the F function is computed as

$$\mathrm{F}(I) = \frac{1}{1000(N \cdot M)} \sqrt{R} \sum_{i=1}^{R} \frac{e_i^2}{A_i} \tag{5}$$

with I being the segmented image, $N \cdot M$ the image size, R the number of segmented regions, and A_i and e_i the area of region i and its average colour error, respectively. The Q function is defined by

$$\mathrm{Q}(I) = \frac{1}{1000(N \cdot M)\sqrt{R} \sum_{i=1}^{R} \left[\frac{e_i^2}{1+logA_i} + (\frac{R(A_i)}{A_i})^2 \right]} \tag{6}$$

with $R(A_i)$ being the number of segmented regions with area equal to A_i. This measure penalizes more rigidly the existence of small regions.

For comparison purposes, five irregular pyramids are employed: the BIP [14], the localized pyramid (LP) [7], the segmentation approach proposed by Lallich et al [10] (MP), the hierarchy of image partitions (HP) [6], and the combinatorial pyramid (CoP) [2]. Four features have been evaluated: the F and Q functions, the SV measure, and the execution time. The images used are a set of 50 images from the Waterloo and Coil 100 databases [13]. The algorithms were run on a 3GHz Pentium IV PC. Table I presents the quantitative results. One appreciates that the shift variance value of the

Fig. 5. a) Reference image and detected features;b)–h) images matched against the reference image. The colour of the ellipses determines if the associated region has been matched to a reference feature (displayed in yellow) or not (displayed in blue).

proposed approach is significantly reduced, providing better results than the rest of approaches. The F and Q values are also improved with respect to the values provided by the original BIP, although they are greater than the ones provided by the HP, the MP and the CoP. Finally, aathough the computation time is slightly greater than with the original BIP, it is still at least ten times less than in the rest of the irregular pyramids.

TABLE I
QUANTITATIVE SEGMENTATION RESULTS: HIERARCHY HEIGHT, NUMBER OF REGIONS OBTAINED, F, Q, AND SHIFT VARIANCE (SV) VALUES, AND EXECUTION TIME.

	F	Q	SV	Time (sec)
LP	743.2	1011.5	30.2	2.75
MP	650.1	818.5	29.3	3.42
HP	670.3	955.1	28.4	4.23
CoP	630.7	870.2	30.5	2.85
BIP	720.2	1090.1	44.1	0.20
Proposed	700.1	950.3	24.3	0.23

B. Testing the approach in an environment mapping framework

The proposed approach has been tested on an ActiveMedia Pioneer 2AT robot. Among other sensors, this robot is mounted with a STH-MDCS stereoscopic camera from Videre Design. This is 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 robot at a constant orientation, looking forward. The robot was driven through different environments while capturing real-life stereo images. Images were restricted to 640×480 or 320×240 pixels.

The viewpoint invariance of our approach has been also qualitatively tested. Images of an scene starting from head on (reference pose) and gradually increasing the viewing angle

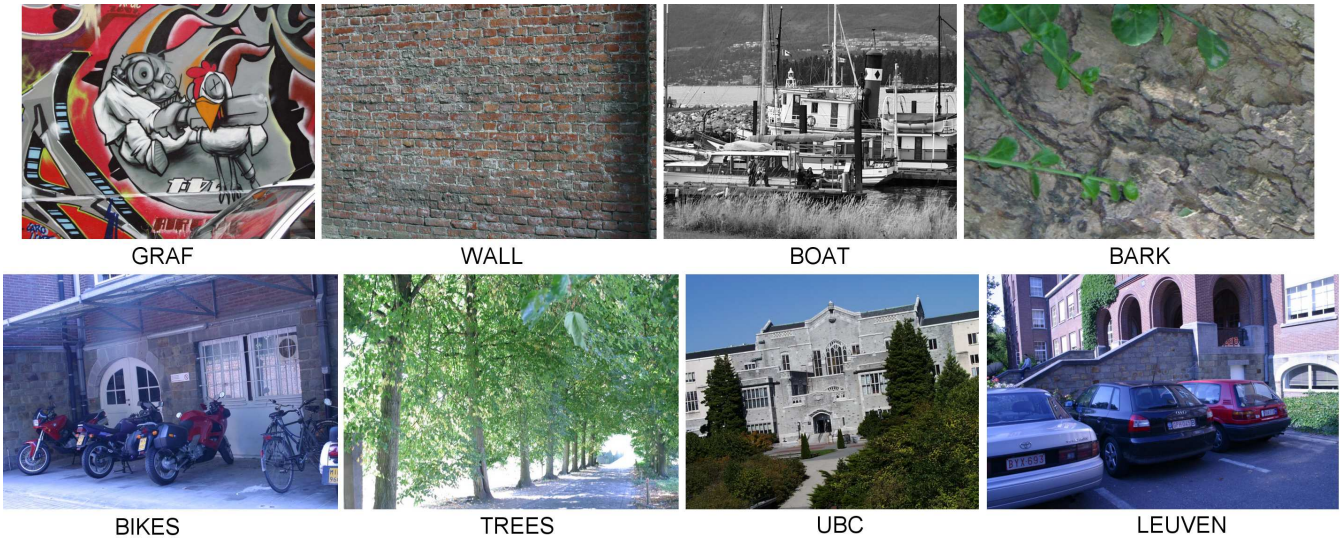

Fig. 6. Image examples of the eight sets used for comparison purposes.

and/or the distance to the reference pose have been captured. Fig. 5 shows one of these experiments, where each visual feature is represented by an ellipse. For each image, visual features are extracted and matched to the features found in the zero degrees reference image (Fig. 5a). A nearest neighbour-based matching strategy has been used, i.e. two regions **A** and **B** are matched if the descriptor D_B is the nearest neighbour to D_A and if the distance between them is below a threshold U. With this approach, a descriptor has only one match. The colour of the ellipses represented in Figs. 5b-h determines if the associated region has been matched to a reference feature (displayed in yellow) or not (displayed in blue). Fig. 11 shows some frames of an experiment conducted in an indoor environment. In this test, detected regions were matched during the trajectory using the same nearest network algorithm. As can be seen, corresponding regions are matched when the same scene is observed from different viewpoints conditions.

Experimental results show that the system can deal with changes in viewpoint up to 50 or 60 degrees and with scale changes of 2 to 2.5. It can be also noted that the number of matches found slightly decreases with increasing scale change.

V. A COMPARATIVE STUDY

Our approach has been also compared with other similar methods employed the protocol described by Mikolajczyk et al. [18]. Images, Matlab code to carry out the performance tests, and binaries of the approaches have been downloaded from http://www.robots.ox.ac.uk/∼vgg/research/affine. Specifically, 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). These image sets can be grouped into two different scene types: one scene type contains homogeneous regions which present distinctive boundaries (structured scenes), meanwhile the other type 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.

To evaluate the described detector, we use the repeatability score as described by Mikolajczyk et al. [18]. This indicates how many of the detected affine regions are found in both images, relative to the lowest total number of regions found (where only the part of the image that is visible in both images is taken into account). It must be noted that 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) [12], the Hessian-affine detector [16], the *maximally stable extremal region* detector (MSER) [15], the *intensity extrema-based region* detector (IBR) [19] and the Fast-Hessian [1]. For all experiments, the default parameters given by the authors are used for each detector. From Table II, 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

TABLE II
NUMBER OF DETECTED REGIONS AND COMPUTATION TIMES FOR DIFFERENT DETECTORS FOR GRAF IMAGE (SEE FIG. 6).

detector	Number of regions	Run time (sec)
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

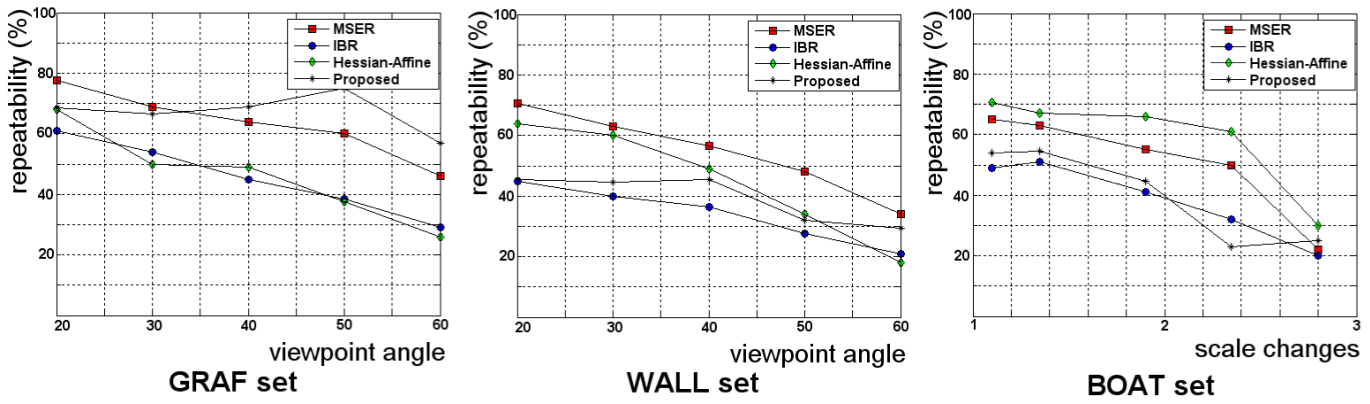

Fig. 7. Repeatability scores for GRAF, WALL and BOAT sequences (see Fig. 6).

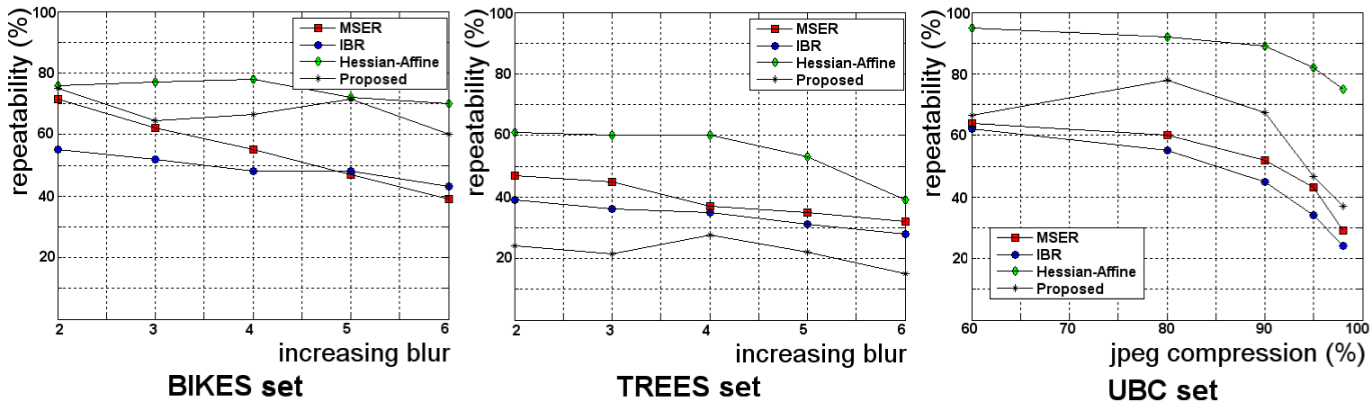

Fig. 8. Repeatability scores for BIKES, TREES and UBC sequences (see Fig. 6).

scenes (e.g. the proposed approach and the MSER) and others to more textured scenes (e.g. Hessian-affine). Table II shows that computation times are also very different. They have been measured on a Pentium 4.2GHz Linux PC, for the GRAF image, which is 800×640 pixels.

The repeatability for six sets of images are illustrated in Figs. 7 and 8. 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 PCA 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 [17]. Fig. 9 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 [17] for further details). Compared descriptors are the SIFT [11], colour SIFT [3] and GLOH [17]. From the results, it can be noted that the PCA 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.

Fig. 10 shows the relationship between the matching accuracy of the proposed descriptor and the dimensionality of the feature space. As expected, increasing the dimensionality of the feature vector results in better accuracy. However, when this dimension exceeds a certain size, the matching accuracy of the algorithm remains approximately constant.

VI. Conclusions

This paper describes an affine region detector whose performance is similar to the current state-of-the art, both in speed and accuracy. To obtain these regions, a hierarchical grouping approach has been performed, generating from the input image an irregular pyramid. 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. In the obtained hierarchy, the receptive field of a pyramid node is considered as a salient regions if this pyramid node is high-contrasted with respect to its neighbours. This detection is conducted over the different pyramid levels, allowing to detect salient

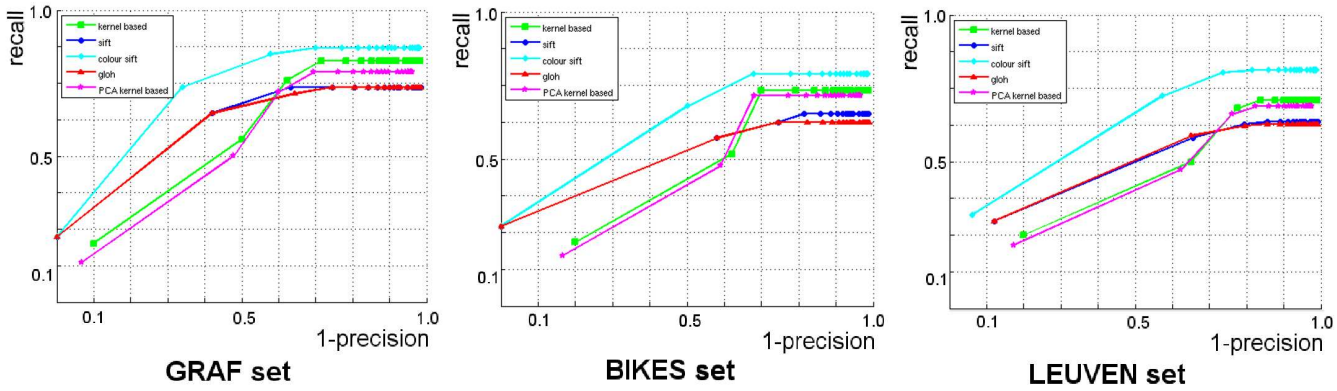

Fig. 9. Recall vs. 1-precision curves for GRAF, BIKES and LEUVEN sequences (see Fig. 6).

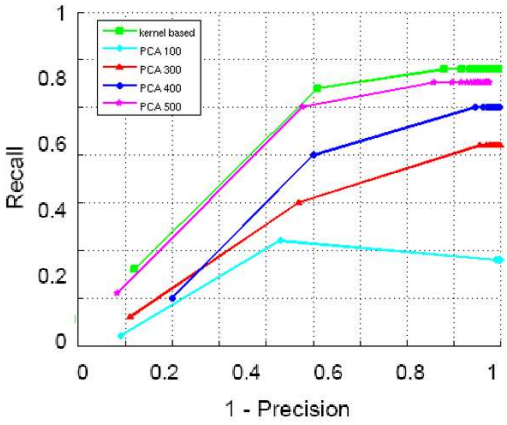

Fig. 10. Performance of the proposed descriptor as PCA dimension varies.

regions at different scales. On the other hand, salient regions have been characterized by a kernel-based descriptor. In order to reduce the large size of this descriptor, we have applied PCA to the kernel-based histograms. The performance of the proposed descriptor is comparable to other similar approaches.

ACKNOWLEDGMENT

This work has been partially granted by the Spanish Ministerio de Ciencia e Inovación (MCINN) and FEDER funds, and by Junta de Andalucía, under projects no. TIN2008-06196 and P07-TIC-03106, respectively.

REFERENCES

[1] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In *In: Proc. European Conf. on Computer Vision*, pages 407–417, 2006.

[2] L. Brun and W. Kropatsch. Construction of combinatorial pyramids. In *In: Proc. of Graph-based Representation in Pattern Recognition*, pages 1–12, 2003.

[3] G. J. Burghouts and J. M. Geusebroek. Performance evaluation of local colour invariants. *Comput. Vision and Image Understanding*, 113.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. on Pattern Anal. and Machine Intell.*, 25(5):564–577, 2003.

[5] Y. Haxhimusa, R. Glantz, M. Saib, G. Langs, and W. Kropatsch. Logarithmic tapering graph pyramid. In *Proc. 24th German Association for Pattern Recognition Symposium*, pages 117–124, 2002.

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

[7] J. Huart and P. Bertolino. Similarity-based and perception-based image segmentation. In *In: Proc. IEEE Int. Conf. on Image Processing 3*, pages 1148–1151, 2005.

[8] J. Jolion. Stochastic pyramid revisited. *Pattern Recognition Letters*, 24(8):1035–1042, 2003.

[9] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *In: Proc. Computer Vision Pattern Recogn. 2*, pages 560–513, 2004.

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

[11] D. Lowe. Object recognition from local scale invariant features. In *In: Proc. 7th Int. Conf. on Computer Vision*, pages 1150–1157, 1999.

[12] D. Lowe. Distinctive image features from scale-invariant keypoints, cascade filtering approach. *Int. Journal Computer Vision*, 60:91–110, 2004.

[13] R. Marfil, L. Molina-Tanco, A. Bandera, J. Rodríguez, and F. Sandoval. Pyramid segmentation algorithms revisited. *Pattern Recognition*, 39(8):1430–1451, 2006.

[14] R. Marfil, L. Molina-Tanco, A. Bandera, and F. Sandoval. The construction of bounded irregular pyramids with a union-find decimation process. *Lecture Notes on Computer Science*, 4538:307–318, 2007.

[15] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. In *In: Proc. British Machine Vision Conference*, pages 384–393, 2002.

[16] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *In: Proc. 7th European Conference on Computer Vision*, pages 128–142, 2002.

[17] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis Machine Intell.*, 27(10):1615–1630, 2005.

[18] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *Int. Journal Computer Vision*, 65:43–72, 2006.

[19] T. Tuytelaars and L. V. Gool. Matching widely separated views based on affine invariant regions. *Int. Journal on Computer Vision*, 59(1):61–85, 2004.

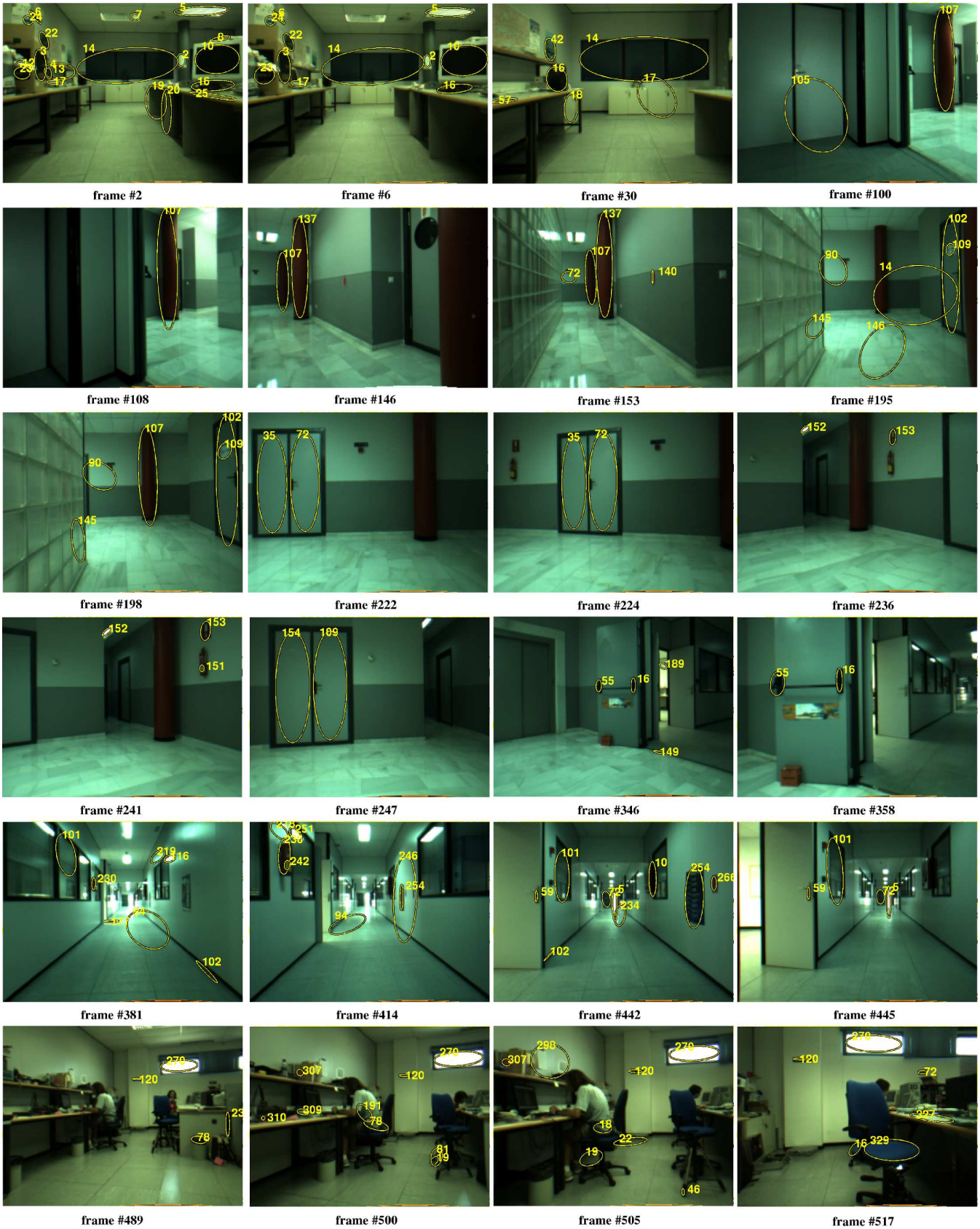

Fig. 11. Experiment in an indoor environment. Detected regions were matched during the trajectory using a nearest neighbour algorithm. It can be seen how corresponding regions are matched when the same scene is observed, e.g. in the following frame sets: (#2, #6, #30), (#100, #108), (#146, #153, #198), (#222, #224), (#236, #241), (#346, #358), (#414, #442, #445) and (#489, #500, #505, #517)