A Novel Local Invariant Descriptor Adapted to Mobile Robot Vision

Guangzhi Cao, Jiaqian Chen, and Jingping Jiang, *Senior Member, IEEE*

***Abstract*—In the past years, recognition algorithms based on local features have received much attention; and their advantages over traditional recognition methods in dealing with similarity transformation, partial occlusion and extraneous noise, have been verified. But due to the difficulty of most current local-feature-based methods in realtime implementation, their application in mobile robot vision has been deferred. Motivated by it, this paper is expected to provide a vision system especially adapted to mobile robots based on local features. Compared with previous works, in this paper a novel local invariant descriptor combined with gradient information is provided; and accordingly a new matching method of interest points is introduced. The novel local invariant, which is characterized by Gaussian derivatives, also consists of descriptions of some non-interest points localized according to the average gradient at interest points, so that the novel local invariant descriptor achieves a high discrimination under various similarity transformation. And the new matching method greatly increases the robustness of the algorithm by employing a segmentation correlation and eliminating some possibly wrong matching results. Experimental results demonstrate our algorithm has a good recognition ability and robustness in case of rotation, partial occlusion, various similarity transformations, extraneous noise, etc. And it can be implemented in real time, which makes it very appropriate for mobile robot use.**

I. Introduction

A. Recognition Methods

Although laser finder has shown good performance in indoor robot navigation, a vision system is in a great need for mobile robot to finish some sophisticated tasks, such as landmark recognition, mail delivering, button pushing, etc. As it is well known, recognition algorithms play an important role in a robot vision system. So far, a lot of object recognition methods have been provided in the past decades, which can mainly be divided into two groups: those based on geometric models of an object [1][2] and those based on the appearance feature of an object [3]. Because it is often difficult to obtain geometric models for objects in the real world, the appearance-based methods are usually much preferred for a vision system which is aimed at practical applications. And the appearance-based algorithms such as histogram methods [4][5][6], eigenspace methods [7][8] and so on, have been successfully developed. However, since all the appearance-based approaches are global, they usually have difficulty in dealing with partial occlusion and extraneous noise. Therefore, recently local features have been receiving more and more attention because of its superior character in describing local image patches and dealing with extraneous noise. And the work based on local invariants by Schmid and Mohr [9] has been pointed out in [3] as one of the most successful object recognition systems so far, which was developed to address the problem of retrieving images from large image databases.

B. Robot Vision Based on Local Features

Since a robot vision system is required to deal with partial occlusion and other local disturbances, the introduction of local features is quite favorable and promising. Based on the image retrieval system in [9], Baerveldt [10] developed an object verification and localization system. Different to [9], in [10] the author introduced the local characterizations of extra points around the interest points but with lower orders to increase the discrimination. By this modification, the computation of the local descriptor is simplified, and meanwhile it does show good performance. However, this system is unable to recognize rotated objects due to the uninformed positions of the extra introduced points. Obviously, it is a drawback since object rotation commonly exists. With regard to the merits and defects of previous works, this paper is mainly aimed to develop a local-feature-based recognition algorithm appropriate for mobile robot use with good performance and robustness under various view conditions, such as partial occlusion, similarity transformations, rotation, extraneous noise and minor viewpoint variations. And the improvements are achieved mainly by a novel local invariant and a corresponding new matching method which will be detailed in section 2. Experimental results convincingly confirmed the effectiveness of this improved algorithm.

Manuscript received on September 22, 2003.

Guangzhi Cao is with the College of Electrical Engineering, Zhejiang Uuniversity, Hangzhou, CO 310027 P.R. China (phone: 86-571-87994809; e-mail: guanzhicao@hotmail.com).

Jiaqian Chen and Jingping Jiang are also with the College of Electrical Engineering, Zhejiang Uuniversity, Hangzhou, CO 310027 P.R. China (e-mail: streetren@hotmail.com; eejiang@dial.zju.edu.cn).

II. OUR APPROACH

Due to their respective weakness of appearance-based and geometric methods in recognition as mentioned above, we introduce local features in our mobile robot vision. Similar to most methods based on local features, our algorithm can be briefly described as follows. First, we extract interest points, i.e. points with high information content. Then a small neighborhood around the interest points is described by with a novel local invariant descriptor. The object model, consisting of a list of interest points and their local invariants, is then matched with the interest points extracted from the scene image on a point to point basis. Finally, some geometric constraints are applied to eliminate the false matching and refine the matching results. The specific steps of our algorithm are described below in detail together with the changes and improvements compared with the previous works [9][10][11].

A. Interest Points

Obviously, computing an image descriptor for each pixel in the image creates too much information. Interest points are those points in the image which contain a high degree of image information, such as corners and junctions. The signal usually changes two-dimensionally at interest points, so the derivatives are usually high at these points.

A wide variety of detectors for automatic extraction of interest points have been presented in [12], and the comparative performance of different detectors is provided in [13]. In terms of object recognition, the most important quality of detectors is repeatability. A detector called SUSAN [14] has been proved to be one of the best detectors, and it is also faster than most other detectors. That's why it was employed in our application. The basic idea of SUSAN detector is that each image point has associated with it a local area of similar brightness. As simply illustrated in Fig. 1, by counting the number of pixels that have almost the same grayvlaue with the nucleus in the circle, the nucleus is decided whether it is an interest point or not.

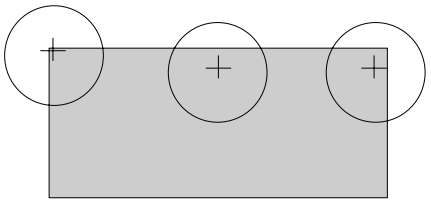

Fig. 1 Principle of SUSAN detector

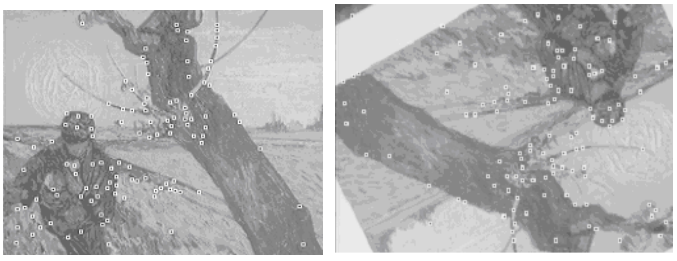

Fig. 2 Interest points of a painting detected by SUSAN
Fig. 3 Interest points of the same painting in Fig. 2 but rotated

Fig. 2 and Fig. 3 show interest points detected on the same painting but under rotation. We can find most of the interest points are repeated. What's more, SUSAN detector is also robust against all disturbance factors such as scale and extraneous noise.

B. A Novel Local Invariant Descriptor

In order to match the interest points of an object model and of a scene image, we need a discriminative local description of each interest point, which was proved to be the key step of the algorithm during the experiments. The quality of the local descriptor will have a great influence in the final performance of the algorithm. Different descriptors have been presented in the literature, such as Gabor filters [3], wavelets [15], Gaussian derivatives [3][9][10][11] and etc. Among all of them, Gaussian derivatives have been frequently used due to their desirable properties, such as robustness to minor scale changes (up to 20%), possible extension to obtain a rotational invariant and a mutliscale approach when a larger scale variation is expected. Therefore, we also employed Gaussian derivatives in our approach.

The image in a neighborhood of a point can be described by the set of its derivatives [16]. But the computation of the derivatives is usually ill-conditioned as it lacks robustness due to the presence of noise in the image. Simply, consider the functions $f(x)$ and $F(x)=f(x)+\varepsilon\sin(\omega x)$. $F(x)$ and $f(x)$ are very close for a small ε, but $f'(x)$ can be very different from $F'(x)$ if ω is big. So a high frequency noise can significantly modify the first order derivative and the higher order derivatives even more. The stable computation of derivatives is achieved by convolution with Gaussian derivatives [17]. Such a set of derivatives is named "local jet" in [16] and defined as follows:

Let I be an image and σ a given scale. The "local jet" of order N at a point $\mathbf{x}=(x,y)$ is defined by

$$J^N[I](\mathbf{x},\sigma)=\{L_{i_1\ldots i_n}(\mathbf{x},\sigma)\,|\,(\mathbf{x},\sigma)\in I\times\Re^+;n=0,\ldots,N\}, \qquad (2\text{-}1)$$

where $L_{i_1\ldots i_n}(\mathbf{x},\sigma)$ is the convolution of image I with the Gaussian derivatives $G_{i_1\ldots i_n}(\mathbf{x},\sigma)=\frac{\partial^n}{\partial_{i_1}\ldots\partial_{i_n}}G(\mathbf{x},\sigma)$, i.e. $L_{i_1\ldots i_n}(\mathbf{x},\sigma)=(G_{i_1\ldots i_n}*I)(\mathbf{x},\sigma)$, $i_k\in\{x,y\}$.

The σ of the Gaussian function determines the quality of smoothing. It also coincides with a definition of scale-space, which we will later find is very important for our multiscale approach.

Now let us discuss our novel local invariant descriptor

based on local jet specifically. A complete set of differential invariants can be computed to locally characterize a signal, which has been theoretically studied in [16] and [17]. The set of differential invariants used in our experiments is limited to 2nd order because the higher order derivatives will introduce a large computational load inappropriate for real-time implementation and they are also sensitive to high frequency noise. The differential invariants are stacked in a vector denoted by V which is given in Einstein summation convention as shown in (2-2). We can note that the descriptor V is rotation invariant and its first component represents the average luminance, the second component the square of the gradient magnitude and the fourth the Laplacian.

$$V[0...4] = \begin{bmatrix} L \\ L_i L_i \\ L_i L_{ij} L_j \\ L_{ii} \\ L_{ij} L_{ji} \end{bmatrix} = \begin{bmatrix} L \\ L_x L_x + L_y L_y \\ L_{xx} L_x L_x + 2L_{xy} L_x L_y + L_{yy} L_y L_y \\ L_{xx} + L_{yy} \\ L_{xx} L_{xx} + 2L_{xy} L_{xy} + L_{yy} L_{yy} \end{bmatrix}, \quad (2\text{-}2)$$

where L_i is the element of the local jet.

As pointed in [10] and [11], such a local description was usually not discriminative enough. In [11], the solution was to apply the filters at multiple scales. This, however, led to a relatively large local region around the interest points; and it is undesired in our application where robustness against partial occlusion and other local disturbances is crucial. And in [10], the solution was to include the local jets of some other points near the interest points in the local descriptor as well. This did make the descriptor more discriminative, but lost invariance to rotated objects because the extra points could not be correctly localized when rotation happens in a scene, which greatly degrades its practicability. It is also a potential drawback we want to avoid. To make the descriptor discriminative enough, the local derivative information of some extra points near the interest points is also introduced here. But instead of uninformed choice of the extra points like in [10], in our algorithm these non-interest points are localized based on the gradient information at the interest points. More specifically, we propose to use four extra points equally distributed at the gradient orientation and its perpendicular direction on a circle with its center placed at the interest points, as illustrated in Fig. 4. Thus a composed jet, which consists of 25 elements, is formed to describe the local image patch around the interest points. So V in (2-2) should be replaced by $V\{i\}$ ($i=0,1,2,3,4$) where $V\{0\}$ represents the previous local invariant at the interest points. Since the gradient is invariant to placement, rotation and scale changes, this descriptor is also guaranteed to be invariant to these changes. (But it is necessary to make some amendment in order to obtain invariance to scale changes; see subsection *E*) With

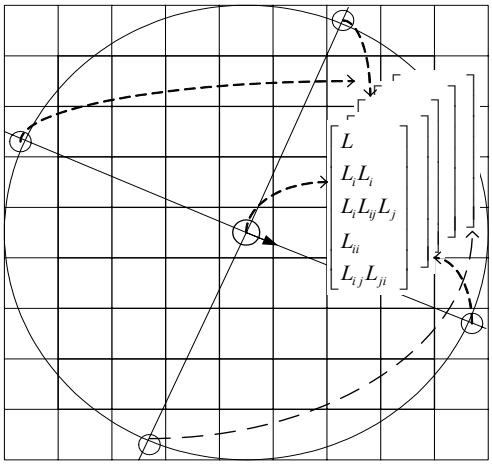

Fig. 4 The novel local invariant (the arrow denotes the gradient orientation at the interest point)

such a characterization, the discrimination of the local invariant descriptor is greatly increased without losing any favourable properties of the Gaussian derivatives. To make the gradient computation more stable, in our experiments the average gradient in a small neighborhood of an interest point was used instead of the specific gradient at the interest point.

What's more, we know there are two ways for Gaussian derivatives to obtain a rotation invariant descriptor: one is to produce a rotation invariant template, and the other is to steer the derivatives to a special direction. Here we chose the first one because of its less computational complexity but slight inefficiency compared with the latter, which is favourable for hardware implementation. And during our tests, we have tried to decrease the dimension of the descriptor by excluding some elements from the local descriptor $V\{i\}$, but it was proved not to be worthwhile due to certain degradation in the discrimination of the descriptor. Therefore, we can see this novel local invariant is a tradeoff between discrimination, robustness and computation cost.

C. Matching of Local Descriptors

After obtaining the local descriptor of interest points, the object model is matched with the scene image on a point to point basis. A robust matching method is very helpful to eliminate extraneous noise and discriminate between different interest points. And here we provide a new matching method adapted correspondingly to the novel local invariant descriptor introduced above.

In [9], the Mahalanobis distance is used as the matching criterion, but the computation is prohibitively complex for realtime implementation in despite of its good precision. So usually the normalized dot-production (or correlation) is used to represent the matching degree in real applications because the dot-production operation can be efficiently implemented using convolution by hardware. But as we

found, it was undesired and also unreasonable to directly compute as usual the correlation of two composed sets which here consist of 5 different points, 25 elements, respectively, because these 5 points are uncorrelated and the intensities of their derivatives can be very different from each other. Considering it, in our experiments the composed sets are segmented so that the five points are matched one by one according to their location with regard to the interest points using the gradient information, and the average correlation value is taken as their similarity criterion.

What's more, we note that sometimes instability of the gradient information is inevitable and even the gradients at two close points can be very different where an abrupt grayvalue change happens. Taking it into consideration, we try to make the matching more robust by excluding the worst matching result of the four non-interest neighbor points (generally, the interest point is relatively stable due to its speciality as we know above). The specific matching strategy can be described by (2-3).

$$\begin{cases} d_i = \dfrac{V\{i\} \cdot V'\{i\}}{\|V\{i\}\| \|V'\{i\}\|}, \quad i = 0,1,2,3,4; \\ d_{final} = \dfrac{d_0 + \sum_{i=1...4} d_i - \min(d_i)}{4}; \quad i = 1,2,3,4 \end{cases}, \tag{2-3}$$

where $V\{0\}$ denotes the differential invariants of interest points and $V\{1...4\}$ denote the differential invariants of the other four non-interest points, respectively. From (2-3), we can see the worst matched one among the four non-interest points is excluded from the computation of the final matching result d_{final}.

Here we present an example to show the effectiveness of our matching method. Fig. 5 gives an object and a scene image where the object is rotated by 152 degree, which almost represents the most difficult case for recognition. Fig. 6 provides the novel local descriptors of the two corresponding interest points marked in Fig. 5. In Fig. 7, the correlation histogram by the usual correlation method and the histogram after the exclusion of the worst matched points are provided. As it is shown, this routine makes the mean correlation value bigger, but the discrimination keeps high. There are few points whose correlation values are higher than 0.9 in both cases, and the number even decreased in the second one though it does not always happen. And by each of the methods, the highest correlation value is achieved by the real corresponding one indicated by the small circle in the scene image in Fig.5, and here the value by the new method is even slightly higher than the one by the usual method. So this example effectively demonstrated the robustness and discrimination of the new matching method though it was obtained at some certain expense of the mean correlation value.

More specifically, for each local invariant at an interest

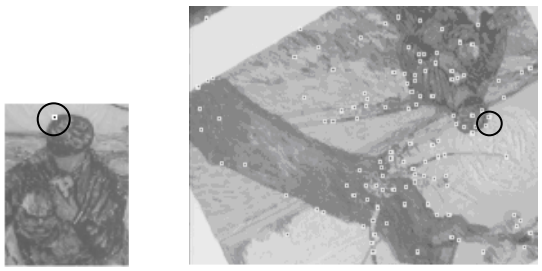

Fig.5 A interest point of a model and the corresponding one in a scene image (indicated by a small circle)

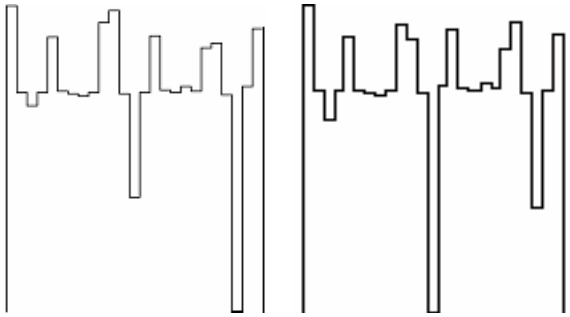

Fig. 6 The normalized local invariants of the two corresponding interest points in Fig. 5

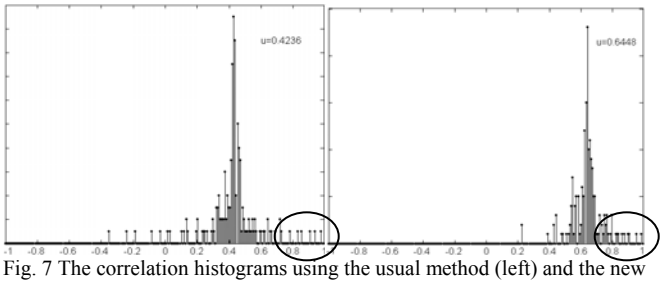

Fig. 7 The correlation histograms using the usual method (left) and the new method (right)

point **x** of the model, the local invariant at an interest point **x'** in the scene image is determined such that $d_{final}(V(\mathbf{x}, \sigma), V(\mathbf{x}', \sigma))$ is minimal provided that the correlation exceeds 0.9. The process is then reversed. Thus we obtain two lists of matched points. Finally, only those points which choose each other mutually are kept.

D. Geometric Constraints

When an object lies in a very complex background or there are many similar ones in a scene image, there is a probability that some interest points of the object may be close to several points in the scene and it may lead to false matching. Our tests confirmed that even a good matching of local invariants in last subsection was sometimes insufficient to discriminate many similar points. With regard to it, some researchers [16][18] suggested using longer vectors to decrease this probability. But the use of higher

order derivatives is not practical for our application. We don't either employ global features since they are usually sensitive to extraneous features and partial occlusion, and also time-consuming. Instead, we prefer to take the local geometric constraints used in [19] and [9] but in a simplified version. The basic idea is that the local shape configurations will keep consistent under a similarity transformation and scale change. More specifically, after the matching step in last subsection, we compare the angles between an interest point and its *p* nearest interest points as simply demonstrated in Fig. 8. If half of the neighbor interest points passed the angle test (i.e. α_1 approximates α_2 in Fig. 8), then the matching is confirmed, or else it will be rejected. With such a constraint, the recognition reliability of the algorithm was definitely increased.

E. Multiscale Approach

For a large scale change, it can equivalent to a scale change of the Gaussian derivatives; and this makes a multiscale approach possible [20][9][10]. For two images I_1 and I_2 where I_2 is only changed by a scale factor α, i.e. $I_1(x) = I_2(u) = I_2(\alpha x)$, there exists

$$\int_{-\infty}^{+\infty} I_1(x) G_{i_1 \ldots i_n}(x,\sigma)\,dx = \alpha^n \int_{-\infty}^{+\infty} I_2(u) G_{i_1 \ldots i_n}(u,\sigma\alpha)\,du\,, \qquad (2\text{-}4)$$

where $G_{i_1 \ldots i_n}$ is the derivatives of the Gaussian. So by adjusting the scale of the Gaussian derivatives σ proportionally to the image scale factor α, we can obtain a local invariant descriptor at different scales. Since the multiscale invariants have been well addressed in the previous literature, we hereby don't discuss it in detail any more. What has to be mentioned is that it is necessary for our novel local invariant also to accordingly adjust the radius of the circle on which the four non-interest points lie, so that the composed local descriptors can keep invariant at different scales.

III. Experimental Results and Conclusion

To demonstrate the performance of the algorithm, experiments have been conducted to test its recognition rate. Some of the images in the experiments are derived from the literature so that we can have an intuitive comparison about their difference and respective performance. And the rest images taken by ourselves are usually daily objects, which are typical for robot vision tests, such as landmark, extinguisher, lamp, etc. The algorithm has shown a high recognition ability and good robustness to object rotation, scale change, minor viewpoint variation, partial occlusion and extraneous noise.

In the experiments, we first applied a Gaussian filter (σ = 1) to smooth the scene image before detection of interest points and gradient computation. Because the multiplication of derivatives, which is necessary to obtain the rotational invariance, increases the instability of the local descriptor, a relatively large σ is needed for Gaussian derivatives calculation; and we set σ = 2 during the tests. Although a larger σ might make the multiplication in the descriptor more stable, it would also decrease the discrimination severely. For the same reason, the matching rate of interest points is usually relatively low where there exist rotated objects. The whole computation time is about 0.3 second by a Pentium III 800MHz processor, which can nearly satisfy realtime requirements.

During the experiments, we found that a matching rate of interest points higher than 30 percent could lead to a reliable recognition result. It can be explained by the good discrimination of the local invariant. The algorithm showed a very strong recognition ability during the tests whenever the objects were partially visible, in a complex background, or with a minor viewpoint variation. Only in a few cases where a scene image contained very little information of an object, the algorithm might fail to recognize the object correctly. Some typical experimental results are provided below with all the initially extracted interest points marked by white squares and the final matched ones marked by black squares.

Fig. 9 gives an example of a 2-D painting where the object was rotated and a small part was lost. A 62 percent matching rate was finally obtained for it. In Fig.10, the reading lamp was in a very complex background, part of it was covered and there also existed a view variation of about 15 degree. It finally resulted in a 46 percent matching rate. In Fig. 11, the extinguisher was taken at different distances in order to provide a scale change (about 1.5). It also underwent a small

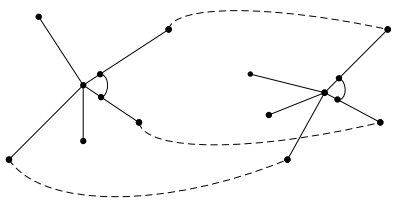

Fig. 8 Geometric constraint based on angle tests

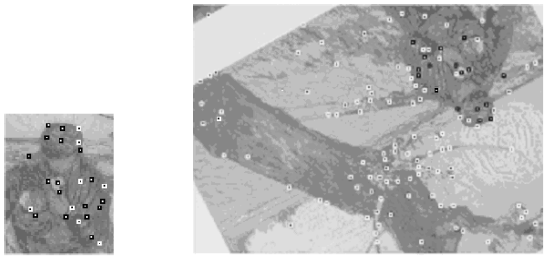

Fig. 9 A painting rotated by 152° and part lost

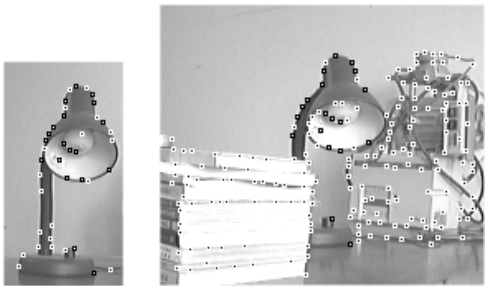

Fig. 10 A reading lamp in a very complex background, partially occluded and with a view variation of about 15°

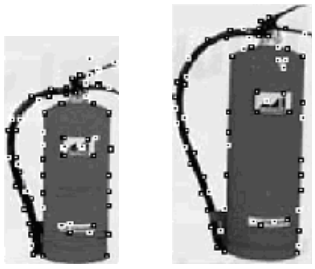

Fig. 11 An extinguisher at different scales, with a small view variation and low resolution

view variation and relatively low resolution, which finally resulted in a 67 percent matching rate.

From above, we can see the novel local invariant descriptor shows very good performance and robustness in recognition. And its application in mobile robot vision will necessarily be very promising because of its special quality. But there are still much work remained, such as to further increase the discrimination of the local descriptor, make it more stable in various view conditions and facilitate its realtime implementation. And more extensive tests should be conducted to verify its recognition ability in different conditions. We are also interested to implement the algorithm in a real mobile robot in order to see its applicability and precision in realtime localization and recognition. Some global methods can possibly be combined, for example, to deal with the situation that several objects exist in one scene image.

REFERENCES

[1] P.J. Besl and R.C. Jain, "Three-Dimensional Object Recognition," *ACM Computing Surveys*, vol. 17, no. 1, 1985, pp. 75–145.

[2] R.T. Chin, H. Smith, and S.C. Fralick, "Model-Based Recognitionin Robot Vision," *ACM Computing Surveys,* vol. 18, no. 1, 1986, pp. 67–108.

[3] B. Schiele, "Object recognition using multidimensional receptive field histograms", *Ph.D. Thesis*, Institut National Polytechnique de Grenoble, Grenoble, July, 1997.

[4] M.J. Swain and D.H. Ballard, "Color Indexing," *Int'l J. Computer Vision,* vol. 7, no. 1, 1991, pp. 11–32.

[5] B.V. Funt and G.D. Finlayson, "Color Constant Color Indexing," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, 1995, pp. 522–529.

[6] K. Nagao, "Recognizing 3D Objects Using Photometric Invariant," *in Proc. Fifth Int'l Conf. Computer Vision,* 1995, pp. 480–487.

[7] M.A. Turk and A.P. Pentland, "Face Recognition Using Eigenfaces," *in Proc. Conf. Computer Vision and Pattern Recognition,* 1991, pp. 586–591.

[8] H. Murase and S.K. Nayar, "Visual Learning and Recognition of 3D Objects From Appearance," *Int'l J. Computer Vision,* vol. 14, 1995, pp. 5–24.

[9] C. Schmid, R. Mohr, "Local gray-value invariants for image retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19 (5), 1997, pp. 530–535.

[10] A.J. Baerveldt, "A Vision System for Object Verification and Localization Based on Local Features". *Journal of Robotics and Autonomous Systems 34 (2001) Elsevier*, 2001, pp. 83-92.

[11] R.P.N. Rao, D.H. Ballard, "An active vision architecture based on iconic representations", *Artificial Intelligence*, 78(1–2), 199, pp. 461–505.

[12] R. Deriche and G. Giraudon, "A Computational Approach for Corner and Vertex Detection," *Int'l J. Computer Vision*, vol. 10, no. 2, 1993, pp. 101–124.

[13] C. Schmid, R. Mohr, C. Bauckhage, "Comparing and evaluating interest points", *in Proceedings of the International Conferenceon Computer Vision*, Bombay, January 1998.

[14] S.M. Smith, J.M. Brady, "SUSAN—A new approach to low level image processing", *International Journal of Computer Vision,* vol. 23(1), 1997, pp. 45–78.

[15] N. Kruger, N. Peters, C. Malsburg, "Object recognition with a sparse and autonomously learned representation based on banana wavelets", *Technical Report IRINI*, December 1996.

[16] J.J. Koenderink and A.J. van Doorn, "Representation of Local Geometry in the Visual System," *Biological Cybernetics*, vol. 55, 1987, pp. 367–375.

[17] L. Florack, "The Syntactical Structure of Scalar Images", *PhD thesi*s, Universiteit Utrecht, The Netherlands, 1993.

[18] A. Califano and R. Mohan, "Multidimensional Indexing for Recognizing Visual Shapes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, no. 4, 1994, pp. 373–392.

[19] Z. Zhang, R. Deriche, O. Faugeras, and Q.T. Luong, "A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry," *Artificial Intelligence*, vol. 78, 1995, pp. 87–119.

[20] J.J. Koenderink, "The Structure of Images," *Biological Cybernetics*, vol. 50, 1984, pp. 363–396.