APPEARANCE BASED SALIENT POINT DETECTION WITH INTRINSIC SCALE-FREQUENCY DESCRIPTOR

Plinio Moreno, Alexandre Bernardino and José Santos-Victor
Instituto de Sistemas e Robótica
Instituto Superior Técnico
1049-001 Lisboa - Portugal
{plinio, alex, jasv}@isr.ist.utl.pt

Abstract

Recent object recognition methods propose to represent objects by collections of local appearance descriptors in several interest points. For recognition, this representation is matched to image data. Interest points (candidates for matching) are usually selected from images in a purely bottom-up manner. However, in many situations, there is a limited number of objects to search for, and some information about object characteristics should be employed in the detection of salient points, to reduce the number of potential candidates. In this paper we propose a methodology for the selection of candidates with prior information of the object local appearance. Points are represented by a rotation and scale invariant descriptor, composed by the response of filters derived from Gabor functions, denoted as "intrinsic scale/frequency descriptor". When compared to classical salient point detectors, like extremal points of laplacian operators at several scales, the proposed methodology is able to reduce the amount of canditates for matching by more than 60%. Since matching is a costly operation, this strategy will improve the efficiency of object recognition methods.

Keywords: Gabor filter, saliency, interest point, object detection

1 Introduction

The object recognition problem has been tackled recently with several successful results [5, 10, 7, 12]. All of these works exploit the idea of selecting various points in the object and building up a local neighborhood representation for each one of the selected points. Two related problems are involved in this process: (i) which points in the image should be considered and (ii) how to represent the information contained in their neighborhood. In previous work [8] we have addressed the second problem by describing each points neighborhood with the response of Gabor filters tuned to several scales, orientations and frequencies. In this work we further exploit the properties of Gabor functions to address the point selection problem.

Usually, the point selection problem, also called keypoint detection [5, 11], interest point detection[7], bottom-up saliency [2], and salient region detection [4], has been addressed in a bottom-up fashion. Salient points are selected to be distinguishable from its neighbors and have good properties for matching and invariance to common image deformations. Several types of saliency functions have been proposed in the literature: minimization of matching error [7, 11], scale-Shanon entropy [3], local maxima of Laplacian of Gaussian [5], winner-take-all of the saliency map [2].

In one of the most influential works on object recognition of the last decade, [5] has proposed the SIFT descriptors as local models of object appearance being invariant to image rotation and scale. Salient points are obtained by local maxima of difference-of-gaussian operators, in scale and space. Then at each salient point, a description of its appearance, composed by gradient histograms at multiple orientations, is matched to previously learned models.

However, in guided visual search problems, where specific objects are searched in the scene, it is convenient to incorporate object related knowledge as soon as possible in the recognition process, to reduce the amount of possible candidates. In this paper we present such an approach, where saliency computation is biased to favor object related points. The objective is to remove points very different from the model and have very few rejections of "good points". The method is based on a scale/frequency signature function, invariant to position, scale and rotation. The signature function is derived from Gabor filters, which fit nicely our Gabor based recognition methods [8]. Also, recent fast methods for Gabor filtering [1] support the feasibility of Gabor based recognition.

2 The Scale/Frequency Signature Saliency

Our biased (top-down) saliency detector is based on the "characteristic texture" of image patterns. Here, we exploit the scale and orientation invariance properties of texture, to derive novel saliency operators.

Our design is based on Gabor functions, that act as low-level oriented edge and texture discriminators and are sensitive to different frequencies and scale information.

The 2D zero mean isotropic Gabor function is:

$$g_{\theta,f,\sigma}(x,y) = \frac{e^{-\frac{x^2+y^2}{2\sigma^2}}}{2\pi\sigma^2}\left(e^{j2\pi f(x\cos(\theta)+y\sin(\theta))} - e^{-2\sigma^2 f^2\pi^2}\right) \quad (1)$$

By selectively changing each of the Gabor function parameters σ, θ and f, we can "tune" the filter to particular patterns arising in the images. By convolving the Gabor function with image patterns $I(x,y)$, we can evaluate their similarity. The Gabor response at point (x_0, y_0) is

$$G_{\theta,f,\sigma}(x_0,y_0) = \int\int I(x,y)g_{\theta,f,\sigma}(x_0-x, y_0-y)dxdy \quad (2)$$

The Gabor response of Eq. (2) can emphasize basically three types of characteristics in the image: edge-oriented characteristics, texture-oriented characteristics and a combination of both. In order to emphasize different types of image characteristics, we must vary the parameters σ, θ and f of the Gabor function. In Figure 1 we can see examples of Gabor functions with several $\gamma = 1/\sigma f$ values.

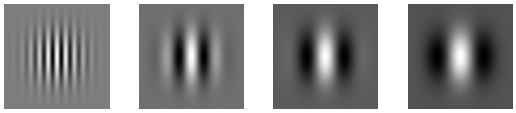

Figure 1. Real part of Gabor functions. γ values from left to right, $\{1/2, 3/2, 5/2, 7/2\}$

However, the Gabor response by itself is not invariant to scale and orientation. To enforce these properties we reinterpret the Gabor filter parameters (scale and frequency), in a joint dimensionless parameter $\gamma = 1/\sigma f$, and we derive operators from integration of Gabor kernels over the scale/frequency parameter and orientation.

In order to compute the scale/frequency signature function at an image point, we build two image representations: (i) scale-space representation and (ii) frequency-space representation. After the computation of the scale and frequency representations, every pixel in the image has two curves: the scale curve and the frequency curve(signature). The scale/frequency signature is defined as the frequency signature, endowed with a map of the frequency values into γ values. We compute the map changing the frequency values by the inverse of the frequency multiplied by the scale at which is the highest local maxima in the scale curve. Finally we explain the saliency model of an object, and how to match that model with signatures computed in new images.

2.1 Scale-Space Representation from Gabor Response

The first step in the construction of a rotation and scale invariant signature for local appearance description consists in computing the intrinsic scale of the pattern. This can be obtained by analysing the scale profile of image points. In order to compute the scale profile of image points we define the Gabor scale-space kernel:

$$GSS_{kernel}(x,y,\sigma) = \int_0^\infty \int_{-\pi}^{\pi} g_{\sigma,\theta,f}(x,y)d\theta df \quad (3)$$

At this stage we are only interested in the scale properties of the image pattern, and the integration of the Gabor function over all orientations and frequencies removes the dependency on these parameters. The closed form expression for the Gabor scale-space kernel is given by:

$$GSS_{kernel}(r,\sigma) = \frac{e^{-\frac{r^2}{2\sigma^2}}}{2\pi\sigma^2}\left(\frac{1}{r} - \frac{\sqrt{\pi/2}}{\sigma}\right) \quad (4)$$

where $r = \sqrt{x^2+y^2}$. However, due to discretization in real images, the frequency sampling in images cannot cover the whole interval $[0,\infty)$. In the case of the lower integral limit, the Gabor wavelength (inverse of frequency) should not be greater than the Gaussian envelope ($\lambda \leq 6\sigma$), so $f \geq 1/6\sigma$. For the upper integral limit, we define a scale minimum value($\sigma = 2$), and use the Nyquist sampling limit($f = 1/2$). Using the relation $\gamma = 1/\sigma f$, $\sigma = 2$ and $f = 1/2$, $\gamma = 1$. Replacing $\gamma = 1$ in $\gamma = 1/\sigma f$, the upper integral limit is $f = 1/\sigma$. With these new integral limits, the scale-space kernel is:

$$GSS_{kernel}(x,y,\sigma) = \int_{1/6\sigma}^{1/\sigma} \int_{-\pi}^{\pi} g_{\sigma,\theta,f}(x,y)d\theta df \quad (5)$$

and a closed form expression is presented in appendix. In

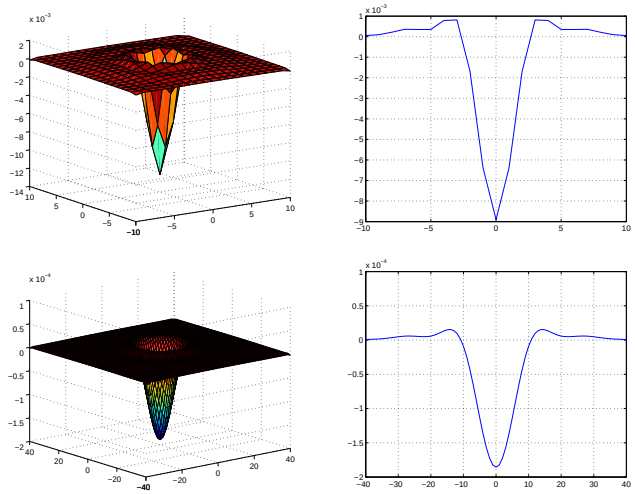

Figure 2. Example of Gabor scale-space kernel. Top figures, 3D plot and 1D slice of $GSS_{kernel}(x,y,4)$.Bottom figures, 3D plot and 1D slice of $GSS_{kernel}(x,y,16)$

Figure 2 we can see two examples of the GSS_{kernel} from

Eq.(5). The shape of the scale-space kernel is very similar to the 2D Laplacian of Gaussian. In order to build the normalized scale-space representation with the Gabor scale-space kernel, we multiply the response of the kernel by the scale, and the normalized Gabor scale-space representation of an image point (x_0, y_0), is:

$$GSS_{norm}(x_0, y_0, \sigma_0) = \sigma_0 |I(x_0, y_0) * GSS_{kernel}(x_0, y_0, \sigma_0)| \quad (6)$$

In Figure 4 we can see an example of GSS_{norm} for the case of an eye's center point.

2.2 Frequency-Space Representation from Gabor Response

To determine the texture-frequency profile of the image patterns we define the Gabor frequency-space kernel as:

$$GFS_{kernel}(x, y, f) = \int_0^{\infty} \int_{-\pi}^{\pi} g_{\sigma,\theta,f}(x, y) d\theta d\sigma \quad (7)$$

Here, the integral is performed over all scales and orientations. The closed form expression for the frequency-space

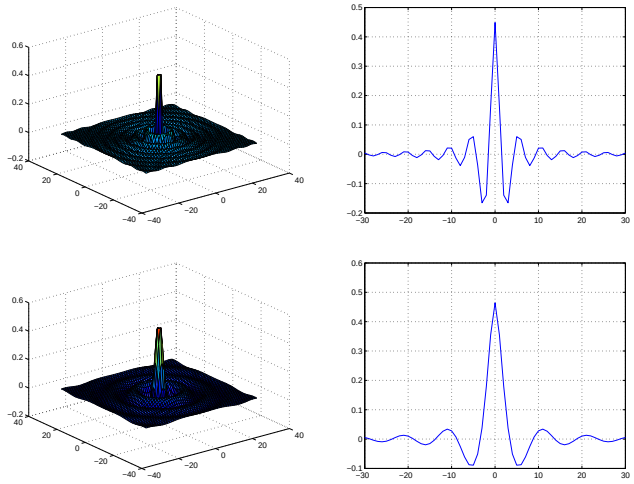

Figure 3. Example of Gabor frequency-space kernel. Top figures, 3D plot and 1D slice of $GFS_{kernel}(x, y, 0.2)$.Bottom figures, 3D plot and 1D slice of $GFS_{kernel}(x, y, 0.1)$

kernel is the following:

$$GFS_{kernel}(r, f) = \frac{\sqrt{\pi/2}}{r} \left(-e^{-2\pi f r} + J_0(2\pi f r)\right) \quad (8)$$

In Eq.(8), $r = \sqrt{x^2 + y^2}$, and $J_0(z)$ is the Bessel function of the first kind. Again, due to discretization, the scale limits are redefined to reasonable values. Changing the integral limits like in Section 2.1, the frequency-space kernel is:

$$GFS_{kernel}(x, y, f) = \int_{1/6f}^{1/f} \int_{-\pi}^{\pi} g_{\sigma,\theta,f}(x, y) d\theta d\sigma \quad (9)$$

and a closed form expression is also presented in appendix. In Figure 3 we can see an example of GFS_{kernel} from Eq.(9), its shape is an exponentially decreasing 2D Bessel function. In order to build the normalized frequency-space representation with the Gabor frequency-space kernel, we multiply the response of the kernel by the frequency, and the normalized Gabor frequency-space representation at point (x_0, y_0) is

$$GFS_{norm}(x_0, y_0, f_0) = f_0(I(x_0, y_0) * GFS_{kernel}(x_0, y_0, f_0)) \quad (10)$$

In Figure 4 we can see an example of GFS_{norm} for the case of an eye's center point.

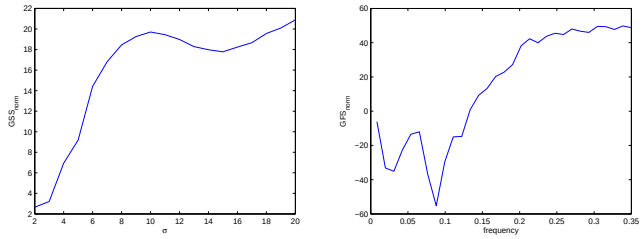

Figure 4. Example of GSS_{norm}(left side) and GFS_{norm}(right side) for the eye's center point

2.3 Scale/frequency Signature Computation

In order to compute the signature, we perform two steps: (i) Compute the frequency signature, and (ii) map the frequency interval into a γ interval. The rationale is to model the object saliency by the mean texture-frequency profile. But the texture-frequency profile itself is not scale invariant, so we map the frequency values into γ values. For the first step, let us define a set of frequency values $F = \{f_1, \ldots, f_n\}$. The frequency signature of an image point (x, y) is

$$FS_{x,y}(f_i) = GFS_{norm}(x, y, f_i), f_i \in F \quad (11)$$

To map the set of frequency values F into γ values, we look for the intrinsic scale($\hat{\sigma}$) of the image point (x, y):

$$\hat{\sigma} = \arg\max_{\sigma} GSS_{norm}(x, y, \sigma) \quad (12)$$

Looking again at the left side of Figure 4, we find that $\hat{\sigma} = 10$ in the eye example. The γ interval of signature is $\Gamma = \{\gamma_1, \ldots, \gamma_n\} = \{1/f_1\hat{\sigma}, \ldots, 1/f_n\hat{\sigma}\}$. Thus, the scale/frequency signature of an image point (x, y) is

$$SFS_{x,y}(\gamma_i) = FS_{x,y}(1/\gamma_i\hat{\sigma}), \gamma_i \in \Gamma \quad (13)$$

In Figure 6 we can see the scale/frequency signature of the eye's center point.

2.4 Top-down Saliency Model

The saliency model of an image point (x, y) is the mean value of the scale/frequency signature computed in a training set, $\overline{SFS}$. In order to compute the signature, we (i) compute the mean frequency signature, (ii) find the intrinsic scale in the mean scale-space representation, and (iii) map the frequency values into γ values. The mean frequency-signature of an image point is:

$$\overline{FS}_{x,y}(f_i) = \overline{GFS}_{norm}(x, y, f_i), f_i \in F \quad (14)$$

We can observe the $\overline{FS}_{eye}$ in the left side of Figure 5. Now we look for the intrinsic scale of the mean scale-space representation:

$$\hat{\sigma}_{object} = \arg\max_{\sigma} \overline{GSS}_{norm}(x_{object}, y_{object}, \sigma) \quad (15)$$

In the right side of Figure 5, we see the intrinsic scale of the eye located at $\hat{\sigma}_{eye} = 8$.

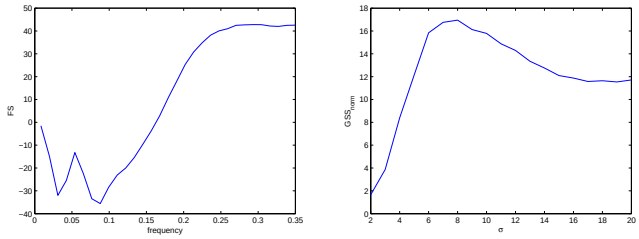

Figure 5. In the left side, $\overline{FS}_{eye}$, and in the right side $\overline{GFS}_{norm}$ for the eye's center point

The γ interval of the signature model is $\Gamma = \{\gamma_1, \ldots, \gamma_n\} = \{1/f_1\hat{\sigma}_{object}, \ldots, 1/f_n\hat{\sigma}_{object}\}$. The signature object model is:

$$\overline{SFS}_{object}(\gamma_i) = \overline{FS}_{object}(1/\gamma_i\hat{\sigma}_{object}), \gamma_i \in \Gamma \quad (16)$$

In the right side of Figure 6 we can see the mean scale/frequency signature of the eye's center point.

2.5 Matching Signatures

In order to match a signature $S_{x,y}$ with the saliency model $\overline{SFS}_{object}$, we perform the following steps: (i) Find the intersection of the γ interval between the two signatures, (ii) subsample the longest signature, (iii) translate and normalize both signatures, and (iv) compute the distance between signatures(Euclidean and earth movers distances).

Let us define two interval of γ values: $\Gamma_S = [\gamma_{i_S}, \gamma_{f_S}]$ of signature $S_{x,y}$, and $\Gamma_{\overline{SFS}} = [\gamma_{i_{\overline{SFS}}}, \gamma_{f_{\overline{SFS}}}]$ of the object model $\overline{SFS}_{object}$, where i stands for initial

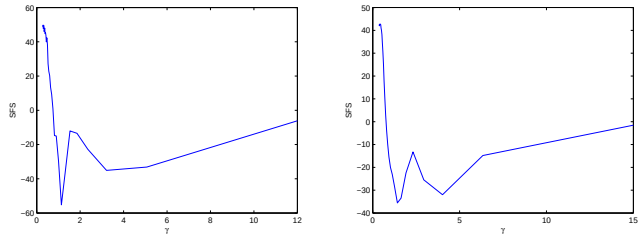

Figure 6. SFS for an eye's center point(left side), and $\overline{SFS}$ for the training set

value, and f stands for final value. The segment of the signature for computing the distance is the intersection of the two intervals, $\Gamma_S \cap \Gamma_{\overline{SFS}} = [\gamma_{i_\cap}, \gamma_{f_\cap}]$.

The number of signature elements within the interval $[\gamma_{i_\cap}, \gamma_{f_\cap}]$ could be different in $S_{x,y}$ and $\overline{SFS}_{object}$. Therefore, subsample the signature segment with more elements in $[\gamma_{i_\cap}, \gamma_{f_\cap}]$ to have equal sized signature segments. In the third step, to avoid negative values we translate both signatures by the distance

$$d = \left| \min_{S<0, \overline{SFS}<0} (S_{x,y}, \overline{SFS}_{object}) \right| \quad (17)$$

Secondly, we normalize the signature in order to achieve constrast invariance. Each signature will add up one after normalization.

In the last step, we compare two metrics when computing the distance between signatures: Euclidean distance and earth movers distance. Earth movers distance[9] reflects the minimal amount of work that must be performed to transform one signature into the other.

3 Experimental Results

We present the results of tests performed in an eye and nose point pre-selection. The tests were performed using 112 subjects from AR face database [6], where half of them are used for learning the signature object model, and the half remaining for the selective attention test. We are looking for two facial landmarks: Eyes and noses' center points.

3.1 Looking for Eye and Nose signatures

We use the training set for learning the mean scale/frequency signature, and also set the adequate distance threshold value. The threshold value is the maximum distance of the facial point signature to the mean signature in the training set. We compute the reduction of number of candidates with respect to a multi-scale Difference of Gaussians operator. So in the test images we apply first the multi-scale DoG, and keep the local maxima in every DoG image. Now we have a set of bottom-up salient points, and we compute the scale-frequency signature for

all these points. The points with distance to the model less than the threshold are the candidates for further processing. To evaluate the performance of each experiment we count the number of hits (successful detections) in the test set. Given an object part model, a distance function and an image point, a hit exists if there is a distance to the model less than the threshold inside a circle of radius r around the image point.

Table 1. List of the performed tests

Facial Point	Performance	% of bottom-up SP	distance
Eye	98.21	17.23	Euclidean
Eye	100	37.30	Earth movers
Nose	98.21	22.57	Euclidean
Nose	98.21	37.31	Earth movers

In Table 1 we can see the performance for the different set-ups, when $r = 5$. The earth movers distance has a better performance, but keep almost double of the points that the euclidean distance. However, as we mention in the introduction, we want to reduce the number of candidates without losing the object. In Figure 7 we can see an example of the points selected by the eye signature.

3.2 Scale Invariance

To check the scale invariance of the scale-frequency signature, we compute the success rate in rescaled images mantaining the signature model learned in the original images. In Figure 8 we observe that the methodology proposed mantain the performance constant with the earth movers distance, coping scale variations up to 40%(≈ 0.5 octaves).

4 Conclusions

We present a top-down saliency detector that is able to reduce the number of candidates for matching image regions. The saliency representation is based on Gabor filter response and is able to remove points different from the model, having very few rejections of interest points. In our saliency representation, every image point has a texture-frequency profile, mapped into a scale invariant and dimensionless parameter γ. The rotationally invariant shape of the filters utilized to compute the signature, and the scale invariance of γ, allow us to have a saliency detector invariant to scale transformations as well to rotation transformations.

The tests presented illustrate the variation of the performance for two different facial features. We also illustrate the scale invariance of the saliency detector. The performance varies according to the amount of information present in the region.

Figure 7. Example of points selected by the bottom-up saliency in the top image. In the middle image, points selected by the scale/frequency signature from the bottom-up saliency points . In the bottom image, points selected by our Gabor recognition method[8] as eye points, from the points selected by the scale/frequency signature.

In future work we want to automate the selection of the regions, designing a learning procedure for selecting regions where there are few or no rejections of our saliency detection method.

5 Acknowledgements

Research partly funded by European project IST 2001 37540(CAVIAR), the FCT Programa Operacional Sociedade de Informação(POSI) in the frame of QCA III, and Portuguese Foundation for Science and Technology PhD Grant FCT SFRH\BD\10573\2002

A Gabor scale-space kernel

The closed form expression of Eq.(5) is:

$$
\begin{aligned}
GSS_{kernel}(r,\sigma) = \frac{e^{-\frac{r^2}{2\sigma^2}}}{12\pi\sigma^3}[3\sqrt{2\pi}(\mathrm{erf}(\pi/3\sqrt{2}) \\
-\mathrm{erf}(\sqrt{2}\pi)) - \pi^2 J_1(\pi r/3\sigma)H_0(\pi r/3\sigma) \\
+6\pi^2 J_1(2\pi r/\sigma)H_0(2\pi r/\sigma) \\
+\pi J_0(\pi r/3\sigma)(-2+\pi H_1(\pi r/3\sigma)) \\
-6\pi_0 F_1(1;-(\pi^2 r^2)/\sigma^2)(-2+\pi H_1(2\pi r/\sigma))]
\end{aligned}
$$

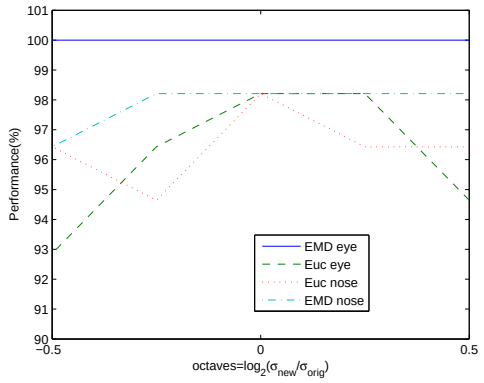

Figure 8. Scale invariance test

where $r = \sqrt{x^2 + y^2}$, $\mathrm{erf}(z)$ is the error function, $J_0(z)$ and $J_1(z)$ are Bessel functions of the first kind, $H_0(z)$ and $H_1(z)$ are Struve functions, and ${}_0F_1(a;b)$ is the regularized confluent hypergeometric function.

B Gabor frequency-space kernel

The closed form expression of Eq.(9) is:

$$
\begin{aligned}
&GFS_{kernel}(r,\sigma) = \frac{J_0(2\pi fr)\sqrt{2\pi}}{2} \\
&(\mathrm{erf}(3\sqrt{2}fr) - \mathrm{erf}(fr/\sqrt{2})) + \\
&\frac{1}{4r^2}\left[-24e^{-\pi^2/18 - 18f^2r^2} + 4e^{-2\pi^2 - 1/2f^2r^2}\right. \\
&-2\sqrt{2}e^{-2\pi fr}\pi^{3/2}\mathrm{erf}\left(\frac{\pi - 18fr}{3\sqrt{2}}\right) - \\
&\frac{1}{fr}\left(e^{-2\pi fr}(1+2\pi fr)\mathrm{erf}\left(\frac{-2\pi + fr}{\sqrt{2}}\right)\right) + \\
&2\sqrt{2}e^{2\pi fr}\pi^{3/2}\mathrm{erf}\left(\frac{2\pi + fr}{\sqrt{2}}\right) - \\
&\frac{e^{2\pi fr}\sqrt{2\pi}\mathrm{erf}\left(\frac{2\pi+fr}{\sqrt{2}}\right)}{fr} + \\
&\frac{e^{-2\pi fr}\sqrt{2\pi}\mathrm{erf}\left(\frac{-\pi+18fr}{3\sqrt{2}}\right)}{fr} + \\
&2\sqrt{2}e^{2\pi fr}\pi^{3/2}\mathrm{erf}\left(\frac{\pi + 18fr}{3\sqrt{2}}\right) + \\
&\left.\frac{e^{2\pi fr}\sqrt{2\pi}\mathrm{erf}\left(\frac{\pi+18fr}{3\sqrt{2}}\right)}{fr}\right]
\end{aligned}
$$

where $r = \sqrt{x^2 + y^2}$, $\mathrm{erf}(z)$ is the error function, $J_0(z)$ is the Bessel function of first kind.

References

[1] A. Bernardino and J. Santos-Victor. A real-time gabor primal sketch for visual attention. In *Proc. IbPRIA'05*, Estoril, Portugal, 2005.

[2] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Journal on Pattern Analysis and Machine Inteligence*, 20(11):1254–1259, November 1998.

[3] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[4] T. Kadir and M. Brady. An affine invariant salient region detector. In *European Conference on Computer Vision*, volume 1, pages 228–241, 2004.

[5] D. Lowe. Object recognition from local scale-invariant features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1150–1157, 1999.

[6] A.M. Martinez and R. Benavente. The ar face database. Technical report, CVC, June 1998.

[7] K. Mikolajcyzik and C. Schmid. An affine invariant interest point detector. In *European Conference on Computer Vision*, pages 128–142. Springer, 2002.

[8] P. Moreno, A. Bernardino, and J. Santos-Victor. Gabor parameter selection for local feature detection. In *Proc. IbPRIA'05*, Estoril, Portugal, 2005.

[9] Y. Rubner, C. Tomasi, and L. Guibas. A metric for distributions with applications to image databases. In *Proc. of IEEE International Conference on Computer Vision*, pages 59–66, 1998.

[10] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, May 1997.

[11] B. Triggs. Detecting keypoints with stable position, orientation, and scale under illumination changes. In *European Conference on Computer Vision*, volume 3, pages 100–113, 2004.

[12] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2000.