12

Scene Recognition through Visual Attention and Image Features: A Comparison between SIFT and SURF Approaches

Fernando López-García[1], Xosé Ramón Fdez-Vidal[2],
Xosé Manuel Pardo[2] and Raquel Dosil[2]
[1]*Universidad Politécnica de Valencia*
[2]*Universidade de Santiago de Compostela*
Spain

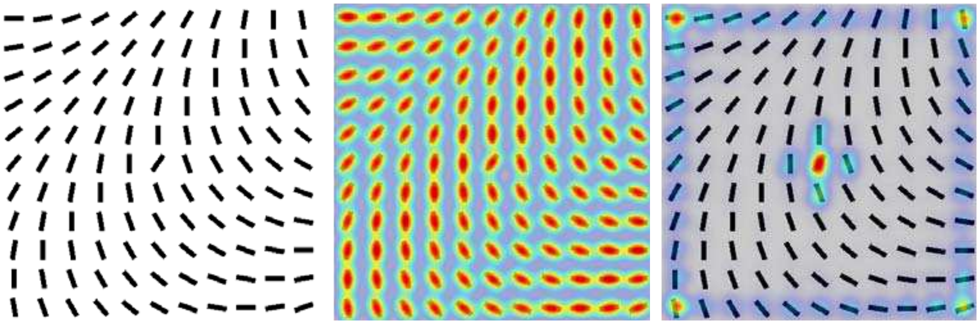
1. Introduction

In this work we study how we can use a novel model of spatial saliency (visual attention) combined with image features to significantly accelerate a scene recognition application and, at the same time, preserve recognition performance. To do so, we use a mobile robot-like application where scene recognition is carried out through the use of image features to characterize the different scenarios, and the Nearest Neighbor rule to carry out the classification. SIFT and SURF are two recent and competitive alternatives to image local featuring that we compare through extensive experimental work. Results from the experiments show that SIFT features perform significantly better than SURF features achieving important reductions in the size of the database of prototypes without significant losses in recognition performance, and thus, accelerating scene recognition. Also, from the experiments it is concluded that SURF features are less distinctive when using very large databases of interest points, as it occurs in the present case.

Visual attention is the process by which the Human Visual System (HVS) is able to select from a given scene regions of interest that contain salient information, and thus, reduce the amount of information to be processed (Treisman, 1980; Koch, 1985). In the last decade, several computational models biologically motivated have been released to implement visual attention in image and video processing (Itti, 2000; García-Díaz, 2008). Visual attention has also been used to improve object recognition and scene analysis (Bonaiuto, 2005; Walther, 2005). In this chapter, we study the utility of using a novel model of spatial saliency to improve a scene recognition application by reducing the amount of prototypes needed to carry out the classification task. The application is based on mobile robot-like video sequences taken in indoor facilities formed by several rooms and halls. The aim is to recognize the different scenarios in order to provide the mobile robot system with general location data.

The visual attention approach is a novel model of bottom-up saliency that uses local phase information of the input data where the statistic information of second order is deleted to achieve a Retinoptical map of saliency. The proposed approach joints computational mechanisms of the two hypotheses largely accepted in early vision: first, the *efficient coding*

(Barlow, 1961; Attneave, 1954), which postulates that the mission of the first stages of the visual processing chain is to reduce the redundancy or predictability in the incoming data; and second, in the visual cortex relevant attributes of the image are early detected using local phase or energy analysis, such as edges of objects. At those points where these features are located there is an alignment of the local phase of the Fourier harmonics (Phase Congruency). The model of local energy to detect features (Morrone & Burr, 1988; Morrone & Owens, 1987; Kovesi, 1999) is based on this idea and demonstrated its suitability for perceptual appearance and image segmentation. Nevertheless, it is not able to prioritize the features with regards to the visual saliency. This fact is illustrated in Figure 1, where the input image is formed by bars that increment its orientation in steps of 10^{o} from left to right and top to bottom, except for the central bar that breaks this periodicity creating a pop-out effect for the HVS.



a) Original Image. b) Salience from PC. c) Salience from our model.

Fig. 1. Saliency maps of the original image; (a) from Phase Congruency (b) and the proposed model (c).

In Fig. 1b we see the map of saliency achieved using Kovesi's model (Kovesi, 1999) based on Phase Congruency (PC). It provides approximately equal weight to all features clearing away the pop-out effect. We think the reason for that is the high redundancy in images which implies correlations and thus Gaussianism in chromatic and spatial components. It is known that to handle information about phase structure is equivalent to use non-Gaussian information in data distribution (Hyvärinen et al., 2009). Thus, to focus in the information that does not depend on covariances (local phase) it is necessary to reduce redundancy, that is, to decorrelate the data. One way is through data whitening. Redundancy in RGB color components is deleted through PCA and spatial redundancy is avoided using an strategy of filter-based whitening in frequency domain. In Fig. 1c it is shown that this hypothesis works making possible to prioritize the salience of visual features from local phase.

Scene recognition is performed using SIFT (Lowe, 2004) and SURF (Bay, 2008) for image featuring (two different approaches that we compare) and the Nearest Neighbor rule for classification. SIFT features are distinctive image features that are invariant to image scale and rotation, and partially invariant to change in illumination and 3D viewpoint. They are fast to compute and robust to disruptions due to occlusion, clutter or noise. SIFT features have proven to be useful in many object recognition applications and currently they are considered the state-of-the art for general purpose and real-world object learning and recognition, together with SURF features. SURF is a robust image descriptor used in computer vision tasks like object recognition or 3D reconstruction. The standard version of SURF is several times faster than SIFT and it is claimed by its authors to be more robust against different image transformations than SIFT. However, the results of our experimental work showed that SIFT features perform significantly better than SURF features. In combination with saliency maps, SIFT features lead to drastic reductions in the number of interest points introduced in the database of prototypes (used in 1-NN classification), also achieving very good performance in scene recognition. Thus, since the computing costs of classification are significantly reduced the scene recognition is accelerated.

The chapter is developed as follows. Next Section presents the model of spatial saliency. An overview of the image featuring methods is provided in Section 3. Section 4 deals with the scene recognition application. Experimental work and results are presented in Section 5. Finally, Section 6 is devoted to conclusions.

2. Model of spatial saliency

Figure 2 shows a general flow diagram of the saliency model. Following we describe each stage of the model.

2.1 Early stage

The goal of this initial stage is to delete the statistical information of second order in color components (RGB) and spatial components (between pixels of each color component), through different whitening processes.

The aim of the initial step in this stage is to provide the model with a color space that contains a mechanism, biologically inspired, called *short-term adaptation* (Simoncelli & Olshausen, 2001; Barlow & Foldiak, 1989), which main goal is to achieve a final synchronization in the adaptive process that promotes the most useful aspects for later processing. For that, the color RGB image is decomposed into three channels maximally decorrelated using Principal Component Analysis (PCA). Nevertheless, we are not interested in reducing the color space dimension, thus, we use a transformed space of the original dimension (3 color components). The first component corresponds to opponents channel B/W and the remaining two correspond to opponents similar to R/G and Y/B. However, the space, unlike the opponents space CIE-Lab, is adapted to the specific statistic of the incoming image.

In a second step, the goal is to eliminate the spatial redundancy among the pixels in each color channel. In this case, we use a filter-based strategic in frequency domain called Spectral Whitening (SW). It is consequence of the Wiener-Khinchin theorem: "for a stochastic process, *the average power spectrum is the Fourier Transform of the autocorrelation function*". Thus, a whitened image should have a flat power spectrum. This can be easily achieved using an adaptive filter in the frequency domain that normalizes the spectrum of the transformed Fourier corresponding to the incoming image I(x,y) in the following way:

$$n\left(\omega_x,\omega_y\right)=\frac{\Im\left[I\left(x,y\right)\right]}{\left\|\Im\left[I\left(x,y\right)\right]\right\|}=\frac{f\left(\omega_x,\omega_y\right)}{\left\|f\left(\omega_x,\omega_y\right)\right\|} \tag{1.1}$$

where $\Im[\cdot]$ is the transformed Fourier and $\omega_s=\sqrt{\omega_x^2+\omega_y^2}$ is the spatial frequency. Physically, SW is a redistribution of the spectrum energy that will achieve an enhancement

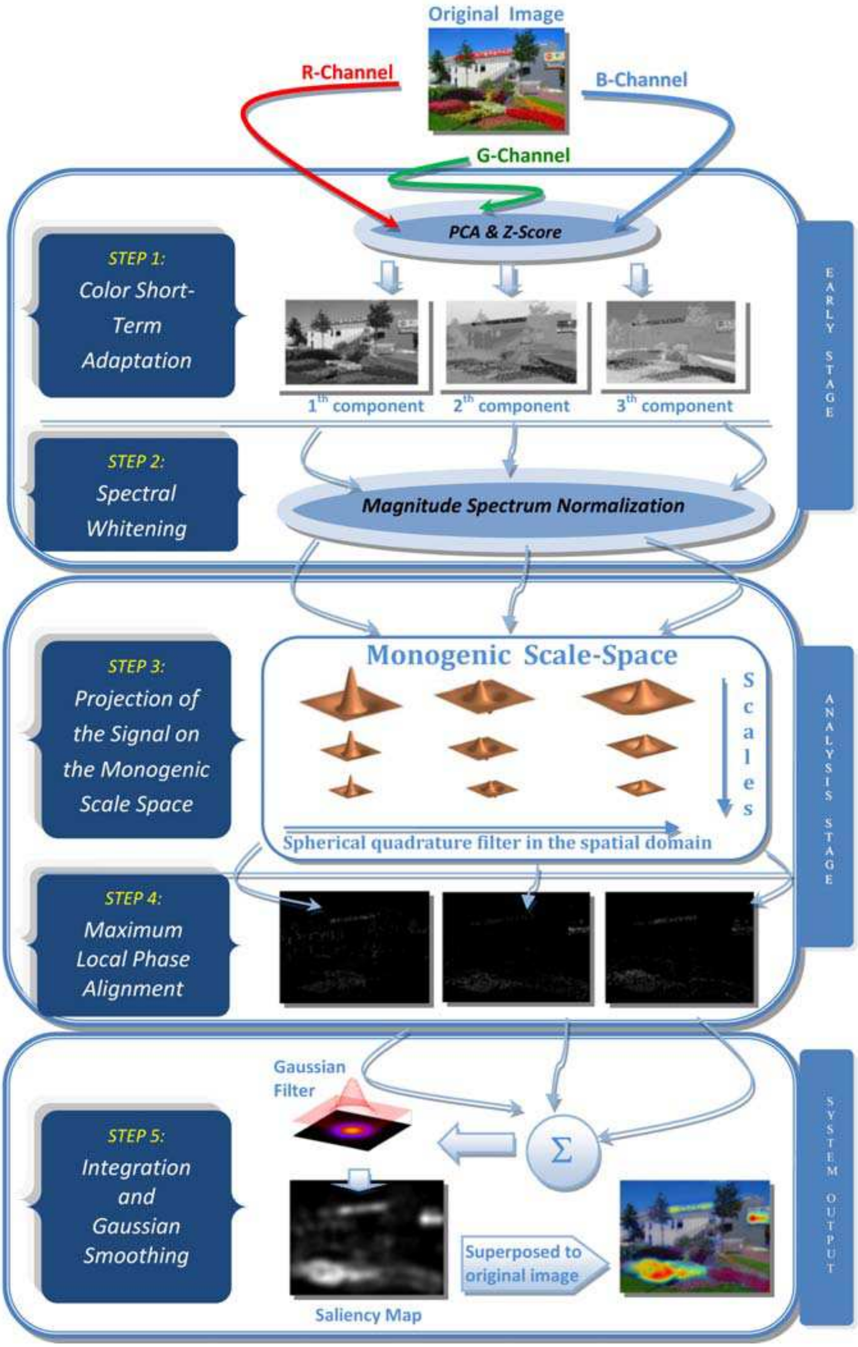

Original Image
R-Channel
B-Channel
G-Channel
PCA & Z-Score
STEP 1:
Color Short-Term Adaptation
1th component
2th component
3th component
STEP 2:
Spectral Whitening
Magnitude Spectrum Normalization
EARLY STAGE
STEP 3:
Projection of the Signal on the Monogenic Scale Space
Monogenic Scale-Space
Scales
Spherical quadrature filter in the spatial domain
STEP 4:
Maximum Local Phase Alignment
ANALYSIS STAGE
Gaussian Filter
Σ
STEP 5:
Integration and Gaussian Smoothing
Superposed to original image
Saliency Map
SYSTEM OUTPUT

Fig. 2. General diagram showing how the data flows through the model.

of the less redundant patterns. This whitening method was previously used in the saliency model proposed by Guo et al. (Guo et al., 2008) which is based on global phase analysis.

2.2 Analysis stage

In this stage, it is analyzed the maximum alignment of the local phase of each pixel in each whitened color channel weighted by the strength of the visual features in the analyzed scale. Classical methodology to estimate the amplitude and phase in a 1D signal is the Analytic Signal. However, the 2D version was achieved partially using a quadrature phase bank filter (Gabor like filters), until the introduction of the Monogenic Signal by Felsberg & Sommer (Felsberg & Sommer, 2001). Our model uses this last methodology that achieves a new 2D analytic signal from the Riesz's transform, which is the 2D isotropic extension of Hilbert's transform. Its representation in Fourier's domain is a set of two simple filters in phase-quadrature that are not selective neither in scale nor orientation:

$$\left(H_1\left(\omega_x,\omega_y\right),H_2\left(\omega_x,\omega_y\right)\right)=\left(\frac{\omega_x}{\sqrt{\omega_x^2+\omega_y^2}}i,\frac{\omega_y}{\sqrt{\omega_x^2+\omega_y^2}}i\right) \tag{1.2}$$

The Monogenic Signal is a vector function of three components formed by the original signal and two components achieved by convolving it with the filters of Riez's transform, that is:

$$\vec{f}_M(x,y)=\left[f(x,y),f(x,y)*h_1(x,y),f(x,y)*h_2(x,y)\right] \tag{1.3}$$

where $h_1(x,y)$ and $h_2(x,y)$ are the representations in the spatial domain of $H_1(\omega_x,\omega_y)$ and $H_2(\omega_x,\omega_y)$ respectively. Because filters H_1 and H_2 are oriented in frequency domain but are not selective in scale, commonly a Gaussian like band-pass filter is used to build scaled versions of Riez's filters. In our case, we used the following log-Gauss filter:

$$G_s\left(\omega_x,\omega_y\right)=e^{-\left(\frac{\log\left(\omega/\omega_o\right)^2}{2\left(\log\left(k/\omega_o\right)\right)^2}\right)} \tag{1.4}$$

where $\omega=(\omega_x,\omega_y)$ is the spatial frequency, $\omega_o=(\omega_{ox},\omega_{oy})$ is the central frequency of the filter and *k* is the parameter that governs the bandwidth of the filter. If $g_s(x,y)$ is the spatial representation of previous filter, the monogenic space of scales is built as follows:

$$\begin{aligned}\vec{f}_{M,s}(x,y)&=\left[f(x,y)*g_s(x,y),f(x,y)*g_s(x,y)*h_1(x,y),f(x,y)*g_s(x,y)*h_2(x,y)\right]=\\&=\left[f_s(x,y),h_{1,s}(x,y),h_{2,s}(x,y)\right]\end{aligned} \tag{1.5}$$

The chosen bank of filters is formed by three scales (s=3) which central wavelengths were distributed in 1 octave from the minimum wavelength (assigned to λ_1=8 pixels), that is λ_i={8, 16, 32} pixels. The *k* parameter was fixed to achieve a bandwidth of 2 octaves in each filter in order to obtain a good spectral coverage in the bank of filters. A simple implementation of the monogenic signal, in the frequency domain, can be found in (Kovesi, 2000).

Once it is achieved the monogenic decomposition, the importance of each visual feature is measured by maximizing in each pixel of the image and for all the scales, the level of local phase alignment of the Fourier Harmonics, weighted by the strength of the visual structure in each scale (measured as local energy $\left\| f_{M,i}(x,y) \right\|$). We call this measure Weighted Maximum Phase Alignment (WMAP), and is the following:

$$
\begin{aligned}
WMPA(x,y) &= \mathrm{w}_{\mathrm{fdn}} \cdot \max_{i=1}^{s} \left\{ \left\| \vec{f}_{M,i}(x,y) \right\| \cdot \cos\theta_i \right\} = \\
&= \mathrm{w}_{\mathrm{fdn}} \cdot \max_{i=1}^{s} \left\{ \vec{f}_{M,i}(x,y) \bullet \left(\frac{\vec{E}_{local}(x,y)}{\left\| \vec{E}_{local}(x,y) \right\|} \right) \right\} = \\
&= \mathrm{w}_{\mathrm{fdn}} \cdot \max_{i=1}^{s} \left\{ \left(f_i(x,y), h_{1,i}(x,y), h_{2,i}(x,y) \right) \cdot \left(\frac{\left(\sum_{i=1}^{s} f_i(x,y), \sum_{i=1}^{s} h_{1,i}(x,y), \sum_{i=1}^{s} h_{2,i}(x,y) \right)}{\left(\sqrt{ \left(\sum_{i=1}^{s} f_i(x,y) \right)^2 + \left(\sum_{i=1}^{s} h_{1,i}(x,y) \right)^2 + \left(\sum_{i=1}^{s} h(x,y) \right)^2 } \right)} \right) \right\}
\end{aligned}
\tag{1.6}
$$

where $\vec{f}_{M,i}(x,y)$ is the monogenic signal for the i-*th* scale and θ_i is the angle between vectors $\vec{f}_{M,i}(x,y)$ and $\vec{E}_{local}$. This angle measures the deviation of the local phase in the monogenic signal at the i-*th* scale respect to the local energy vector in pixel (x,y).

We are only interested on those pixels where local phase is congruent for the most of the used scales. Thus, our measure must incorporate a factor that penalizes too narrow frequency distributions. Factor $\mathrm{w}_{\mathrm{fdn}}$ is achieved as it was proposed by Kovesi (Kovesi, 1999) for his measure of local Phase Congruency (PC).

2.3 Output stage

The final stage of the model has the aim of achieving a Retinoptic measure of the salience of each pixel in the image. For that, we integrate in each pixel the WMPA(x,y) measures of each color channel:

$$
Saliency(x,y) = \sum_{c=1}^{3} WMAP(x,y) \tag{1.7}
$$

Finally, a smoothing is introduced by a Gaussian filter and also a normalization in order to make easy to interpret the saliency map as a probability function to receive attention.

2.4 Computational complexity

The computational efficiency of the model is low due to the load introduced by the PCA analysis, which grows lineally with the number of pixels in the image (N) and cubically with the number of components (color channels), O(M^3+N M^2). The number of components is low and constant, M=3, thus, the asymptotic complexity depends on N. The computational complexity of the model depends on the FFT (Fast Fourier Transform) complexity performed in filtering processing. This complexity is O(N log(N)). On the other hand, the computational timing of the model is low, by example, for an image of 512x384 pixels using an Intel Core2 Quad processor at 2.4 GHz and 4Gb of RAM memory, the algorithm takes 0.91 seconds. We have to take into account that the algorithm is scientific software programmed in MATLAB.

3. Image features

SIFT and SURF belong to a set of methods aimed to detect and describe local features in images. Among these methods we can found (Mikolajczyk, 2005): shape context, steerable filters, PCA-SIFT, differential invariants, spin images, complex filters, moment invariants and gradient location and orientation histograms (GLOH). Nevertheless, SIFT and SURF have captured recent attention of researchers working on applications like object recognition, robot mapping and navigation, image stitching, 3D modeling, video tracking, etc, being its comparison a current issue in literature (Bauer, 2007).

With regards to SIFT features, we used the Lowe´s algorithm (Lowe, 2004) which works as follows. To identify the interest points (keypoints), scale space extrema are found in a difference-of-Gaussian (DoG) function convolved with the image. The extremas are found by comparing each point with its neighbors in the current image and adjacent scales. Points are selected as candidate keypoint locations if they are the maximum or minimum value in their neighborhood. Then image gradients and orientations, at each pixel of the Gaussian convolved image at each scale, are computed. For each key location an orientation, determined by the peak of a histogram of previously computed neighborhood orientations, is assigned. Once the orientation, scale, and location of the keypoints have been computed, invariance to these values is achieved by computing the keypoint local feature descriptors relative to them. Local feature descriptors are 128-dimensional vectors obtained from the pre-computed image orientations and gradients around the keypoints.

SURF features (Bay, 2008) are based on sums of 2D Haar wavelet responses and make a very efficient use of integral images to speed-up the process. As basic image descriptors they use a Haar wavelet approximation of the determinant of Hessian blob detector. There are two versions: the standard version which uses a descriptor vector of 64 components (SURF-64), and the extended version which uses 128 components (SURF-128). SURF are robust image features partly inspired by SIFT, being the standard version of SURF several times faster than SIFT. SURF features provide significantly less keypoints than SIFT, approximately the half of them (see Figure 3).

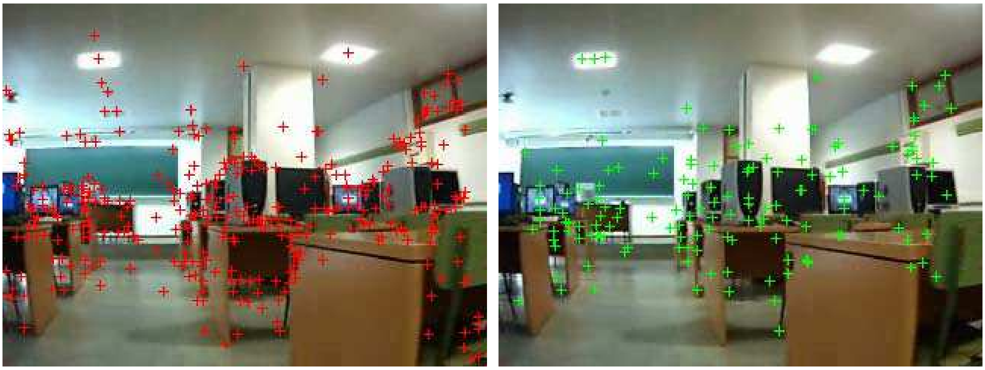


Fig. 3. SIFT (left) and SURF (right) keypoints computed for the same frame.

4. Scene recognition application

Scene recognition is related with the recognition of general scenarios rather than local objects. This approach is useful in many applications such as mobile robot navigation, image retrieval, extraction of contextual information for object recognition, and even to provide access to tourist information using camera phones. In our case, we are interested in recognize a set of different scenarios which are part of university facilities formed by four class rooms and three halls. The final aim is to provide general location data useful for the navigation of a mobile robot system. Scene recognition is commonly performed using generic image features that try to collect enough information to be able to distinguish among the different scenarios. For this purpose we used SIFT and SURF alternatives.

To compute the SIFT features we used the original code by Lowe (http://people.cs.ubc.ca/lowe/keypoints/). We also used the original code for SURF features by Bay et al (http://www.vision.ee.ethz.ch/~surf/). To carry out the classification task we used the 1-NN rule, which is a simple classification approach but fast to compute and robust. For the 1-NN approach, we need to build previously a database of prototypes that will collect the recognition knowledge of the classifier. These prototypes are a set of labelled SIFT/SURF keypoints obtained from the training frames. The class of the keypoints computed for a specific training frame will be that previously assigned to this frame in an off-line supervised labeling process. The database is then incorporated into the 1-NN classifier, which uses the Euclidean distance to select the closest prototype to the test SIFT/SURF keypoint being classified. The class of every test keypoint will be assigned to the class of the closest prototype in the database, and finally, the class of the entire test frame will be that of the majority of its keypoints.

5. Experiments and results

The experimental work consisted in a set of experiments carried out using four video sequences taken in a robot-navigation manner. These video sequences were grabbed in an university area covering several rooms and halls. Sequences were taken at 5 fps collecting a total number of 2,174 frames (7:15 minutes) for the first sequence, 1,986 frames for the second (6:37 minutes), 1,816 frames for the third (6:03 minutes) and 1,753 frames for the fourth (5:50 minutes). First and third sequences were taken in a specific order of halls and rooms: hall-1, room-1, hall-1, room-2, hall-1, room-3, hall-1, hall-2, hall-3, room-4, hall-3, hall-2, hall-1. The second and fourth sequences were grabbed following the opposite order to collect all possible viewpoints of the robot navigation through the facilities. In all the experiments, we used the first and second sequences for training and the third and fourth for testing.

In the first experiment we computed the SIFT keypoints for all the frames of the training video sequences. Then, we labelled these keypoints with the corresponding frame class: room-1, room-2, room-3, room-4, hall-1, hall-2 or hall-3. The whole set of labelled keypoints formed itself the database of prototypes to be used by the 1-NN classifier. For each frame of the testing sequences their corresponding SIFT keypoints were computed and classified. The final class for the frame was set to the majority class among its keypoints. Very good performance was achieved, 95.25% of correct classification of frames. However, an important drawback was the computational cost of classification, which was high despite the fact that 1-NN is known as a low cost classifier. This was due to the very large size of the database of prototypes formed by 1,170,215 samples. In the next experiment, we followed the previous steps but using SURF features instead of SIFT. In this case, recognition results were very bad achieving only 28.24% of recognition performance with SURF-128 features, and 25.05% using SURF-64. In both SURF cases the size of the database of prototypes was of 415, 845.

Although there are well known techniques for NN classifiers to optimize the database of prototypes (e.g. feature selection, feature extraction, condensing, editing) and also for the acceleration of the classification computation (e.g. kd-trees), at this point we are interested in the utility of using the saliency maps derived from the visual attention approach. The idea is to achieve significant reductions of the original database by selecting in each training frame only those keypoints that are included within the saliency map computed for this frame. Also, in the testing frames only those keypoints lying within the saliency maps will be considered for classification. Once the database is reduced in this way, optimizing techniques could be used to achieve even further improvements.

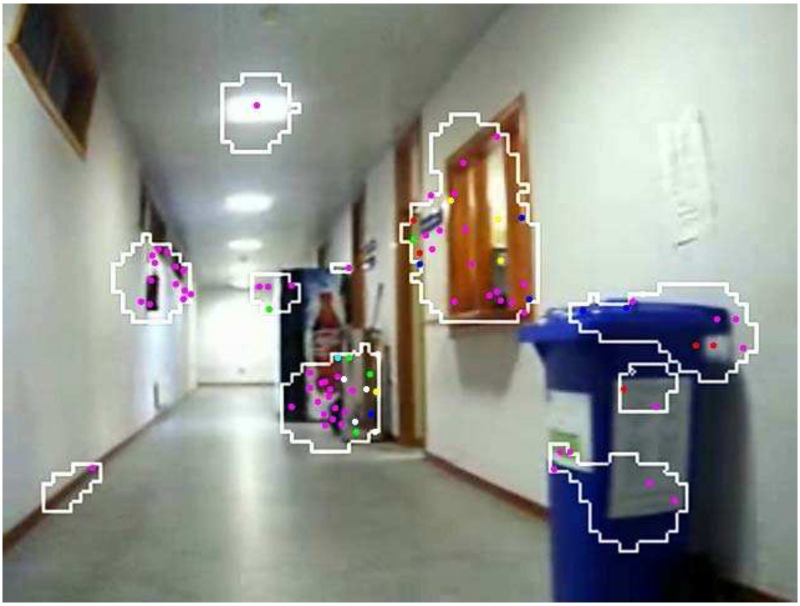


Fig. 4. Saliency regions at threshold 0.250 and corresponding SIFT keypoins.

In next experiments we carried out the idea showed in previous paragraph, although we wanted to explore more in-depth the possibilities of saliency maps. As it was commented, saliency measures are set in a range between 0 and 1, thus, we can choose different levels of saliency by simply using thresholds. We will be the least restrictive if we choose a saliency > 0.0, and more restrictive if we choose higher levels (e.g. 0.125, 0.250, etc). We planned to use eigth different saliency levels: 0.125, 0.250, 0.375, 0.500, 0.625, 0.750 and 0.875. For each saliency level we carried out the scene recognition experiment (see Figure 4) achieving the percentage of recognition performance, and the size of the database of prototypes. Results using SIFT and SURF features are shown in Tables 1, 2 and 3 and Figures 5, 6 and 7.

	Recognition %	Database Size	Database Size %
Original	95.25	1,170,215	100.0
Saliency > 0.125	95.25	779,995	66.65
Saliency > 0.250	94.72	462,486	39.52
Saliency > 0.375	93.45	273,908	23.41
Saliency > 0.500	92.21	157,388	13.45
Saliency > 0.650	89.30	86,161	7.36
Saliency > 0.750	83.31	42,418	3.62
Saliency > 0.875	56.03	15,894	1.36

Table 1. Results achieved using original frames and saliency maps with SIFT features.

	Recognition %	Database Size	Database Size %
Original	28.24	415,845	100.0
Saliency > 0.125	33.51	273,775	65.84
Saliency > 0.250	86.56	157,394	37.85
Saliency > 0.375	32.01	88,059	21.18
Saliency > 0.500	66.55	47,767	11.49
Saliency > 0.650	67.06	24,338	5.85
Saliency > 0.750	35.27	11,040	2.65
Saliency > 0.875	18.33	3,971	0.95

Table 2. Results achieved using original frames and saliency maps with SURF-128 features.

	Recognition %	Database Size	Database Size %
Original	25.05	415,845	100.0
Saliency > 0.125	27.74	273,775	65.84
Saliency > 0.250	51.50	157,394	37.85
Saliency > 0.375	25.64	88,059	21.18
Saliency > 0.500	28.97	47,767	11.49
Saliency > 0.650	67.33	24,338	5.85
Saliency > 0.750	34.89	11,040	2.65
Saliency > 0.875	19.22	3,971	0.95

Table 3. Results achieved using original frames and saliency maps with SURF-64 features.

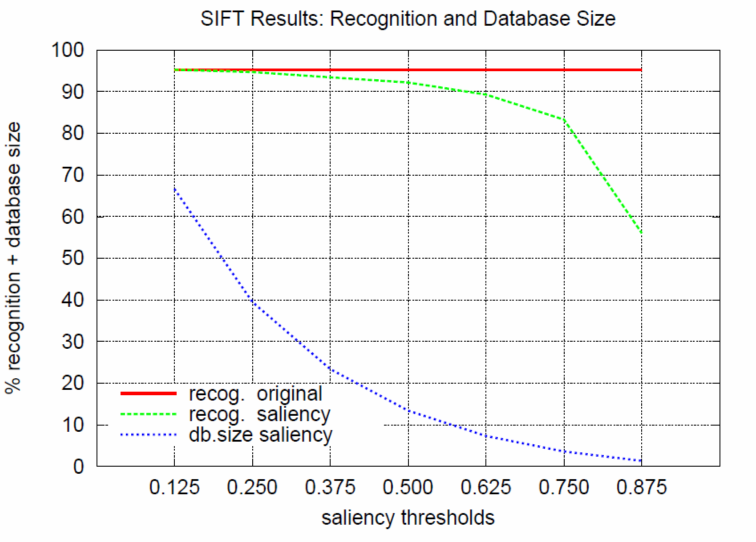


Fig. 5. Graphical results of recognition and database size using SIFT features.

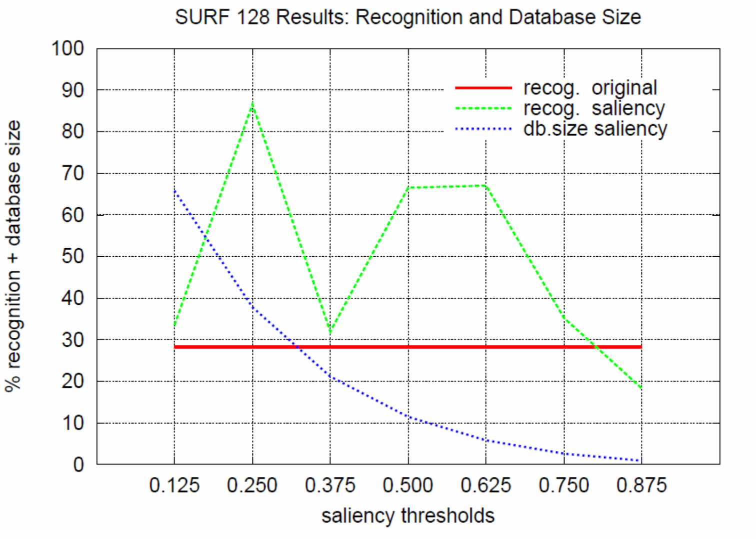


Fig. 6. Graphical results of recognition and database size using SURF-128 features.

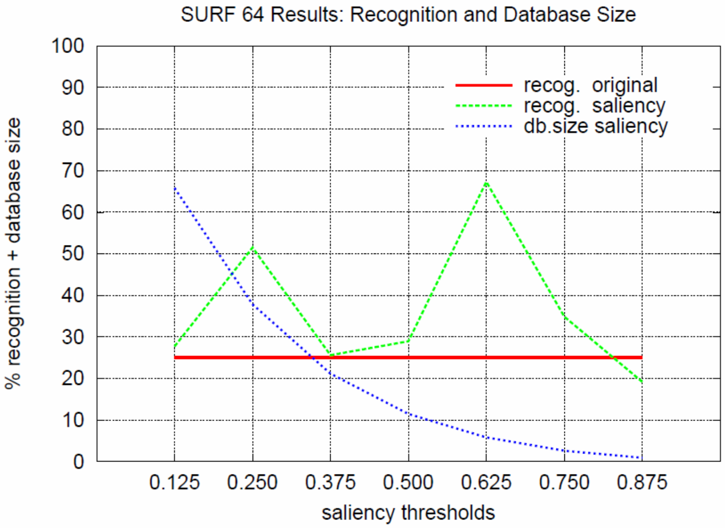


Fig. 7. Graphical results of recognition and database size using SURF-64 features.

Experimental results show that although SURF features collect significantly less interest points than SIFT features (approximately the half of them) their performance is not adequate for the scene recognition application. However, SURF features have proven to be adequate, and faster than SIFT features, in other applications (Bay, 2008). Another interesting result is that recognition performance of SURF features shows an irregular behavior with the saliency thresholds, in both cases, SURF-64 and SURF-128. A maximum peak of 86.56% is reached at saliency level 0.250 in SURF-128, while recognition results provided by SURF-64 features are worse. When using no saliency maps and even with some less restrictive thresholds, recognition results of SURF features are very bad. This means that SURF features loose distinctiveness as more interest points are used. This fact does not occur in SIFT features, thus, SIFT features present more distinctiveness than SURF features in very large databases of interest points. The best results are achieved using SIFT features, which combined with saliency maps can reduce the amount of prototypes in the database up to one order of magnitude, while the recognition performance is held, e.g. saliency level 0.500 in Table 1 and Figure 5. In this case, the performance drops to 92.21% (only 3.04 points from 95.25%) while the database size is drastically reduced from 1,170,215 to 157,388 prototypes.

6. Conclusions

In this work, scene recognition is carried out using a novel biologically inspired approach to visual attention in combination with local image features. SIFT and SURF approaches to image featuring are compared. Experimental results show that despite SURF features imply the use of less interest points the best performance corresponds by far to SIFT features. The SIFT method achieves a 95.25% of performance on scene recognition in the best case, while the SURF method only reaches 86.56%. Another important result is achieved when we use the saliency maps from the visual attention approach in combination with SIFT features. In this case, the database of prototypes, used in the classification task of scene recognition, can be drastically reduced (up to one order of magnitude) with a slightly drop in recognition performance. Thus, the scene recognition application can be significantly speeded-up. In addition, the experiments show that SURF features are less distinctive than SIFT features when we use very large databases of interest points.

7. Acknowledgements

Work supported by the Ministry of Education and Science of the Spanish Government (AVISTA-TIN2006-08447) and the Government of Galicia (PGIDIT07PXIB206028PR).

8. References

Attneave, F. (1954). Informational aspects of visual perception. *Psychological Review*, Vol. 61, No. 3, pp. 183-193, ISSN 0033-295X (print), ISSN 1939-1471 (online).

Bauer, J.; Sünderhauf, N. & Protzel, P. (2007). Comparing Several Implementations of Two Recently Published Feature Detectors. *Proceedings of The International Conference on Intelligent and Autonomous Systems*, (IAV), Toulouse, France.

Barlow, H.B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication*, pp. 217-234.

Barlow, H.B. & Foldiak, P. (1989). Adaptation and decorrelation in the cortex. In *The Computing Neuron*, pp. 54-72, Addison-Wesley, ISBN 0-201-18348-X.

Bay, H.; Ess, A.; Tuytelaars, T. & Gool, L. V. (2008). Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, Vol. 110, No. 3, pp. 346-359, ISSN 1077-3142.

Bonaiuto, J. J. & Itti, L. (2005). Combining Attention and Recognition for Rapid Scene Analysis, *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05)*, pp. 90-90, ISBN 978-88-89884-09-6, San Diego (USA), June 2005, IEEE Computer Society.

Felsberg, M. & Sommer, G. (2001). The Monogenic Signal. *IEEE Transactions on Signal Processing*, Vol. 49, No. 12, pp. 3136-3144, ISSN 1053-587X.

García-Díaz, A.; Fdez-Vidal, X. R.; Dosil, R. and Pardo, X. M. (2008). Local Energy Variability as a Generic Measure of Bottom-Up Salience, In: *Pattern Recognition Techniques, Technology and Applications*, Peng-Yeng Yin (Ed.), pp. 1–24 (Chapter 1), In-Teh, ISBN 978-953-7619-24-4,Vienna.

Guo, C.L.; Ma, Q. & Zhang, L.M. (2008). Spatio-Temporal Saliency Detection Using Phase Spectrum of Quaternion Fourier Transform. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 2008 (CVPR'2008)*, Nº 220, IEEE, pp. 1-8. ISBN 9781424422425.

Hyvärinen, A.; Hurri, J. & Hoyer, P.O. (2009). Natural Image Statistics. A probabilistic approach to early computational vision. Springer. ISBN 978-1-84882-491-1.

Itti, L. & Koch, C. (2000). A Saliency-based Search Mechanism for Overt and Covert Shifts of Visual Attention. *Vision Research*, Vol. 40, pp. 1489–1506, ISSN 0042-6989.

Koch, C. & Ullman, S. (1985). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Human Neurobiology*, Vol. 4, No. 4, pp. 219-227, ISSN 0721-9075.

Kovesi, P.D. (1999). Image features from phase congruency. Videre: *Journal of Computer Vision Research*. MIT Press. Vol. 1, No. 3.
Available from: http://mitpress.mit.edu/e-journals/Videre/001/v13.html

Kovesi, P.D. (2000). MATLAB and Octave Functions for Computer Vision and Image Processing. School of Computer Science & Software Engineering, The University of Western Australia.
Available from: http://www.csse.uwa.edu.au/~pk/research/matlabfns/

Lowe, D. G. (2004). Distinctive Image Features from Scale-invariant Keypoints. *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110, ISSN 0920-5691 (print), ISSN 1573-1405 (online).

Mikolajczyk, K. & Schmid, C. (2005). A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 10, pp. 1615-1630, ISSN 0162-8828.

Morrone, M.C. & Burr, D.C. (1988). Feature detection in human vision: A phase-dependent energy model. *Proceedings of the Royal Society of London B: Biological Sciences*, Vol. 235, No. 1280 , pp. 221–245, ISSN 1471-2954.

Morrone, M.C. & Owens, R. (1987). Feature detection from local energy. *Pattern Recognition Letters*, Vol. 6, No. 5, pp. 303–313, ISSN 0167-8655.

Simoncelli, E.P. & Olshausen, B.A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, Vol. 24, No. 1, pp. 1193-1216, ISSN 0147-006X.

Treisman, A. & Gelade, G. (1980). A Feature Integration Theory of Attention. *Cognitive Psychologhy*, Vol. 12, pp. 97–136, ISSN 0010-0285.

Walther, D.; Rutishauser, U.; Koch, C. & Perona, P. (2005). Selective Visual Attention Enables Learning and Recognition of Multiple Objects in Cluttered Scenes. *Computer Vision and Image Understanding*, Vol. 100, pp. 1-63, ISSN 1077-3142.