

# Where to look next and what to look for

Bernt Schiele and James L. Crowley  
GRAVIR, I.N.P.G., 46, Félix Viallet, 38031 Grenoble  
email: Bernt.Schiele@imag.fr

February 27, 1996

## Abstract

In [Sch 96c] we have introduced the use of Multidimensional Receptive Field Histograms for Probabilistic Object Recognition. In this paper we reverse the object recognition problem by asking the question, "*where should we look?*", when we want to verify the presence of an object, to track an object or to actively explore a scene. This paper describes the statistical framework from which we obtain a network of salient points for an object. This network of salient points may be used for fixation control in the context of active object recognition.

## 1 Introduction

One of the important questions for fixation control in an active vision system is signified by the phrase "*where to look next?*" [Swa 93]. One can divide the different approaches for this problem into three categories: task driven, context driven and feature driven attention mechanisms. In this paper we propose a new statistical approach to answer the question "*where to look next?*". Our approach is based on the use of multidimensional histograms of local neighborhood operators and a network of salient points. The multidimensional receptive field histograms serve as basis to generate a hypothesis for the presence of an object. The network of salient points enables the technique to determine the next region of interest to drive fixation and to verify the generated hypothesis. The network of salient points also tells us "*what to look for?*" in the new region of interest.

Section 2 reviews the multidimensional receptive field histogram approach for probabilistic object recognition. Section 3 develops a technique to obtain the network of salient points and how to use this network for fixation control in the context of active object recognition. Section 4 shows examples and the use of salient points. We also examine the robustness of the salient points to scale changes.

## 2 Probabilistic Object Recognition based on Multidimensional Receptive Field Histograms

In earlier papers we presented a technique to determine the identity of an object in a scene using multidimensional histograms of local neighborhood operators. We showed [Sch 96a] that this technique can be used to determine the most probable object, independent of its position, scale and image-plane rotation. In [Sch 96b] we evaluated the robustness of the approach to image-plane orientation and view points changes. The experimental results demonstrate that the approach is robust to such changes.

In [Sch 96c] we extended this technique to probabilistic object recognition. We developed a technique to determine the probability of each object in an image only based on multidimensional receptive field histograms. In the following section 2.1 we describe briefly the local characteristics which will be used throughout the paper. Section 2.2 reviews the use of this technique for probabilistic object recognition.

## 2.1 Local Characteristics based on Gaussian Derivatives

Multidimensional receptive field histograms can be constructed using a vector of any linear filter. In [Sch 96a] we experimentally compared the invariant properties for a number of receptive field functions, including Gabor filter and local derivative operators. Those experiments showed that Gaussian derivatives provided the most robust and equi-variant recognition results. Accordingly, in the work described in this paper we use filters which are based on equi-variant Gaussian derivatives. Gaussian derivatives permit an explicit selection of scale. This is achieved by adapting the variance  $\sigma$  of the derivative. Given the Gaussian distribution  $G(x, y)$ :

$$G(x, y) = e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1)$$

the first derivative in  $x$ - and  $y$ -direction is given by:

$$D_x(x, y) = -\frac{x}{\sigma^2}G(x, y) \quad (2)$$

$$D_y(x, y) = -\frac{y}{\sigma^2}G(x, y) \quad (3)$$

The Magnitude and Direction of the first derivative are calculated as:

$$Mag(x, y) = \sqrt{(D_x)^2 + (D_y)^2} \quad (4)$$

$$Dir(x, y) = \arctan \frac{D_y}{D_x} \quad (5)$$

The Laplace operator is calculated as:

$$G_{xx}(x, y) = \left(\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2}\right)G(x, y) \quad (6)$$

$$G_{yy}(x, y) = \left(\frac{y^2}{\sigma^4} - \frac{1}{\sigma^2}\right)G(x, y) \quad (7)$$

$$Lap(x, y) = G_{xx}(x, y) + G_{yy}(x, y) \quad (8)$$

## 2.2 Probabilistic Object Recognition

In the context of probabilistic object recognition we are interested in the calculation of the probability of the object  $O_n$  given a certain local measurement  $M_k$ . This probability  $p(O_n|M_k)$  can be calculated by the Bayes rule:

$$p(O_n|M_k) = \frac{p(M_k|O_n)p(O_n)}{p(M_k)} = \frac{p(M_k|O_n)p(O_n)}{\sum_n p(M_k|O_n)p(O_n)} \quad (9)$$

with

- $p(O_n)$  the a priori probability of the object  $O_n$ ,
- $p(M_k)$  the a priori probability of the filter output combination  $M_k$ ,
- $p(M_k|O_n)$  is the probability density function of object  $O_n$ , which differs from the histogram of an object  $O_n$  only by a normalization factor.

Having two local measurements  $M_k$  and  $M_l$  from the same object  $O_n$  we can rewrite this formula:

$$p(O_n|M_k \wedge M_l) = \frac{p(M_k \wedge M_l|O_n)p(O_n)}{\sum_n p(M_k \wedge M_l|O_n)p(O_n)} \quad (10)$$

Under the assumption of independence of  $M_k$  and  $M_l$  we obtain:

$$p(O_n|M_k \wedge M_l) = \frac{p(M_k|O_n)p(M_l|O_n)p(O_n)}{\sum_n p(M_k|O_n)p(M_l|O_n)p(O_n)} \quad (11)$$

Having  $K$  independent local measurements  $M_1, M_2, \dots, M_K$  we can calculate the probability of each object  $O_n$  by:

$$p(O_n | \bigwedge_k M_k) = \frac{p(\bigwedge_k M_k | O_n) p(O_n)}{\sum_n p(\bigwedge_k M_k | O_n) p(O_n)} \quad (12)$$

$$= \frac{\prod_k p(M_k | O_n) p(O_n)}{\sum_n \prod_k p(M_k | O_n) p(O_n)} \quad (13)$$

In our context the local measurement  $M_k$  corresponds to a single multidimensional receptive field vector. Therefore  $K$  local measurements  $M_k$  correspond to  $K$  receptive field vectors which are typically from the same region of the image. To guarantee the independence of the different local measurements we choose the minimal distance  $d(M_k, M_l)$  between two measurements  $M_k$  and  $M_l$  sufficiently large (in the experiments described below we choose the minimal distance  $d(M_k, M_l) \geq 5\sigma$ ).

For the experiments we can assume that all objects do have the same probability  $p(O_n) = \frac{1}{N}$ , where  $N$  is the number of objects. Therefore formula (13) simplifies to:

$$p(O_n | \bigwedge_k M_k) = \frac{\prod_k p(M_k | O_n)}{\sum_n \prod_k p(M_k | O_n)} \quad (14)$$

The probabilities  $p(M_k | O_n)$  are directly given by the multidimensional receptive field histograms. Therefore formula (14) shows a calculation of the probability for each object  $O_n$  only based on the multidimensional receptive field histograms of the  $N$  objects.

### 3 Active Object Recognition

In the previous section we have reviewed the use of Multidimensional Receptive Field Histograms for Probabilistic Object Recognition. In this section we reverse the object recognition problem by asking the question, "*where should we look?*", when we want to verify the presence of an object, to track an object or to actively explore a scene. This section describes a method for deriving a network of salient points for an object which are characteristic of an object. These salient points are literally the points on the object which are the most unique. These points maximize the discrimination between objects. A network of salient points can be used to control fixation in the context of object recognition.

#### 3.1 Network of salient points

This first part of the section introduces a concept to determine the most significant points in an image and/or of an object. We use this concept to determine a network of salient points for a set of objects and use these networks in the context of fixation control for object recognition (see section 3.2). It is worth mentioning that this concept can generally be used as an *interest point detector*. Such interest points of an object or in an image can be e.g. used directly in a geometric hashing approach for object recognition [Wol 90]. We also use the *interest point detector* to find track-able (that is discriminable and easy to detect) points of an object.

As we described in section 2.2 we can calculate the probability for an object  $O_n$ , given the local measurement  $M_k$ :

$$p(O_n | M_k) = \frac{p(M_k | O_n) p(O_n)}{p(M_k)} \quad (15)$$

with

- $p(O_n)$  the a priori probability of the object  $O_n$ ,
- $p(M_k)$  the a priori probability of the filter output combination  $M_k$ ,
- $p(M_k | O_n)$  the probability density function of object  $O_n$ , which differs from the histogram of local measurements of an object  $O_n$  only by a normalization factor.

In this section we are interested in the most significant points for a certain object  $O_n$  (or in a given image). These points can be obtained by maximizing  $p(O_n|M_k)$  over all filter output  $M_k$  (of the object  $O_n$ ). Therefore we maximize the relation between  $p(O_n|M_k)$  and  $p(M_k)$  (as we will see later, we can neglect  $p(O_n)$ , since it is not important for the order of the maxima). This results in highly discriminant and, as the examples show, easy to detect filter combinations.

Such salient points (or interest points) can be used in many ways: Since these points are the most discriminate points one can use them in the context of active vision (as e.g. for tracking or fixation control (section 3.2)). In the context of object recognition one can use the first  $K$  maxima in order to minimize the number of points needed to determine the identity of an object. Such  $K$  maxima define a network of salient points, which we will use in the following section in the context of an active recognition system.

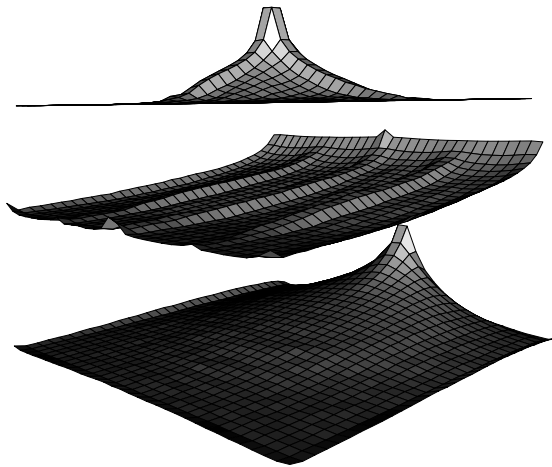


Figure 1: Average Histograms for different filter pairs: top  $Dx-Dy$ , middle  $Mag-Dir$  and bottom  $Mag-Lap$  ( $\sigma = 1.0$ , number of bins per histogram axis is 32).

In order to use formula (15), one has to determine the a priori probability  $p(M_k)$  of each local filter output  $M_k$ . This probability distribution depends not only on the filter combination which is used but also depends on the context of the vision task. Therefore the calculation of the probability distribution  $p(M_k)$  needs careful investigation.

In [Sch 96c] we used the formula  $p(M_k) = \sum_n p(M_k|O_n)p(O_n)$  to calculate the a priori probability of the measurement  $M_k$ . This formula works well in the context of probabilistic object recognition, but depends on the given database and backgrounds. Such a dependency can be justified if one knows the environment of the visual task in advance. This is often the case in a tracking task, where the camera is stationary. This article deals with the more general case, where we do not know all objects/backgrounds in advance. Therefore we have to approximate or estimate the *real* a priori probability  $p(M_k)$ . We propose to calculate the average histogram of a sufficiently large representative image database. Figure 1 shows three such average 2D-histograms for the filter combinations  $Dx-Dy$ ,  $Mag-Dir$  and  $Mag-Lap$ , which have been calculated on a database of 832 images. Even though this average histogram still depends on the database which is used, it seems to be more reliable than a purely analytical approximation.

Using an average histogram for the approximation of the a priori probability  $p(M_k)$  we can rewrite formula (15):

$$p(O_n|M_k) = \frac{p(M_k|O_n)p(O_n)}{p(M_k|O_n)p(O_n) + p(M_k|Average)(1 - p(O_n))} \quad (16)$$

Since in this later formula the real value of the probability  $p(O_n)$  does not affect the order of the maxima of  $p(O_n|M_k)$  we use the following formula to determine the most salient points of the object  $O_n$  (which corresponds to  $p(O_n) = 0.5$ ):

$$p^*(O_n|M_k) = \frac{h(M_k|O_n)}{h(M_k|O_n) + h(M_k|Average)} \quad (17)$$

In this formula  $h(M_k|O_n)$  is the histogram value of the measurement  $M_k$  for the object  $O_n$  and  $h(M_k|Average)$  is the histogram value of the measurement  $M_k$  in the average histogram.

By using the first  $K$  maxima of formula (17), one obtains not only a network of salient points but also a quantitative measure for the significance of each point (in particular  $p^*(O_n|M_k)$ ). This measure can be used to determine if a certain object contains points or regions which are sufficiently discriminant. The higher the value  $p^*(O_n|M_k)$  of the salient points the more they are discriminant. In section 4 we will use these values to determine objects in a database which contain the most discriminant and the objects with the less discriminate points.

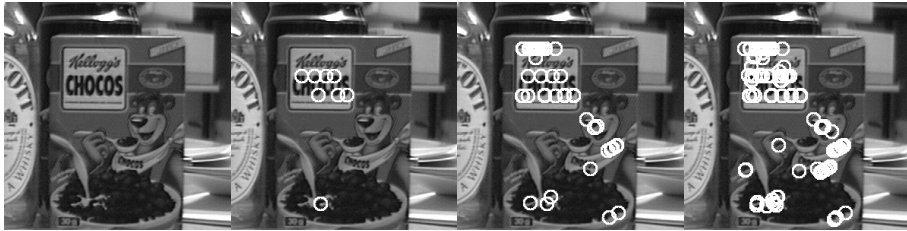


Figure 2: An object and its first 8, 50 and 103 salient points

Figure 2 shows an example of the first salient points of an object (marked with circles). Table 1 shows the corresponding filter output. Here we used the 3D histogram  $Dx-Dy-Lap$  with  $\sigma = 1.0$ . The size of the circles corresponds approximately to the support of the Gaussian filters. The maxima in table 1 correspond to high filter output values, which makes them not only discriminant but also easy to detect in a an image. They correspond to a very low value of  $h(M_k|Average)$  (the first maxima e.g. occurs in average only 0.004 times per image). This results in a very high value for  $p^*(O_n|M_k)$ .

Number of Maxima	$Dx$	$Dy$	$Lap$	$h(M_k O_n)$	$h(M_k Average)$	$p^*(O_n M_k)$
1	34.2	49.0	-99.0	1	0.0445	0.9574
2	-19.4	-50.5	56.8	4	0.2067	0.9509
3	-27.3	-49.0	63.7			
4	-26.6	-48.7	54.4			
5	-25.5	-53.1	62.7			
6	36.0	50.8	6.2	3	0.1635	0.9483
7	35.8	50.7	7.8			
8	40.1	48.2	3.3			
...						
103	16.1	43.8	16.0	19	1.2848	0.9367

Table 1: The filter output values (scaled into the interval  $[-128, +128]$ ) for the first salient points of the object in figure 2, the values of the histogram of the object  $h(M_k|O_n)$ , the average histogram  $h(M_k|Average)$  and the calculated value  $p^*(O_n|M_k)$

In many cases we obtain cluster of salient points (as e.g. in figure 2) so that it seems to be convenient to define a network of salient regions rather than a network of salient points. The concept for fixation control described below can be easily extended to such a network of salient regions.

### 3.2 Fixation control for object recognition

The general scheme which we propose for active fixation control in the context of object recognition can be divided in two major steps:

**Hypothesis generation:** Based on the probabilistic object recognition algorithm introduced above we generate hypothesis for the presence of an object at a certain scale and rotation. After a local scan for a salient point of the best hypothesis objects, we also obtain a hypothesis of the objects position.

**Verification:** In order to verify the generated hypothesis, one can use the *network of salient points* of this object to determine the next region of interest (*where to look next*). Having this network of salient points we also know *what to look for*.

*Hypothesis generation* is mainly based on the described probabilistic object recognition algorithm. The attractive point in using this algorithm is to be able to calculate probabilities for each object at an arbitrary image location. This means we can point the initial region of interest to any part of the image to obtain a hypothesis about the presence of an object. Since we assume that the objects are in the database at different scales and rotations we obtain not only the most probable objects but also there most probable scale and rotation.

By choosing the most probable objects we can scan locally the actual region of interest in order to find a salient point of one of the objects. Doing so one can incorporate the calculated local characteristics to get more evidence for the chosen objects. Once a salient point of one of the most probable objects is found we hypothesize this object. The position of the salient point in the corresponding network of salient points is used to determine the approximate position of the object in the image.

For *verification* of the generated hypothesis (which contains information about the most probable object with its scale, rotation and position in the image) we want to choose the next region of interest in the image, which is the most suitable for the verification step. Using the network of salient points of the hypothesized object one can choose the most significant (or the next most significant) salient point in the network to determine the next region of interest (*where to look next*). Since we know the vector of local characteristics of this salient point we also know *what to look for* at the new region of interest.

Finding the expected salient point ,we can incorporate one or several verification steps until the evidence of the object is sufficient for *confirmation* of the hypothesis. If we cannot find the expected salient point in the new region of interest, we can incorporate the observed local characteristics to calculate and/or modify the probabilities of each object (*recover*).

## 4 Examples for illustration

In this section we want to apply the detector of salient operator to a database of 30 objects. The database of 30 objects is shown in figure 3. Throughout the section we are using a 6-dimensional histogram, namely the filter combination *Dx-Dy-Lap* at two different scales ( $\sigma = 1.0$  and  $2.0$ ). By using the first 30 maxima of  $p^*(O_n|M_k)$  (see equation (17)) we obtain a network of salient points for each object. In the first part of the section we determine the objects with the most discriminant and with the less discriminant salient points. In a second part we examine the robustness of the salient point operator to scale changes.

As mentioned in section 3.1 we can use the value of  $p^*(O_n|M_k)$  to determine the objects of the database which contain the most discriminant points. The first row of figure 4 shows the 5 objects which contain the most discriminant points. The second row of this figure shows the 20 most discriminate point of each object. In figure 5 we show the 5 objects of the database which contain the least discriminate points, corresponding to the worst case for the technique. The second row shows the 20 most salient points for these objects.

In order to examine the robustness of the salient points against scale we used one of objects with the most discriminant points (figure 4) and one of the objects with the less discriminant points (figure 5). Figure 6 shows the result of the same detector of salient points for three different scales (of each object). The approximate scale between each image is 20%. One can see that the most of the salient points are the same, which shows the robustness of the approach to scale changes.



Figure 3: Image database of 30 objects

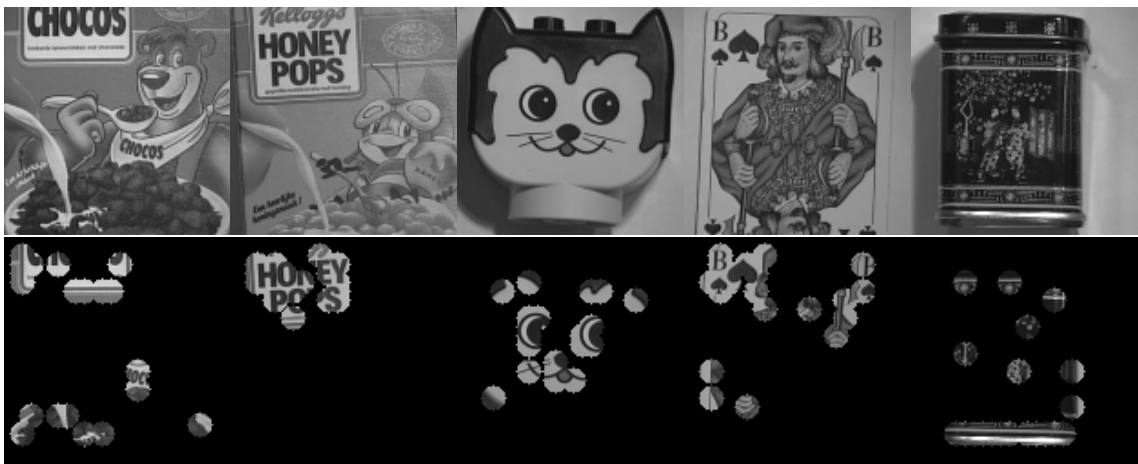


Figure 4: The first row shows the objects which contain the most discriminate points. The second row shows the 20 most-salient points of these objects



Figure 5: The 5 objects which contain the least discriminate points. The second row shows 20 most salient points of these objects. This represents a worst case for the technique.

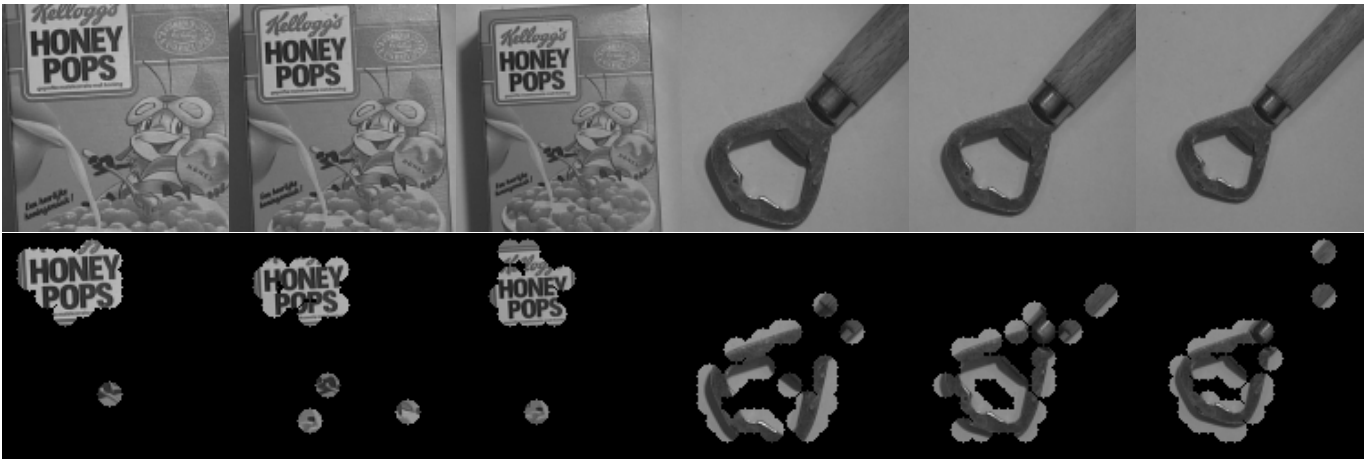


Figure 6: Robustness of the salient point detector to scale. The scale between each image is approximately 20%

## 5 Conclusion and perspective

Under suitable control of lighting, the color vector at a pixel can be an important feature for discriminating objects [Swa 91]. A histogram of color vectors provides a statistically sound manner to fuse such features and improve the probability of recognition. Normalising the color information provides a manner to obtain recognition which is robust in the presence of changes of background light conditions.

Vectors of receptive fields can be used in a manner which is similar, but richer, than color. Such vectors can be formed to include information from different orientations and scales, making them much more discriminant. Individual vectors provide important features for object discrimination. Suitable normalisation render such discrimination robust to noise and to changes in viewpoint. Histograms of such vectors provide a mathematically sound method to estimate the probability of finding an object at a point in a scene. In this paper, we have shown how such histograms can be used to deduce the feature vectors which are maximally discriminant for each object within a set. Networks of such discriminant points can be used to determine "where to look next?" (and "what to look for").

## References

- [Sch 96a] B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In *ECCV'96, Fourth European Conference on Computer Vision*, 14-16 April 1996.
- [Sch 96b] B. Schiele and J. L. Crowley. The robustness of object recognition to rotation using multidimensional receptive field histograms. submitted. available via [www<sup>1</sup>](http://www.pandora.imag.fr/Prima/schiele/), 1996.
- [Sch 96c] B. Schiele and J. L. Crowley. Probabilistic object recognition using multidimensional receptive field histograms. submitted to ICPR'96, August 1996.
- [Swa 91] M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1): 11-32, 1991.
- [Swa 93] M.J. Swain and M.A. Stricker. Promising directions in active vision. *International Journal of Computer Vision*, 11(2): 109-126, 1993.
- [Wol 90] H.J. Wolfson. Model-based object recognition by geometric hashing. In *ECCV'90, First European Conference on Computer Vision*, pages 526-536, 1990.

---

<sup>1</sup><http://pandora.imag.fr/Prima/schiele/>