Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 451389, 11 pages
doi:10.1155/2008/451389

Research Article

Detection of Complex Salient Regions

Sergio Escalera,[1,2] Oriol Pujol,[1,2] and Petia Radeva[1,2]

[1] *Computer Vision Center, Campus UAB, Edifici O, 08193 Bellaterra, Barcelona, Spain*
[2] *Departamento de Matemàtica Aplicada i Anàlisi, Universitat de Barcelona (UB), 08007 Barcelona, Spain*

Correspondence should be addressed to Sergio Escalera, sescalera@cvc.uab.es

Received 16 October 2007; Revised 8 February 2008; Accepted 12 March 2008

Recommended by Irene Gu

The goal of interest point detectors is to find, in an unsupervised way, keypoints easy to extract and at the same time robust to image transformations. We present a novel set of saliency features based on image singularities that takes into account the region content in terms of intensity and local structure. The region complexity is estimated by means of the entropy of the gray-level information; shape information is obtained by measuring the entropy of significant orientations. The regions are located in their representative scale and categorized by their complexity level. Thus, the regions are highly discriminable and less sensitive to confusion and false alarm than the traditional approaches. We compare the novel complex salient regions with the state-of-the-art keypoint detectors. The presented interest points show robustness to a wide set of image transformations and high repeatability as well as allow matching from different camera points of view. Besides, we show the temporal robustness of the novel salient regions in real video sequences, being potentially useful for matching, image retrieval, and object categorization problems.

1. INTRODUCTION

Visual saliency [1] is a broad term that refers to the idea that certain parts of a scene are preattentively distinctive and create some form of immediate significant visual arousal within the early stages of the human vision system. The term "salient feature" has previously been used by many other researchers [2–5]. Although definitions vary, intuitively, saliency corresponds to the "rarity" of a feature [6]. In the framework of keypoint detectors, special attention has been paid to biologically inspired landmarks. One of the main models for early vision in humans, attributed to Neisser [7], is that consisting of preattentive and attentive stages. In the preattentive stage, only "pop-out" features are detected. These are the salient local regions of the image which present some form of discontinuity. In the attentive stages, relationships between these features are found, and grouping takes place in order to model object classes.

Interest point detectors have been used in multiple applications: matching for stereo pairs [8–11], image retrieval from large databases [12, 13], object retrieval in video [14, 15], shot location [16], and object categorization [17–20] to mention just a few. One of the most well-known keypoint detector is the Harris detector [21]. The method is based on searching for edges that are maintained at different scales to detect interest image points. Several variants and applications based on the Harris point detector have been used in the literature such as Harris-Laplacian [22], Affine variants [21], Lowe [23], and so forth. In [9], the authors proposed a novel region detector based on the homogeneity of the parts of the image. Moreover, the definition of the detected regions makes the description of the parts ambiguous when considered in object recognition frameworks. Schmid and Mohr [12] proposed the use of corners as interest points in image retrieval. They compared different corner detectors and showed that the best results were provided by the Harris corner detector [22]. In [24], a method for introducing the cornerness of the Harris detector in the method of [1] is proposed. However, the robustness of the method is directly dependent on the cornerness performance. Kadir and Brady [1] estimate the entropy of the gray levels of a region to measure its magnitude and scale of saliency. The detected regions are shown to be highly discriminable, avoiding the exponential temporal cost of analyzing dictionaries when used in object recognition models, as in [25]. However, using the gray level information, one can obtain regions with different complexity and with the same entropy values. Recently, the authors of [26] proposed the oriented-based SIFT descriptor such as a stability criterion to obtain stable scales for multiscale Harris and Laplacian points, with great success.

In this paper, we propose a model that allows to detect the most relevant image features based on their complexity. We use the entropy measure based on the color or gray level information and shape complexity (defined by means of a novel normalized pseudohistogram of orientations) to categorize the saliency levels. Including simple complexity constraints (the null-orientation concept and the adaptive threshold of orientations), the novel set of features is highly invariant to a great variety of image transformations and leads to a better repeatability and lower false alarm rate than the state-of-the-art keypoint detectors.

The paper is organized as follows. Section 2 explains our complex salient regions. In Section 3, we perform a set of experiments comparing the state-of-the-art region detectors. The validation is done over public image databases [27] and video sequences [28, 29] in order to test the repeatability, false alarm rate, and matching score of the detectors. Finally, Section 4 concludes the paper.

2. CSR: COMPLEX SALIENT REGIONS

In [1], Kadir and Brady introduce the gray-level saliency regions. The key principle behind their approach is that salient image regions exhibit unpredictability in their local attributes and over spatial scale. This section is divided in two parts. Firstly, we describe the background formulation, inspired by [1]. Secondly, we introduce the new metrics to estimate the saliency complexity.

2.1. Detection of salient regions

The approach to detect the position and scale of the salient regions uses a saliency estimation defined by the Shannon entropy at different scales at a given point. In this way, we obtain the entropy as a function in the space of scales. We consider significant saliency regions those that correspond to the maxima of this function, where the maximal entropy value is used to estimate the complex salient magnitude. Now, we define the notation and description of the stages of the process.

Let H be the entropy of a given region, S_p the space of significant scales, and W the relevance factor (weight). In the continuous case, the saliency measure γ is defined as a function of scale s and position x as follows:

$$\gamma(S_p, x) = W^T(S_p, x)H(S_p, x) \tag{1}$$

for each point x and the set of scales at which entropy peaks are obtained (S_p). Then, the saliency is determined by weighting the entropy at those scales by W. The entropy $H(s_i, x)$, where $s_i \in S_p$, is defined as $H(s,x) = -\int p(I,s,x)\log_2 p(I,s,x)dI$, where $p(I,s,x)$ is the probability density function of the intensity I as a function of scale s and position x. In the discrete case, for a region R_x of n pixels, the Shannon entropy is defined as follows:

$$H(R_x) = -\sum_{i=1}^{n} P_{R_x}(i)\log_2 P_{R_x}(i), \tag{2}$$

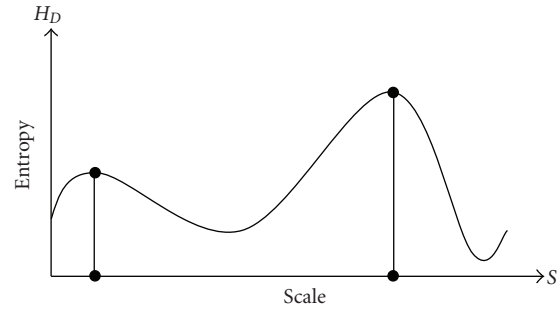


Figure 1: Local maxima of function H_D in the scale space S.

where $P_{R_x}(d_i)$ is the probability of taking the value d_i in the local region R_x. The set of scales S_p is defined by the maxima of the function H in the space of scales $S_p = \{s : \partial H(s,x)/\partial s = 0,\ \partial^2 H(s,x)/\partial s^2 < 0\}$.

The entropy as a function of the scale space S is shown in Figure 1. In the figure, a point x is evaluated in the space of scales, obtaining two local maxima. These peaks of the entropy estimation correspond to the representative scales for the analyzed image point.

The relevance of each position of the saliency at its representative scales is defined by the interscale saliency measure $W(s,x) = s(\partial/\partial s)H(s,x)$.

Considering each scale $s \in S$ that are local maxima ($s \in S_p$) and pixel x, we estimate W in the discrete case as a function of the change in magnitude of the entropy over the scales:

$$W(s,x) = s\frac{|H(s-1,x) - H(s,x)| + |H(s+1,x) - H(s,x)|}{2}. \tag{3}$$

Using the previous weighting factor, we assume that the significant salient regions correspond to that locations with high distortion in terms of the Shannon entropy and its peak magnitude.

2.2. Traditional gray level and orientation saliency

Kadir and Brady [1] used the gray-level entropy to define the saliency complexity of a given region. However, this approach fails in cases where the regions have different complexities. In Figure 2, one can observe different regions with the same amount of pixels for each gray level and different visual complexity. Note that the approach based on the gray-level entropy proposed by [1] gives the same entropy value, thus the same "rarity" level for all of them.

A natural and well-founded measure to solve this pathology is the use of complementary orientation information. In the same work [1], Kadir and Brady show preliminary results applying the orientation information in fingerprint images. However, the use of orientations as a measure of complexity involves several problems. In order to exemplify those problems, suppose that we have the regions (a) and (b) of Figure 3. Both regions have the same pdf (Figure 3(c)), but they contain different number of significant orientations (histograms of Figures 3(d) and

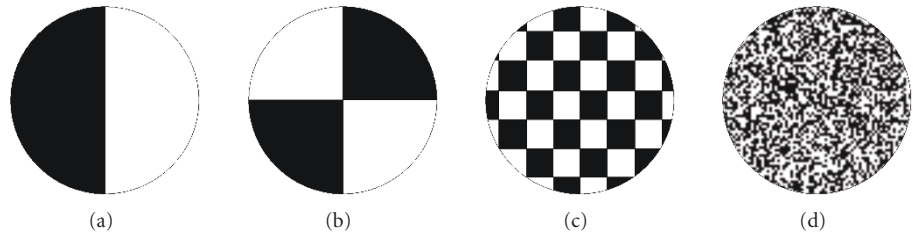


Figure 2: Regions of different complexity with the same gray-level entropy.

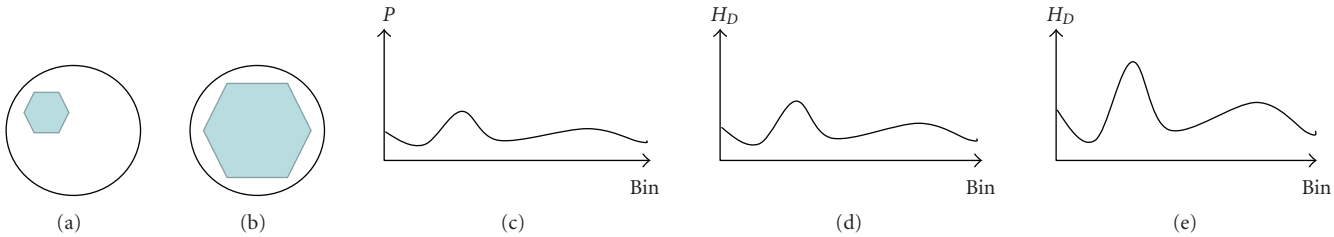


Figure 3: (a), (b) Two circular regions with the same content at different resolutions. (c) Same pdf for the regions (a) and (b). (d) Orientations histogram for (a), and (e) orientations histogram for (b).

3(e)). In a regular histogram, low magnitude gradient is mostly due to noise, and it is distributed uniformly over all bins. However, the pdf obtained in those cases remains the same because of the histogram normalization. We take into account these issues and we incorporate a novel orientations normalization procedure that evaluate properly the complexity level of each image region.

2.3. Normalized orientation entropy measure

The normalized orientation entropy measure is based on computing the entropy using a pseudohistogram of orientations. The usual way to estimate the histogram of orientations of a region is to use a range from 0 to 2π radians. Considering orientation independent from gradient magnitude hide the danger to mix signal with noise (usually, corresponding to low gradient magnitudes). In the limit case, when the gradient is zero, we have a singularity of the orientation function. On the other hand, these pixels normally correspond to homogeneous regions that can be useful to describe parts of the objects. To overcome this problem, we propose to introduce an additional bin that corresponds to the pixels with undetermined orientation that is called *null-orientation bin*. In this case, signal is not mixed with noise and at the same time, homogeneous regions are taken into account. Our proposed orientation metric consists of computing the saliency including the *null-orientations* in the modified orientation pdf.

First of all, we compute the relevant gradient magnitudes of an image to obtain the significant orientations. Instead of using an experimental threshold, we propose an adaptive orientation threshold for each particular image. For a given image, our method computes and normalized the gradient module $|\nabla(I)|$ in the range $[0 \cdots 1]$. Then, we estimate its histogram, and the Otsu method [30] is applied to obtain the adaptive threshold for orientations. The significant orientation locations obtained for two image samples are shown in Figure 4.

Considering the $k \leq K$ most significant orientations using the adaptive threshold, where K is the total number of locations in a given region, we compute the orientations histogram h_O for n orientation bins. In this case, the number of *null-orientation* locations is fixed to $K - k$, and they are added to the histogram h_O as $h_O(n+1) = K - k$.

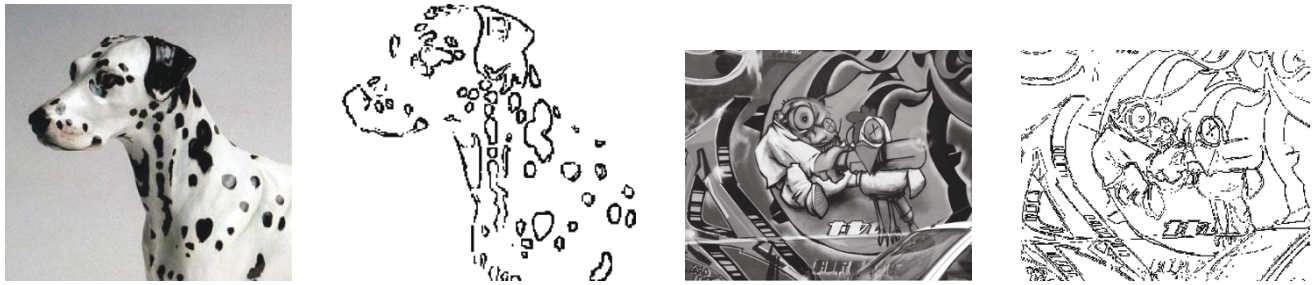
The position $n + 1$ of the histogram h_O is the *null-orientation* bin, and the modified pdf is obtained by means of

$$\mathrm{PDF}_O(i) = \frac{h_O(i)}{\sum_{j=1}^{n+1} h_O(j)}, \quad \forall i \in [1, \ldots, n]. \tag{4}$$

Finally, the pdf PDF_O is used to estimate the orientation entropy value of a given region. Note that the *null-orientation* bin $n + 1$ is not included in the entropy evaluation, since its goal is to normalize the first n bins according to the patch complexity (Observe that the entropy measure of the *null-orientation* bin usually makes the first n bins insignificant.)

2.4. Combining the saliency

In our particular case, the gray-level histogram is combined with the pseudohistogram of orientations. We experimentally tested that the performance of both information offers better performance that only uses the orientations or the gray-level entropy criterion. In this way, once estimated the two corresponding pdf, we apply (1), (2), and (3) to each one in the same way. The final measure is obtained by means of the simple addition $\gamma = \gamma_G + \gamma_O$, where γ_G and γ_O are estimated by (1) for the gray and orientation saliency, and γ is the result, which contains the final significant saliency positions, magnitudes (level of complexity), and scales. Other strategies, such as the product and logarithmic combinations of gray-level and orientation complexities,



(a) (b) (c) (d)

Figure 4: Relevant orientations estimation.

(a) (b) (c)

Figure 5: (a) First maximal complexity region for gray-level entropy, (b) orientations entropy, (c) and combined entropy.

have also been tested to detect salient regions. However, the results were not satisfactory since these combinations were made to discard salient regions if one of the two saliency values is too small, independently of the dominance of the other component. This effect is unsatisfactory since the dominance of one component over the other may produce enough visual complexity to be considered as a salient region. On the other hand, a simple addition showed to maintain the salient regions in the cases, where one of the two measures is predominant enough. At the same time, it also allows to consider regions, where both saliency values introduce moderate complexity. The effect of the combined saliency measure is shown in the toy problem of Figure 5. Figure 5 has three representative objects of different complexities. We applied the gray-level entropy, the orientation entropy, and the combined saliency using simple addition. One can observe that the combined saliency measure selects the region with higher visual complexity (Figure 5(c)).

This new saliency measure gives a high complexity value, when the region contains different gray levels information (nonhomogeneous region) and the shape complexity is high (high number of gradient magnitudes at multiple orientations). The complexity to estimate the regions saliency is $O(nl)$, where n is the number of image pixels, and l is the number of scales searched for each pixel. The complexity of the second step is $O(e)$, where e is the number of extrema detected at the previous step. Note that an exhaustive search is not always required, and not all pixels and possible scales have to be estimated. However, the exhaustive search is relatively fast to compute (less than 1 second in an 800 $\times$ 640 medium resolution image).

An example of CSR responses for an image sample under different transformations is shown in Figure 6. Rotation, white noise addition, and affine distortion transformations are shown. Observe that the CSR regions are maintained in the set of transformations.

The mean number of detected regions and the mean average region size for the traditional gray-level saliency and the novel salient criterion using the Caltech database samples [27] of Figure 7 are shown in Figure 8. All images are of medium resolution (approximately 600 $\times$ 600 pixels). The size of the regions corresponds to the radius of the detected circular regions in 20 bins between radius of length 5 and 100 pixels. Note that the number of detected regions considerably increase using the new metric, in particular it is about three times more. At the same time, the preferred regions for the novel salient regions are of intermediate complexity sizes, which typically implies a higher discriminable power [31].

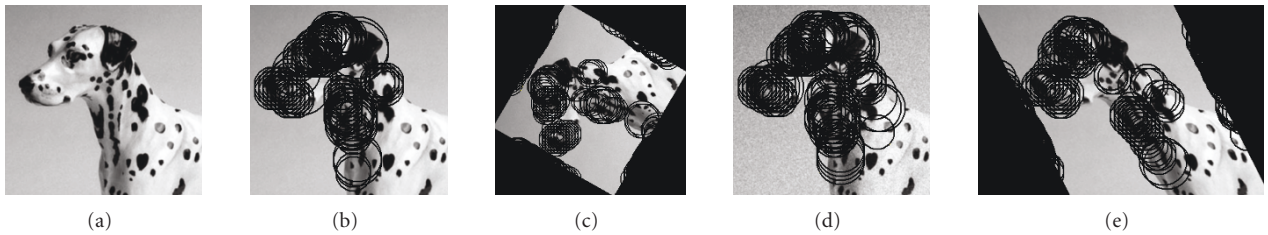
As our orientations strategy normalize the input image it offers invariance to scalar changes in image contrast. The use of gradients is also robust to an additive contrast change in brightness, which makes the technique relatively insensitive to illumination changes. Invariance to scale is obtained by the scale search of local maximums, and the use of circular regions takes into account the global complexity of the inner of the regions, which also makes the strategy invariant to rotation.

3. RESULTS

To validate the presented methodology, we should determine data, measurements for the experiments, state-of-the-art methods to compare, and applications.

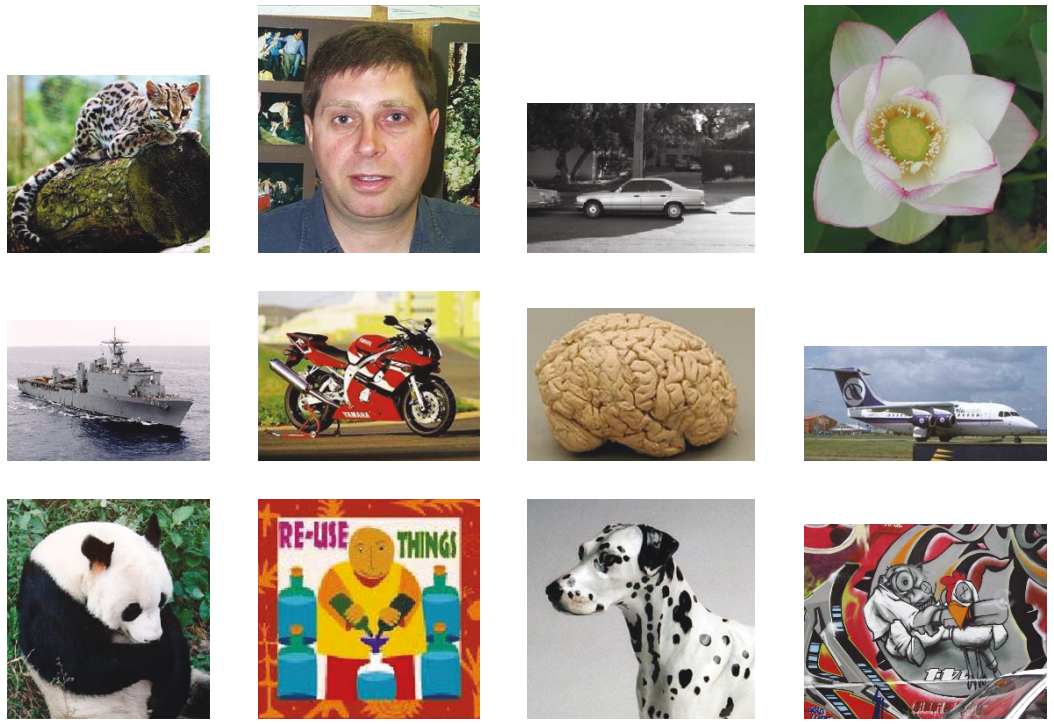
(a) *Data*. Images are obtained from the public Caltech repository [27] and the video sequences from [28, 29].

(b) *Measurements*. To analyze the performance of the proposed CSR, we perform a set of experiments to show the robustness to image transformations of the novel regions in terms of repeatability, false alarm rate, and matching score. The repeatability and matching score criteria are based on the evaluation framework of [31]. Besides, we include the false alarm rate measurement.



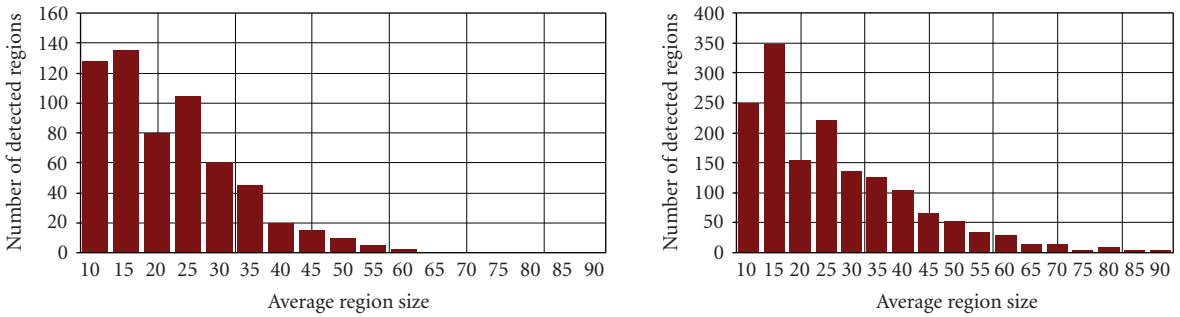
(a) (b) (c) (d) (e)

Figure 6: Image transformation tests for CSR responses: (a) input image, (b) initial CSR region detection, (c) 60 degree rotation, (d) white noise, and (e) affine transformation.



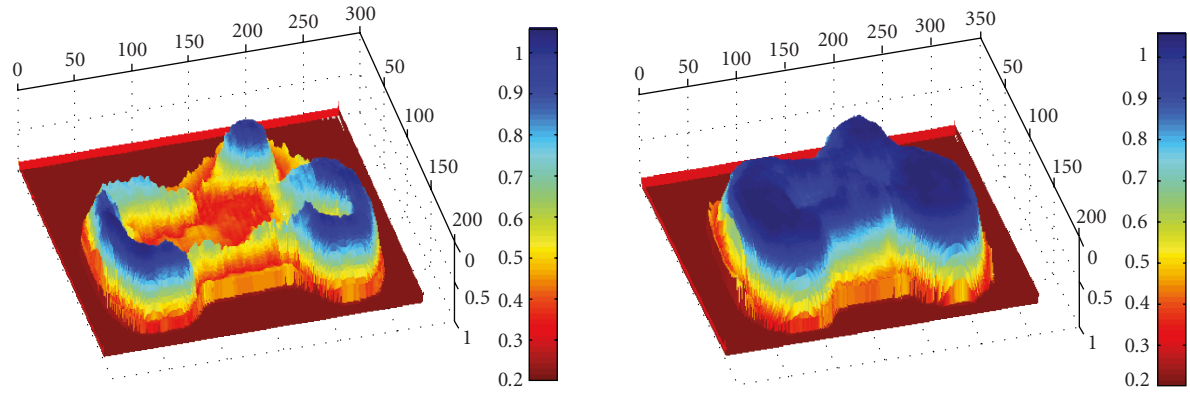
(a) (b) (c) (d)

Figure 7: Caltech database samples.



(a) Grey-level saliency (b) Complex salient regions

Figure 8: Histograms of mean region size and number of detected regions for the samples of Figure 7.



(a) (b)

Figure 9: Mean volume image for the most relevant detected landmarks on the set of Caltech motorbike database for gray saliency (a) and our proposed CSR (b).

(c) *State-of-the-art methods*. We compare the presented CSR with the Harris-Laplacian, Hessian-Laplacian, and the gray-level saliency. The parameters used for the region detectors are the default parameters given by the authors [1, 9, 21]. For the salient criteria of [1] and our CSR, we use 16 bins for the gray-level and orientations histograms. The number of regions obtained by each method strongly depends on the image since each one can contain different type of features.

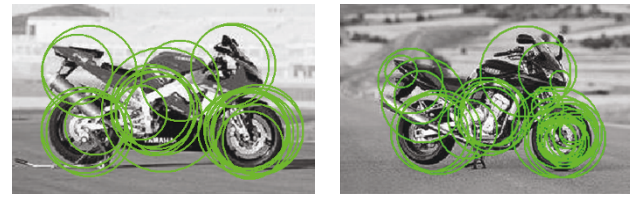
(d) *Applications*. To show the wide applicability of the proposed CSR, we designed a broad set of experiments. First, we compare the performance of the presented CSR with the traditional approach of [1]. Second, we show the robustness to image transformations of the novel regions. Third, we match the detected regions of images taken from different camera points of view. And finally, we apply the technique on video sequences to analyze the temporal behavior by matching the detected regions in different frames.

3.1. Gray-level saliency versus CSR

We selected a set of 250 random motorbike samples from the motorbike Caltech database (the motorbike database was chosen to compare the salient responses of both detectors in a visual distinctive problem, and do not to try to solve a difficult problem) [27] and we estimated the highest saliency responses for each image using the gray-level saliency and the CSR regions. The mean volume image V of detected regions is shown in Figure 9. The volume image V is defined as

$$V = \frac{1}{i}\sum_{i=1}^{N} I_{R_i}, \tag{5}$$

where I_{R_i} is the binary image with value 1 at those positions that fall into the detected circular regions in image I_i, and N is the total number of image samples. One can observe that the CSR responses recover better the motorbike, and the probability to detect each object part is higher. In Figure 10, two examples of detected CSR for the motorbike database are shown.



(a) (b)

Figure 10: Detected CSR from Caltech motorbike images.

3.2. Repeatability and false alarm

In order to validate our results, we selected the samples showed in Figure 7 from the public Caltech repository database [27]. In this set of samples, we applied a set of transformations: rotation (10 degrees per step up to 100), white noise addition (0.1 of the variance per step up to 1.0), scale changes (15% per step up to 150), affine distortions (5 pixels x-axis distortion per step up to 50), and light decreasing (-0.05 per step of β down to -0.5, where the brightness of the new image is raised to the power of γ, where γ is $1/(1+\beta)$).

Over the set of transformations, we apply the evaluation framework of [31] for the repeatability criterion. The repeatability rate measures how well the detector selects the same scene region under various image transformations. As we have a reference image for each sequence of transformations, we know the homographies from each transformed image to the reference image. Then, the accuracy is measured by the amount of overlap between the detected region and the corresponding region projected from the reference image with the known homography. Two regions are matched if they satisfy

$$1 - \frac{R_{\mu_a} \cap R_{H^T \mu_b H}}{R_{\mu_a} \cup R_{H^T \mu_b H}} < \epsilon_O, \tag{6}$$

where R_μ is the circular region obtained by the detector and H is the homography between the two images. We set the maximum overlap error ϵ_O to 40%, as in [31]. Then, the

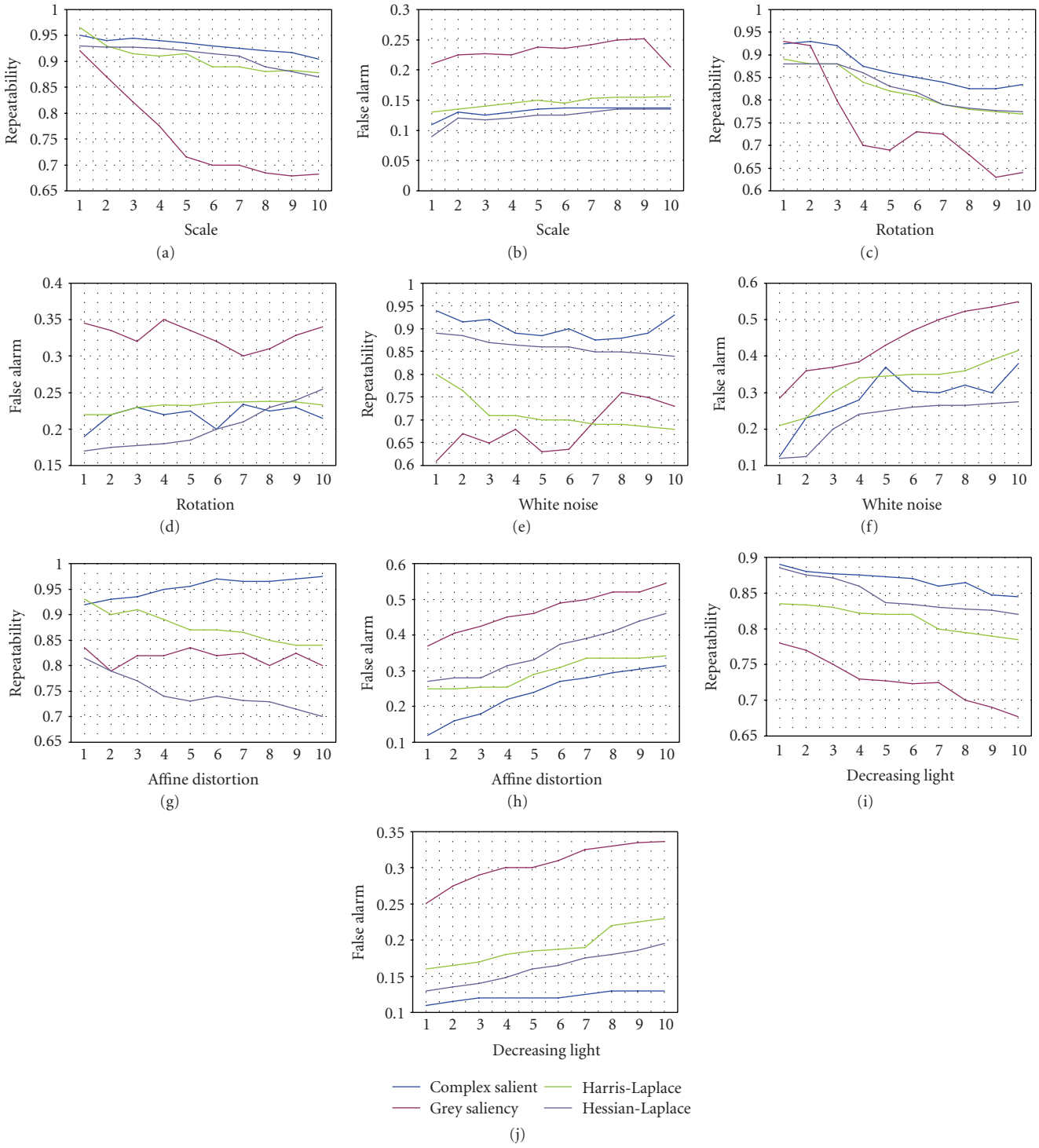


FIGURE 11: Repeatability and false alarm rate in the space of transformations: (a), (b) scale, (c), (d) rotation, (e), (f) white noise, (g), (h) affine invariants, and (i), (j) decreasing light.

repeatability becomes the ratio between the correct matches and the smaller number of detected regions in the two images. Besides, to take into account the amount of regions from the two images that do not produces matches, we introduce the *false alarm* rate criterion, defined as the ratio between the number of regions from the two images that do not match and the total number of regions from the two images. This measure is desirable to be as small as possible.

The mean results for all images checking the repeatability and false alarm ratios for gradually increasing transformations are shown in Figure 11. It is found that white noise addition (Figure 11(e)) and affine transformations

FIGURE 12: (a)–(c) Original images and region detection for (d)–(f) complex salient features, (g)–(i) gray-level entropy, (j)–(l) Harris-Laplacian, and (m)–(o) Hessian-Laplacian for a set of vehicle images from different camera points of view.

(Figure 11(g)) applied to some types of region detectors increase the amount of detected regions. Then, the general behavior in those cases is also the increment of repeatability because of the higher number of overlapping regions. In the same way, this effect also produces a higher increment in their corresponding false alarm curves. Observing the figures, one can see that Harris and Hessian Laplace normally obtain similar results, and Hessian Laplace tends to outperform the Harris Laplace detector. Gray-based salient regions give relatively low repeatability and high false alarm rate, and it is dramatically improved with the CSR regions, which obtain better performance than the rest of detectors in terms of repeatability, obtaining the highest percentage of correspondences for all types of image distortions. For the case of false alarm ratio, the CSR and the Hessian Laplace methods offer the best results, obtaining lower false alarm rate than the Harris Laplace and gray-level salient detectors.

3.3. *Matching under different camera points of view*

In this experiment, we considered different points of view of a camera on the same object. We used a set of 30 real samples from a vehicle. The set of images has been taken with a digital camera of 4 mega pixels from different points of views. Some used samples and the detected regions using the different region detectors are shown in Figure 12.

The matching evaluation is based on the criterion of [31]. A region match is deemed correct, if the overlap error ϵ_O is less than a given threshold. This provides the ground truth for correct matches. Only a single match is allowed for each region. The matching score is computed as the ratio between the number of correct matches and the smaller number of detected regions in the pair of images. Instead of fixing the ϵ_O value, we compute the matching score for a set of ϵ_O values, from 0.65 up to 0.2 decreasing by 0.05.

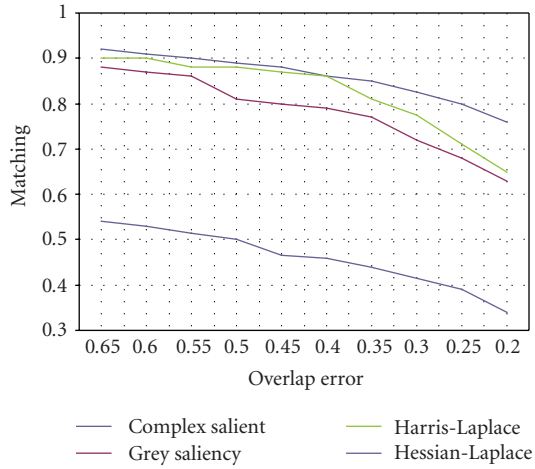


FIGURE 13: Matching percentage of the region detectors for the set of 30 car samples of different points of views in terms of regions intersection percentage.

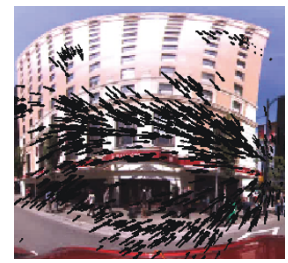
The regions are described using the SIFT descriptor [23] and compared with the Euclidean distance. The overlap value is estimated using a warping technique to align manually the different samples. In Figure 13, the matching score for the region detectors for different ϵ_O thresholds are shown. One can see the low matching percentage of the Hessian-Laplace due to the locality of the detected regions. The gray-level entropy and Hessian-Laplace detectors obtain better matching results. Finally, the CSR regions obtain the highest percentage of matching for all overlap errors values.

3.4. *Temporal robustness*

The next experiment is to apply the CSR regions to video sequences to show their temporal robustness. The temporal robustness of the algorithm is determined by a high score of matching salient features in a sequence of images. This matching is used in order to approximate the optical flow, and thus perform the tracking of the object features. We used the video images from the Ladybug2 spherical digital camera from Point Grey Research group [28]. The car system has six cameras that enable the system to collect video from more than 75% of the full sphere [28]. Furthermore, we also tested the method with road video sequences from the Geovan mobile mapping process from the Institut Cartogràfic de Catalunya [29], that has a stereo pair of calibrated cameras, which are synchronized with a GPS/INS system.

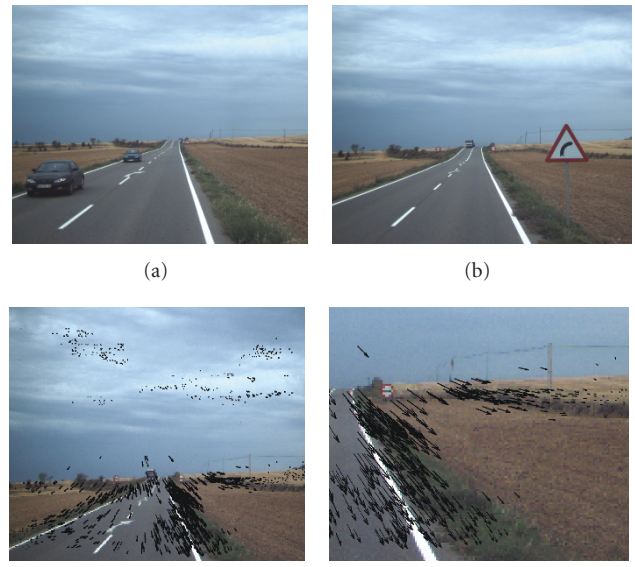
For both experiments we analyzed 100 frames using the SIFT descriptor [23] to describe the regions. The matching is done by similar regions descriptors in terms of the Euclidean distance in a neighborhood two times the diameter of the detected CSRs. The smoothed oriented maps from CSR matchings are shown in Figures 14 and 15. The smoothed oriented maps are obtained by filtering with a gaussian of size 5×5 and $\sigma = 3$ over the map of vectors obtained from the distances of matching each pair of regions. Figure 14(a) shows the oriented map in the first analyzed frame of [28]. Figure 14(b) focuses on the right region of (a). One can see

(a)



(b)

FIGURE 14: (a) Smoothed oriented CSR matches, (b) zoomed right region.



(a) (b) (c) (d)

FIGURE 15: (a), (b) Samples, (c) smoothed oriented CSR matches, and (d) zoomed right region.

that the matched complex regions correspond to singularities in the video sequence and they approximate roughly the video movement. From the road experiment of Figure 15, the oriented map is shown in Figure 15(c). In this video sequence cars and traffic signs appear (Figures 15(a) and 15(b)). The amplified right region is shown in Figure 15(d). One can observe the correct movement trajectory of the road video sequences.

4. CONCLUSIONS

We presented a novel set of salient features, the complex salient regions. These features are based on complex image regions estimated using an entropy measure. The presented CSR analyzes the saliency of the regions using the gray-level and orientations information. We introduced a novel procedure to consider the anisotropic features of image pixels that makes the image orientations useful and highly discriminable in object recognition frameworks. We showed that simply including proper complexity constraints (the null-orientation concept and the adaptive threshold of orientations), the novel set of features is highly invariant to a great variety of image transformations and leads to a better repeatability and lower false alarm rate than the state-of-the-art keypoint detectors. These novel salient regions show robust temporal behavior on real video sequences and can be potentially applied to matching under different camera points of view and image retrieval problems.

We are currently adapting the CSR regions to be invariant to affine transformations [32] and evaluating the methodology to design a multiclass object recognition approach.

ACKNOWLEDGMENT

This work has been supported in part by TIN2006-15308-C02 and FIS ref. PI061290.

REFERENCES

[1] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[2] P. J. Flynn, "Saliencies and symmetries: toward 3D object recognition from large model databases," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '92)*, vol. , pp. 322–327, Champaign, Ill, USA, June 1992.

[3] B. Schiele and J. L. Crowley, "Probabilistic object recognition using multidimensional receptive field histograms," in *Proceedings of the 13th International Conference in Pattern Recognition (ICPR '96)*, vol. 2, pp. 50–54, Vienna, Austria, 1996.

[4] N. Sebe and M. S. Lew, "Salient points for content-based retrieval," in *Proceedings of the 12th British Machine Vision Conference (BMVC '01)*, pp. 401–410, Manchester, UK, September 2001.

[5] K. N. Walker, T. F. Cootes, and C. Taylor, "Locating salient object features," in *Proceedings of the 9th British Machine Vision Conference (BMVC '98)*, pp. 557–566, Southampton, UK, September 1998.

[6] D. Hall, B. Leibe, and B. Schiele, "Saliency of interest points under scale changes," in *Proceedings of the 13th British Machine Vision Conference (BMVC '02)*, Cardiff, UK, September 2002.

[7] U. Neisser, "Visual search," *Scientific American*, vol. 210, no. 6, pp. 94–102, 1964.

[8] A. Baumberg, "Reliable feature matching across widely separated views," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '00)*, vol. 1, pp. 774–781, Hilton Head Island, SC, USA, June 2000.

[9] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," in *Proceedings of the 13th British Machine Vision Conference (BMVC '02)*, vol. 1, pp. 384–393, Cardiff, UK, September 2002.

[10] P. Pritchett and A. Zisserman, "Wide baseline stereo matching," in *Proceedings of the 6th IEEE International Conference on Computer Vision (ICCV '98)*, pp. 754–760, Bombay, India, January 1998.

[11] T. Tuytelaars and L. Gool, "Wide baseline stereo matching based on local, affinely invariant regions," in *Proceedings of the 11th British Machine Vision Conference (BMVC '00*, pp. 412–425, Bristol, UK, September 2000.

[12] C. Schmid and R. Mohr, "Local gray value invariants for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–535, 1997.

[13] T. Tuytelaars and L. Van Gool, "Content-based image retrieval based on local affinely invariant regions," in *Proceedings of the 3rd International Conference on Visual Information and Information Systems (VISUAL '99)*, pp. 493–500, Amsterdam, The Netherlands, June 1999.

[14] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV '03)*, vol. 2, pp. 1470–1477, Nice, France, October 2003.

[15] J. Sivic, F. Schaffalitzky, and A. Zisserman, "Object level grouping for video shots," in *Proceedings of the 8th European Conference on Computer Vision (ECCV '04)*, vol. 3022 of *Lecture Notes in Computer Science*, pp. 85–98, Prague, Czech Republic, May 2004.

[16] F. Schaffalitzky and A. Zisserman, "Automated location matching in movies," *Computer Vision and Image Understanding*, vol. 92, no. 2-3, pp. 236–264, 2003.

[17] G. Csurka, C. Dance, C. Bray, and L. Fan, "Visual categorization with bags of keypoints," in *Proceedings of the International Workshop on Statistical Learning in Computer Vision (ECCV '04)*, pp. 1–22, Prague, Czech Republic, May 2004.

[18] G. Dorko and C. Schmid, "Selection of scale-invariant parts for object class recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '03*, vol. 1, pp. 634–639, Nice, France, October 2003.

[19] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. 2, pp. 264–271, Madison, Wis, USA, June 2003.

[20] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer, "Weak hypotheses and boosting for generic object detection and recognition," in *Proceedings of the 8th European Conference on Computer Vision (ECCV '04)*, vol. 3022 of *Lecture Notes in Computer Science*, pp. 71–84, Prague, Czech Republic, May 2004.

[21] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[22] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the 4th Alvey Vision Conference*, pp. 147–151, Manchester, UK, August-September 1988.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[24] F. Fraundorfer and H. Bischof, "Detecting distinguished regions by saliency," in *Proceedings of the 13th Scandinavian Conference on Image Analysis (SCIA '03)*, vol. 2749 of *Lecture*

Notes in Computer Science, pp. 208–215, Springer, Halmstad, Sweden, June-July 2003.

[25] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio, "A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex," in *Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM '05)*, vol. 36, Monterey, Calif, USA, July 2005.

[26] G. Dorkó and C. Schmid, "Maximally stable local description for scale selection," in *Proceedings of the 9th European Conference on Computer Vision (ECCV '06)*, vol. 3954 of *Lecture Notes in Computer Science*, pp. 504–516, Graz, Austria, May 2006.

[27] http://www.vision.caltech.edu/html-files/archive.html.

[28] http://ptgrey.com/products/ladybug2/samples.asp.

[29] http://www.icc.es/.

[30] N. Otsu, "Threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[31] K. Mikolajczyk, T. Tuytelaars, C. Schmid, et al., "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[32] T. Kadir, A. Zisserman, and M. Brady, "An affine invariant salient region detector," in *Proceedings of the 8th European Conference on Computer Vision (ECCV '04)*, vol. 3021 of *Lecture Notes in Computer Science*, pp. 228–241, Prague, Czech Republic, May 2004.