Complex Salient Regions for Computer Vision Problems

Sergio Escalera, Petia Radeva
Computer Vision Center, UAB
Edifici O, 08193 Bellaterra, Spain
{sescalera,petia}@cvc.uab.es

Oriol Pujol
Dept. Matemàtica Aplicada, UB, CVC
Gran Via 585, 08007, Barcelona, Spain
oriol@cvc.uab.es

Abstract

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 grey-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 allows 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 [9] is a broad term that refers to the idea that certain parts of a scene are pre-attentively 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 [17][9]. Although definitions vary, intuitively, saliency corresponds to the 'rarity' of a feature [5]. 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 [14], is that consisting of pre-attentive and attentive stages. In the pre-attentive 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: baseline matching for stereo pairs, image retrieval from large databases, object retrieval in video, shot location, and object categorization [3][16], to mention just a few. One of the most well-known keypoint detector is the Harris detector [12]. 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 [6], Affine variants [12], DoG [10], etc. In [11], 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 [6]. In [4], a method for introducing the cornerness of the Harris detector in the method of [9] is proposed. Nevertheless, the robustness of the method is directly dependent on the cornerness performance. Kadir et al [9] estimate the entropy of the grey 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 [17]. However, using the grey level information, one can obtain regions with different complexity and with the same entropy values. Recently, the authors of [2] proposed the oriented-based SIFT descriptor such as a stability criterion to obtain stable scales for multi-scale 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 grey level information and shape complexity (defined by means of a novel normalized pseudo-histogram of orientations) to categorize the saliency levels. In literature, orientations have been previously used for saliency definition with very few success [9]. Our approach defines a normalized procedure

that makes this measure very relevant and robust.

The paper is organized as follows: chapter **??** 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 [8] and video sequences [1][7] 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 [9], Kadir et. al. introduce the grey-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 [9]. And, 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:

$$\gamma(S_p, x) = W^T(S_p, x)H(S_p, x) \quad (1)$$

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:

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

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 $S_p = \{s : \frac{\partial H(s,x)}{\partial s} = 0, \quad \frac{\partial^2 H(s,x)}{\partial s^2} < 0\}$

The entropy as a function of the scale space S is shown in fig. 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.

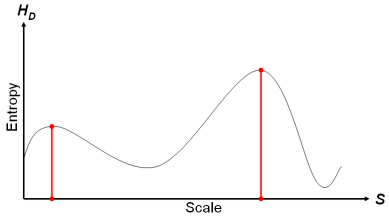


Figure 1. Local maxima of function H_D in the scale space S

The relevance of each position of the saliency at its representative scales is defined by the inter-scale saliency measure $W(s, x) = s\frac{\partial}{\partial s}H(s, x)$.

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:

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

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 grey-level and orientation saliency

Kadir et. al. [9] used the grey-level entropy to define the saliency complexity of a given region. However, this approach falls short in front of clear cases of different complexities. In fig. 2 one can observe different regions with the same amount of pixels for each grey level and different visual complexity. Note that the approach based on the grey-level entropy proposed by [9] gives the same entropy value, thus the same 'rarity' level for all of them.

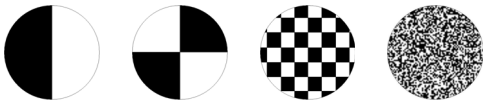
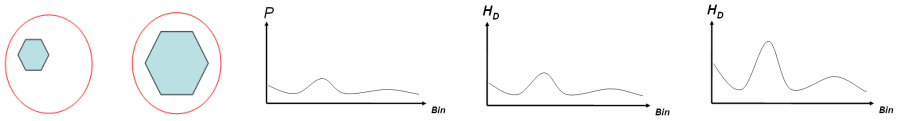


Figure 2. Regions of different complexity with the same grey level entropy.

A natural and well founded measure to solve this pathology is the use of complementary orientation information. In the same work [9], Kadir et. al. shows 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 fig. 3. Both regions have the same pdf (fig. 3(c)), but they contain different number of significant orientations (histograms of fig. 3(d) and (e)). In a regular histogram, low magnitude gradient is mostly due to noise, and it is distributed uniformly over all bins. Nevertheless, the pdf obtained in those cases remains the same because of the histogram



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

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).

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 pseudo-histogram 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 use 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 $[0..1]$. Then, we estimate its histogram, and the Otsu method [15] is applied to obtain the adaptive threshold for orientations. The significant orientation locations obtained for two image samples are shown in fig. 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:

$$PDF_O(i) = \frac{h_O(i)}{\sum_{j=1}^{n+1} h_O(j)}, \forall i \in [1,..,n] \qquad (4)$$

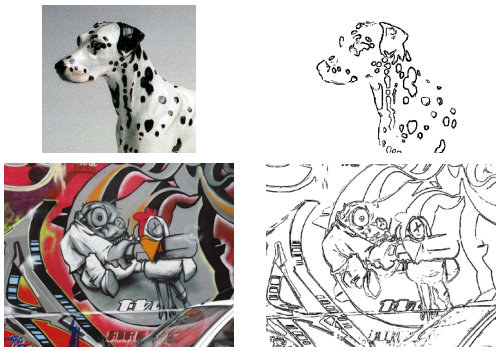


Figure 4. Relevant orientations estimation.

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[1].

2.4. Combining the saliency

In our particular case, the grey-level histogram is combined with the pseudo-histogram of orientations. We experimentally tested that the performance of both information offers better performance that only using the orientations or the grey-level entropy criterion. In this way, once estimated the two corresponding pdf, we apply equations (1), (2), and (3) to each one in the same way. The final measure is obtained by means of the simple addition[2] $\gamma = \gamma_G + \gamma_O$, where γ_G and γ_O are estimated by equation (1) for the grey and orientation saliency, and γ is the result, which contains the final significant saliency positions, magnitudes (level of complexity), and scales. This new saliency measure gives a high complexity value when the region contains different grey levels information (non-homogeneous 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

[1] Observe that the entropy measure of the *null-orientation* bin usually makes the first n bins non-significative.

[2] We experimentally observed that this simple combination obtains the most relevant results in comparison with other kinds of combinations.

pixels and possible scales have to be estimated. Nevertheless, the exhaustive search is relatively fast to compute (less than 1 second in a 800×640 medium resolution image).

To illustrate the effect of the combined saliency measure, we designed the toy problem of fig. 6. Figure 5 has 3 representative objects of different complexities. We applied the grey-level entropy, the orientation entropy, and the combined saliency. One can observe that the combined saliency measure selects the region with higher visual complexity (fig. 6(c)).

(a) (b) (c)

Figure 6. First maximal complexity region for grey-level entropy (a), orientations entropy (b), and combined entropy (c).

An example of CSR responses for an image sample under different transformations is shown in fig. 5. 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 grey-level saliency and the novel salient criterion using the Caltech database samples [8] of fig. 7 are shown in fig. 8. All images are of medium resolution (approximately 600×600 pixels). The size of the regions correspond 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 same time, the preferred regions for the novel salient regions are of reasonable sizes, which tipically implies have a higher discriminable power [13].

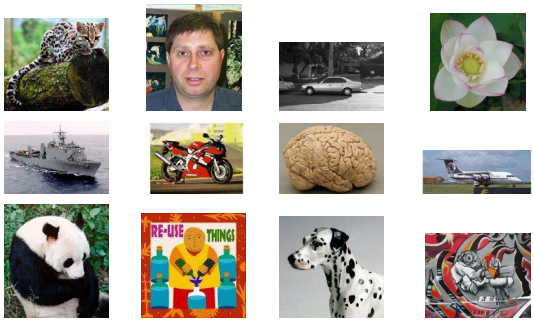


Figure 7. Caltech database samples.

As our orientations strategy normalize the input image it offers invariance to scalar changes in image contrast. The use of gradients is also invariant to an additive contrast change in brightness, which makes the technique invariant to illumination changes. Invariance to scale is obtained by the scale search of local maximums, and the use of circular

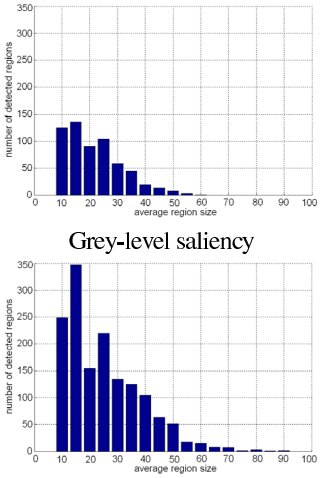


Figure 8. Histograms of mean region size and number of detected regions for the samples of fig. 7.

regions takes into account the global complexity of the inner of the regions, which also makes the strategy invariant to rotation.

3. Results

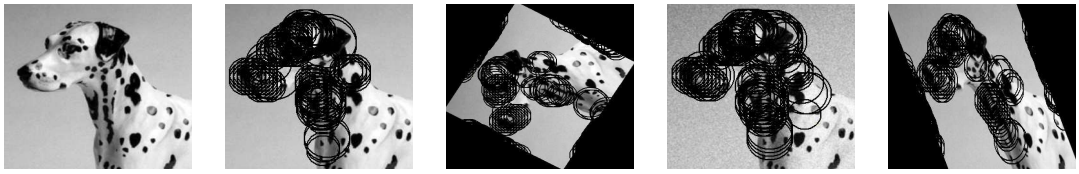
To validate the presented methodology, we should determine: data, measurements for the experiments, state-of-the-art methods to compare, and applications.

a) *Data*: Images are obtained from the public Caltech repository [8], and the video sequences from [1] and [7].

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 [13]. Besides, we include the false alarm rate measurement.

c) *State-of-the-art methods*: We compare the presented CSR with the Harris-Laplacian, Hessian-Laplacian, and the grey-level saliency. The parameters used for the region detectors are the default parameters given by the authors [11][9][12]. For the salient criteria of [9] and our CSR we use 16 bins for the grey-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 [9]. 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.



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

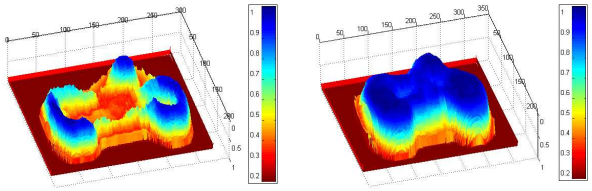
Figure 5. Image transformation tests for CSR responses: (a) input image, (b) initial CSR region detection, (c) 60 degree rotation, (d) white noise, and (e) affine transformation.

3.1. Grey-level Saliency versus CSR

We selected a set of 250 random motorbike samples from the motorbike Caltech database [8][3] and we estimated the highest saliency responses for each image using the grey-level saliency and the CSR regions. The mean volume image V of detected regions is shown in fig. 9. The volume image V is defined as:

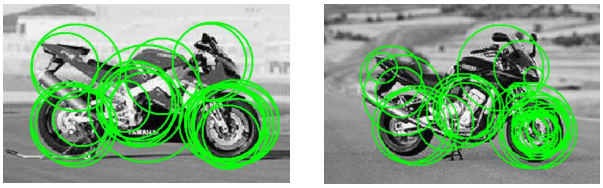
$$V = \frac{1}{i}\sum_{i=1}^{N} I_{R_i} \tag{5}$$

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 fig. 10, two examples of detected CSR for the motorbike database are shown.



(a) (b)

Figure 9. Mean volume image for the most relevant detected landmarks on the set of Caltech Motorbike database for grey Saliency (a) and our proposed CSR (b).



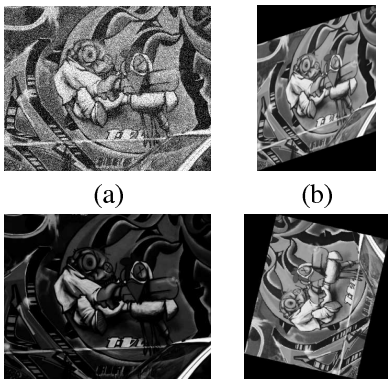
(a) (b)

Figure 10. Detected CSR from Caltech motorbike images.

[3]Note that 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.

3.2. Repeatability and False Alarm

In order to validate our results, we selected the samples showed in fig. 7 from the public Caltech repository database [8]. 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 $1/(1+\beta)$). Some examples of image transformations applied on the samples are shown in fig. 11.



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

Figure 11. Image transformations examples: (a) white noise addition, (b) affine distortion, (c) decreasing light, and (d) image rotation.

Over the set of transformations we apply the evaluation framework of [13] 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 [13]. Then, the 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 fig. 12. 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. Grey-based salient regions give relatively low repeatability and high false alarm rate, and it is dramatically improved with the 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 grey 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 are shown in figure 13.

The matching evaluation is based on the criterion of [13]. 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. The regions are described using the SIFT descriptor [10] and compared with the Euclidean distance. The overlap value is estimated using a warping technique to align manually the different samples. In fig. 14, 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 grey-level entropy and Hessian-Laplace detectors obtain better matching results. Nevertheless, the CSR regions obtain the highest percentage of matching for all overlap errors values.

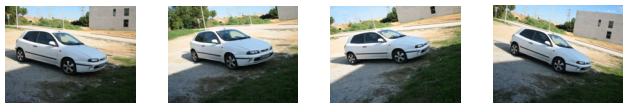


Figure 13. Car samples under different camera points of view.

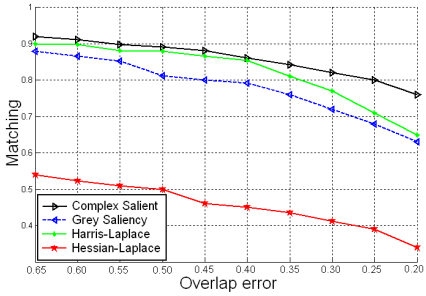


Figure 14. Matching percentage of the region detectors for the set of 30 car samples of different points of views in terms of regions intersection percentage.

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 [1]. The car system has six cameras that enable the system to collect video from more than 75% of the full sphere [1]. Furthermore, we also tested the method with road video sequences from the Geovan Mobile Mapping process from the Institut Cartogràfic de Catalunya [7], 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 [10] 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 fig. 15 and fig. 16. 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. Fig. 15(a) shows the oriented map in the first analyzed frame of [1]. Fig. 15(b) focuses on the right region of (a). One can see that the matched complex regions correspond to singularities in the video sequence and they approximates roughly the video movement. From the road experiment of fig. 15, the oriented map is shown in fig. 15(c). In this video sequence cars and traffic signs appear (fig. 15(a) and (b)). The amplified right region is shown in fig. 15(d). One can observe the correct movement trajectory of the road video sequences.

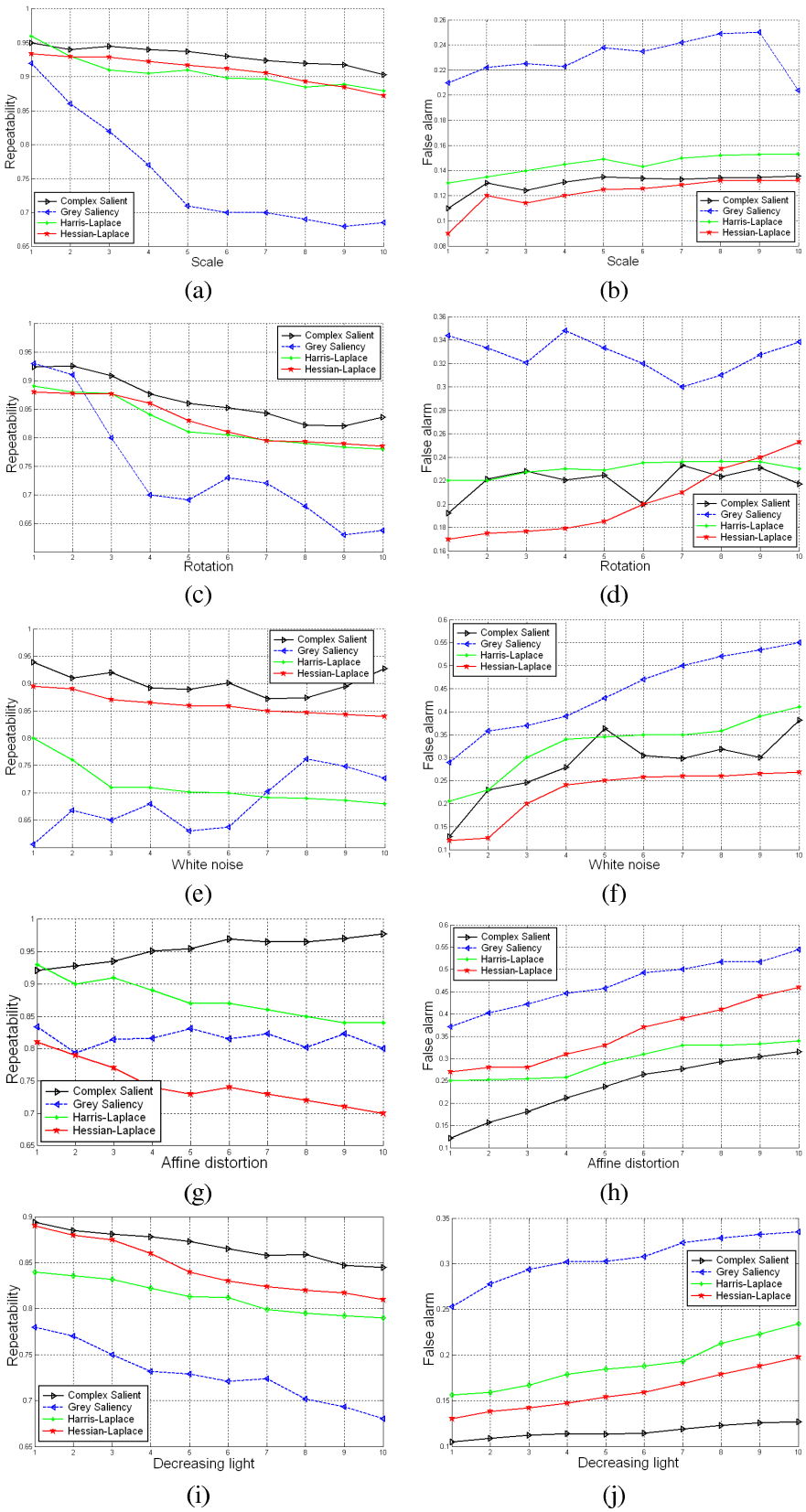
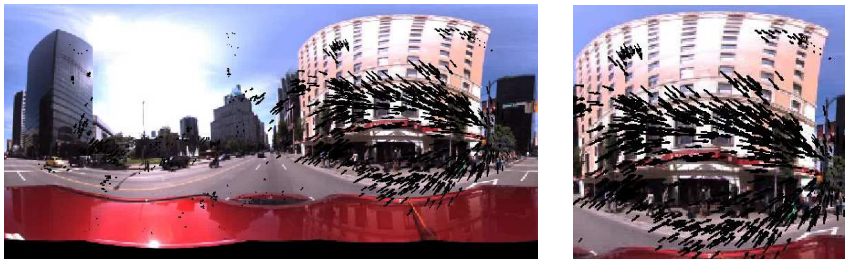


Figure 12. Repeatability and false alarm rate in the space of transformations: (a)(b) scale, (c)(d) rotation, (e)(f) white noise, (g)(h) affine invariants, and (i)(j) decreasing light.



(a) (b)

Figure 15. (a) Smoothed oriented CSR matches, (b) Zoomed right region.

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

Figure 16. (a)(b) Samples, (c) Smoothed oriented CSR matches, (d) Zoomed right region.

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 complexity of the regions using the grey-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. 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 now evaluating the methodology to design a multi-class object recognition approach. We want to categorize the CSR by complexity and group regions to model a recognition framework.

5. Acknowledgements

This work was supported in part by the projects, FIS-PI031488, MI-1509/2005, and TIN2006-15308-C02-02.

References

[1] adasdsa. 2, 4, 6

[2] G. Dorkó and C. Schmid. Maximally stable local description for scale selection. In *ICCV*, 2006. 1

[3] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *Proceedings IEEE Conference on CVPR*, 2003. 1

[4] F. Fraundorfer and H. Bischof. Detecting distinguished regions by saliency. *Image Analysis*, 2749:208–215, 2003. 1

[5] D. Hall, B. Leibe, and B. Schiele. Saliency of interest points under scale changes. *proc. of the British Machine Vision Conference*, 2002. 1

[6] C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, pages 147–151, 1999. 1

[7] http://www.icc.es. 2, 4, 6

[8] http://www.vision.caltech.edu/html files/archive.html. 2, 4, 5

[9] T. Kadir and M. Brady. Saliency, scale and image description. *IJCV*, 45(2):83–105, 2001. 1, 2, 4

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 20:91–110, 2003. 1, 6

[11] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. *Proc. of the British Machine Vision Conference*, 1:384–393, 2002. 1, 4

[12] K. Mikolajczyk and C. Schmid. Affine invariant interest point detectors. *International Journal of Computer Vision*, 60:63–86, 2004. 1, 4

[13] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. In *International Journal of Computer Vision*, volume 65, pages 43–72, 2005. 4, 5, 6

[14] U. Neisser. Visual search. *Scientific American*, 210(6):94–102, 1964. 1

[15] N. Otsu. A threshold selection method for gray level histograms. In *IEEE transactions on SMC*, 1979. 3

[16] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997. 1

[17] 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. *AIM*, 36, 2005. 1