Machine Graphics and Vision Journal, Special Issue, 9(2), 2000

View Invariant Object Recognition using Coloured Receptive Fields

D. Hall, J.L. Crowley, V. Colin de Verdière

Projet PRIMA — Lab. GRAVIR–IMAG, INRIA Rhnes–Alpes, Grenoble, France

Abstract. This paper describes a technique for the recognition and tracking of every day objects. The goal is to build a system in which ordinary desktop objects serve as physical icons in a vision based system for man-machine interaction. In such a system, the manipulation of objects replaces user commands.

This method is based on sampling a local appearance function at discrete viewpoints by projecting it onto a vector of receptive fields which have been normalised to local scale and orientation. This paper reports on the experimental validation of the approach, and of its extension to the use of receptive fields based on colour. The experimental results indicate that the technique does indeed provide a method for building a fast and robust recognition technique. Furthermore, the extension to coloured receptive fields provides a greater degree of local discrimination.

The coloured receptive field approach is applied to the recognition of objects under changing view points. Appearance of objects depends strongly on the view point and the lighting. In the experiments we show that the developed technique based on coloured receptive fields allows the recognition of objects invariant from the view point of the camera. This is obtained by training images from view points that sample the view sphere. This experiment shows that the approach is suitable for the recognition of general objects as physical icons in an augmented reality.

Key words: Object Recognition, View Point Changes, Appearance-Based Vision, Phicons

1 Introduction

A phicon is a physical object to which a virtual entity can be attached, such as commands and their parameters. This allows to manipulation of any physical object to serve as a computer interface device [5, 13]. Several typical situations can be constructed in which the use of phicons is easier than the use of keyboard and mouse. For example, an application in an intelligent environment is started when the user picks up the corresponding physical object. Another example is a space mouse. The user can turn a CAD object or navigate in a virtual reality world by manipulating an object that serves as space mouse phicon. The use of a phicon space mouse is more natural than the use of an ordinary space mouse. This meets the paradigm of natural, graspable, wireless, easy to use, human computer interfaces. Our problem is to build such a system to investigate the improvement in usability provided by phicons.

In an augmented reality system, one or more cameras observe a region of interest in which interaction can take place. Such a region can be a desk or more general a three dimensional space within a room. In such an environment the background and the lighting is variable. Translation of objects invoke differences in the view point of the camera and object pose. These problems require a system that is robust to such differences and make the recognition and pose estimation of phicons in an augmented reality an interesting challenge for computer vision.

An important constraint in a phicon based interface is that almost any physical object can serve as phicon and that the user may select the object of his personal interface.

This imposes the constraint that the computer vision system can not be engineered for specific classes of objects. The system must be completely general. In addition, the computer vision system must not interfere with natural interaction. Thus the vision system must have a very low latency (on the order of 50 milliseconds in the case of tracking), and a very low failure rate.

The acceptance of objects with a wider variety of features increases the difficulty of recognition and pose estimation. Although there already exist many different approaches, most established methods work well for restricted classes of objects.

In this article an approach is proposed that allows the recognition of objects invariant to their pose. A possible solution would be provided by colour histograms [12, 10]. However, this approach is not suitable for pose estimation. The extension to pose estimation in 2D and 3D is an important factor for the design of the approach, because the approach will later be extended for the recognition of the manipulation manner of phicons. For this reason receptive fields are preferred to colour histograms.

Colin de Verdière [3] has recently demonstrated a technique for the recognition of objects over changes in view-point and illumination which is robust to occlusions. In this approach, local scale and orientation are estimated at each point in an image. A vector of receptive fields is then normalised to this scale and orientation. The local neighborhood is projected onto this vector. This provides a representation which can be used by a prediction-verification algorithm for fast recognition and tracking, independent of scale and image orientation. View invariant recognition is obtained by sampling this representation at regular intervals over the view sphere. Because the method uses local receptive fields, it is intrinsically robust to occlusions.

In this article we adapt this technique to the problem of recognising and tracking physical icons. The technique is extended by employing coloured receptive fields. The proposed approach allows the recognition of a wide variety of common objects, including objects with features that make recognition difficult, such as specularity and transparency. Evaluation of the experiments show that good results are obtained, even when the object is rotated in 3D in front of the camera.

The next section reviews the description of the local appearance function by projection onto normalised receptive fields vectors. We then describe how this approach can be extended to coloured receptive fields. We provide experimental results which validate the approach using grey scale receptive fields, and then demonstrate the contribution of colour. The coloured receptive fields are then applied to images sampling the entire view sphere.

2 Describing local appearance

In 1991 Adelson and Bergen [2] reported a function that derives the basic visual elements from structural visual information in the world. This function is called the plenoptic function (from "plenus", full or complete, and "opticus", to see). The plenoptic function is the function of everything that can be seen. In machine vision the world is projected onto an image, which is a sample of the plenoptic function:

$$P(x, y, t, \lambda, V_x, V_y, V_z) \tag{1}$$

where (x, y) are the image coordinates, t, the time instant, λ the response wavelength, and (V_x, V_y, V_z) the view point. If the plenoptic function for an object is known it would be possible to reconstruct every possible image of the object; that is from every possible view, at every moment, for every image pixel, at every wavelength.

Adelson and Bergen propose to analyze samples of the plenoptic function using low order derivatives as feature detectors. Koenderink [8] expands the image signal by the first terms of its Taylor decomposition, that is in terms of the derivatives of increasing order. The vector of this set is called "Local Jet". The Local Jet is known to be useful for describing and recognising local features [11]. The signal derivatives are obtained by convolution of the signal by a set of basis functions.

2.1 Gaussian derivatives

Gaussian derivatives provide a basis for a Taylor series expansion of a local signal. This means that a local image neighborhood can be reconstructed by a linear combination of weighted Gaussian derivative filters. This reconstruction becomes an approximation which increases in error as the number of filters is reduced. The formula for the n^{th} 1D Gaussian derivative with respect to the dimension, x, is:

$$\delta_{x^n} g(x,\sigma) = \frac{d^n g(x,\sigma)}{dx^n} = \left(\frac{-1}{\sigma}\right)^n He_n\left(\frac{x}{\sigma}\right) g(x,\sigma), \quad \text{with } g(x,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \tag{2}$$

where He_n stands for the n^{th} Hermite "type e"polynomials [1].

Gaussian derivatives have an explicit scale parameter, σ, and can though be generated at any scale. With steerable filters proposed by Freeman [6] Gaussian derivatives can be oriented in any arbitrary direction. With automatic scale selection [9] the local scale of a feature can be determined. The object in an image can be normalised by scale which allows recognition under scale changes. The determination of the dominant orientation of a neighborhood allows to normalise by orientation. These two properties are used by all techniques presented in this article.

3 Sampling local appearance

In the technique proposed in [3] a training set consists of all overlapping image neighborhoods, referred to as imagettes, of all model images. An imagette is projected onto a single point in the descriptor space R. Each model image can be represented as a grid of overlapping imagettes. The projections of these imagettes form a surface, a local appearance grid, which models the local appearance of the image in R (see figure 1).

Each object is represented by a set of images from different view points. As every image results in a local appearance grid, each object is modeled by the set of surfaces in R. The recognition process equals the search of the corresponding surface for the projection of a newly observed imagette. The basis of all surfaces in R are stored in a structural way, so that the searched surface can be obtained by table lookup. The resulting surface contains information about the object identity, the view point of the camera and information about the relative location of the imagette to the object position. The information from several points allow to estimate the pose of the object.

The approach based on Gaussian derivatives proposed in [3] serves as benchmark for the evaluation of the results. This approach is fast due to efficient storage and recursive filters [14], rotation invariant due to steerable filters [6], invariant to scale due to automatic scale selection [9], and robust to occlusions due to receptive fields. It produces good results for compact textured objects (see section 5.1). The approach

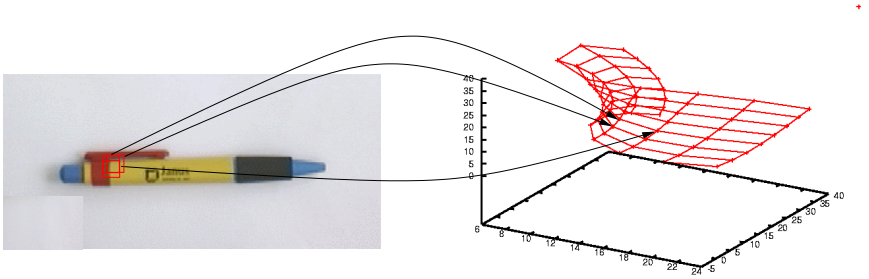

Figure 1: An image as a surface in a subspace of R

fails completely for objects with sparse texture or objects of small sizes or with holes. The reason is that the Gaussian derivatives are computed only from the luminance image. In the luminance image the structure is very well preserved but the chromatic information is lost, and thereby the ability to distinguish objects by their colour. Small or non compact objects can not be recognised because the imagette contains part of the variable background. If the portion of the background is significant the imagette is projected on a different point within the descriptor space. The detection of a surface belonging to another object or no surface at all is possible.

The approach described in this section serves as a starting point for the development of an improved approach. For the discrimination of poorly structured objects, chromatic information is indispensable. In the case of other objects, chrominance improves discrimination. A system that employs structural and chromatic information describes an additional dimension of the plenoptic function. Because this dimension includes more information, it can be expected to produce superior recognition results, at the cost of increased computation. Most of the additional cost may be avoided by keeping the number of receptive fields constant. We compensate the addition of receptive fields for chrominance with a reduction in the number of receptive fields for higher order derivatives. Our experiments show that chrominance is more effective than third order derivatives in discrimination of local neighborhoods.

4 Coloured receptive fields

A new descriptor space is needed that is based on Gaussian derivatives and capable of processing colour images. A direct approach would be to filter each colour channel separately. The advantage would be that no information is lost and no new technique needs to be developed. The disadvantage is that the normalisation process would need to be duplicated independently for each colour channel.

An alternative is to maintain the use of the luminance channel, and to complement this with two channels based on chrominance. The chrominance channels are described using colour-opponent receptive fields. Luminance is known to describe object geometric structure while chrominance is primarily useful for discrimination. Thus a receptive field vector is used in which chrominance receptive fields are normalised with the scale and orientation parameters computed from the luminance channel.

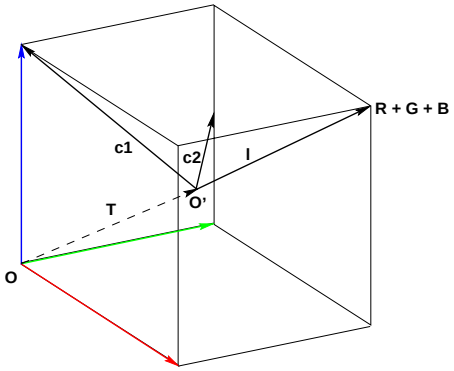

Figure 2: Transformation of the RGB coordinate system.

4.1 Selection of an appropriate colour space

This section addresses the problem of designing the colour opponent receptive fields for chrominance.

The RGB coordinate system is transformed according to following transformation

$$\begin{pmatrix} l \\ c_1 \\ c_2 \end{pmatrix} = T \begin{pmatrix} r \\ g \\ b \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & -1 \end{pmatrix} \begin{pmatrix} r \\ g \\ b \\ 1 \end{pmatrix} \tag{3}$$

This transformation, illustrated in figure 2, moves the origin to the center of the colour cube. One axis corresponds to the luminance axis, which will be used for structure analysis. The other two axis are orthogonal to the luminance axis and are used for colour analysis. We note that the two axis coding colour information are sensitive to red green differences and blue yellow differences, inspired by models of the human visual system [7].

Projection of the image neighborhood onto the luminance axis provides a description of geometric structure. Projection onto the colour difference channel improves discrimination and is less sensitive to specularities and shadows than the image projected onto the luminance axis.

5 Experimental Results

The experiment is based on 8 ordinary objects form an office desktop, that are appropriate to serve as physical icons (shown in figure 3). This set of objects is used to demonstrate the capability of the approach to cope with general objects, among them objects with difficult features and that the approach can even be extended to the recognition of objects invariant from the view point of the camera. The set contains textured and uniform objects, compact objects and objects with holes, specular and transparent objects. Some of the objects can be discriminated easily by their structure (eraser, sweets box), or by their colour (pen, scissors). Other objects exhibit specularities and transparencies which would render most object recognition techniques unreliable (tape, pencil sharpener, protractor). Recognition of such objects is difficult, because small changes

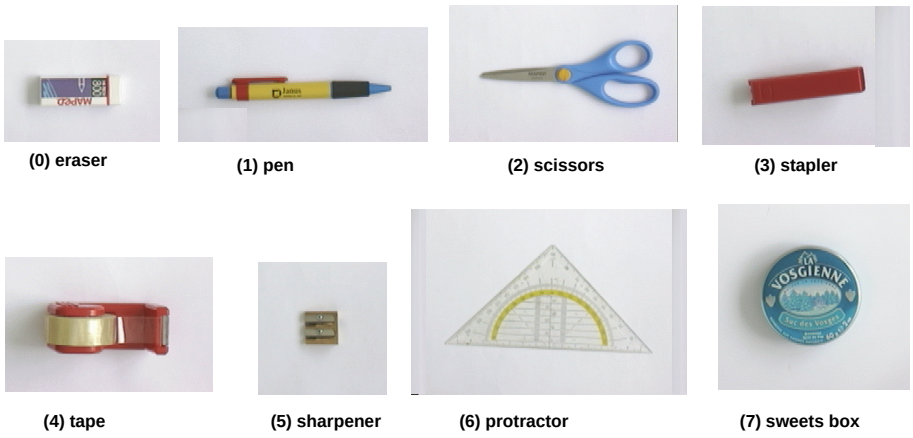

Figure 3: Object set used in the experiments.

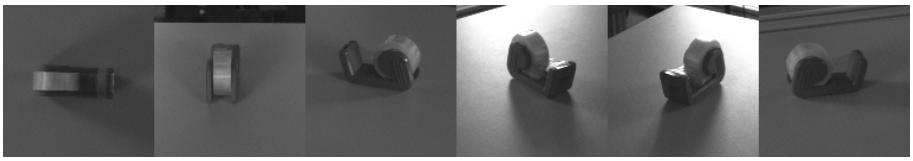

Figure 4: Training images of object tape for experiment 5.3.

of illumination or background conditions invoke significant changes in the appearance of these objects.

For a view invariant recognition of objects an excellent object recognition system is required. Sections 5.1 and 5.2 display the improvement of the addition of chrominance to the receptive fields. Due to the good results obtained in the preliminary experiments, the developed technique is applied to the recognition of the objects in figure 3 sampled over the entire view sphere. The results are shown in section 5.3.

The training phase results in a separate data structure for each experiment. In section 5.1 this data structure contains purely luminance based receptive field vectors up to third order. In section 5.2 and 5.3 the structure contains receptive field responses which include both luminance and chrominance, but are limited to second order. A recognition cycle was run on the test images. A set of 9 test images are used that contain between 2 to 6 different objects of the test set. The orientation and the position of the objects in the test images is different from the orientation and position in the training images. The distance from the camera is constant. In the experiments 5.1 and 5.2 the camera position is static. For the experiment invariant from view points (section 5.3) training and test images on a view sphere were taken using a portique robot [1]. The training base samples the entire view sphere of each object (see Figure 4). The test base contains images that are on the view sphere but are different from the training view points.

For the evaluation of the experiments a grid of image neighborhood locations on the test images were selected using a step size of 5 pixels between neighborhoods. At

[1]The database consists of 8 objects. 5 object are sampled over the entire sphere with 357 images, 3 object are sampled over the half sphere with 186 images. The database is available at `ftp://ftp.inrialpes.fr/pub/prima/images/`

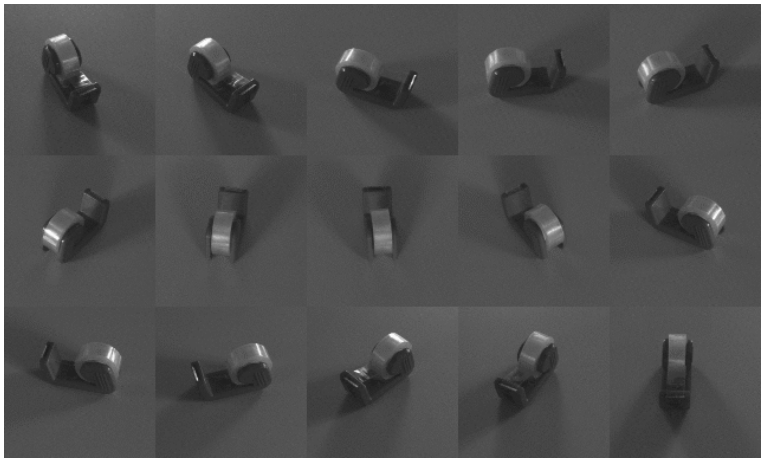

Figure 5: 15 out of 97 test images of object tape on the view sphere.

each neighborhood, the local scale and orientation are determined. The local neighborhood is then projected onto a vector of receptive fields which has been normalised to this scale and orientation. The vector is then used as an index to generate a list of hypotheses for possible objects and image neighborhoods having similar appearance.

For each neighborhood, the method produces a sorted list of image neighborhoods from all the trained objects with a similar appearance. Similarity in appearance is determined by the distance between the vector of responses to the receptive fields. A list of neighborhoods within a tolerance distance (epsilon) are returned. This list is sorted by similarity. If the list is too large, then the neighborhood is judged to be non-discriminant and is rejected. Similarly, if no neighborhoods are found within a tolerance, the neighborhood is judged to be unstable, and is rejected. Neighborhoods for which a small number of similar matches are found are labeled as "accepted" in the experiments below.

The recognition rates must be seen in combination with the acceptance rate. The goal is to obtain high acceptance rates together with high recognition rates. Thus, to evaluate the results of the techniques, three values are presented. First, the percentage of neighborhoods that produced a hypothesis are displayed. The number of such neighborhoods is labeled as the "acceptance rate". This is the percentage of neighborhoods which are both unambiguous and stable. Secondly, we display the number of neighborhoods for which the most similar recalled neighborhood is from the correct object. These cases are labeled "1st answer correct". A third value presents the number of returned neighborhoods for which the correct object and neighborhood was in the best three returned neighborhoods (correct answer among first 3). Such slightly ambiguous neighborhoods can be employed by a prediction-verification algorithm for recognition.

5.1 Local appearance technique based on luminance

This experiment is computed on luminance images according to the technique described in section 3 using recursive filters, automatic scale selection, and steerable filters. This experiment is the benchmark for the following experiments.

Neighborhoods from objects eraser (0), scissors (2), stapler (3), protractor (6), and sweets box (7) have produced good acceptance rates. The acceptance rates for neigh-

object number	0	1	2	3	4	5	6	7
acceptance rate	0.30	0.41	0.65	0.04	0.47	0.54	0.07	0.23
1st answer correct	0.40	0.27	0.62	0.59	0.28	0.12	0.91	0.43
correct answer among first 3	0.77	0.51	0.83	0.82	0.62	0.47	1	0.81

Table 1: Results of technique based on luminance receptive fields. Neighborhoods of objects with discriminant structure are easily recognised. However, luminance provides poor discrimination for uniform and specular objects.

object number	0	1	2	3	4	5	6	7
acceptance rate	0.88	0.87	0.91	0.98	0.83	0.98	0.23	0.99
1st answer correct	0.91	0.98	0.86	0.97	0.74	0.77	0.96	1
correct answer among first 3	0.98	0.99	0.94	0.99	0.90	0.97	0.99	1

Table 2: Results of technique extended to 0^{th} and 1^{st} order Gaussian derivatives in chrominance channels. High recognition rates are obtained for all objects. Average results are slightly superior than those in section 5.2.

borhoods from the stapler (3) and protractor (6) are somewhat lower which indicates that for most of the observed neighborhoods are unstable or ambiguous. These two objects are very hard to recognise by a system using only luminance.

Objects eraser (0), scissors (2) and sweets box (7) produce sufficiently high recognition rates and a simple voting algorithm could be used for recognition. A prediction-verification approach would produce a robust recognition for these objects, as reported by Colin de Verdiére [4]. Poor results for recognising neighborhoods are obtained for objects pen (1), tape (4) and sharpener (5).These objects are either uniform or specular, which makes the recognition using only luminance difficult.

5.2 Object recognition using coloured receptive fields

In this experiment chrominance information is added to the grey scale receptive fields. The two chrominance channels are filtered using a Gaussian and a 1^{st} order Gaussian derivative to capture the average color of the neighborhood and the color gradients that are characteristic for the object. The structure analysis is performed in the 1^{st} and 2^{nd} order derivatives. The 3^{rd} order derivative is abandoned, because its analysis is only interesting when the 2^{nd} order derivative is significant [8]. The descriptor space has than 8 dimension which helps to avoid the problems that occur in high dimensional spaces. The comparison of table 1 and table 2 validates that the improvement by using colour is much superior to the loss in structure recognition by abandoning the 3^{rd} order derivative.

The addition of chrominance information raises the acceptance rates from an average of 0.34 in the previous experiment to an average of 0.83. Many fewer neighborhoods are rejected because of ambiguous or unstable structure. This is an important improvement because even for difficult objects many windows produce a result, which was not the case in the previous experiment. The only object with a low acceptance rate is object protractor (6), which is transparent and particularly difficult to describe.

object number	0	1	2	3	4	5	6	7
acceptance rate	0.88	0.64	0.62	0.76	0.80	0.60	0.66	0.94
1st answer correct	0.56	0.48	0.62	0.58	0.74	0.66	0.78	0.93
correct answer among first 3	0.61	0.52	0.70	0.65	0.80	0.73	0.84	0.95

Table 3: Results for objects on view sphere. 6 training images are used for half sphere.

object number	0	1	2	3	4	5	6	7
acceptance rate	0.74	0.42	0.42	0.54	0.60	0.38	0.19	0.63
1st answer correct	0.73	0.64	0.77	0.72	0.80	0.72	0.50	0.88
correct answer among first 3	0.79	0.69	0.82	0.81	0.86	0.81	0.58	0.92

Table 4: Results for objects on view sphere. 26 training images are used for half sphere.

Very good recognition rates are obtained for all objects. The lowest first answer recognition rates are obtained for objects tape (4) and sharpener (5). These objects are highly specular and thus change their appearance with pose and illumination. Even for these objects the recognition rates are sufficiently high that a simply voting scheme could be used for recognition in restricted domains.

5.3 View point invariant object recognition

For this experiment images sampling the view sphere are used. Two experiments are performed that differ in the number of used training images.

The acceptance rates in table 3 can be compared to those of the previous experiment. The recognition rates are slightly inferior. This is expected, because the identification of objects under different view points is much more difficult. All test images undergo a change in view point between 3° and 22° in longitude and between 8° and 26° in latitude using 6 training images and between 3° and 16° in longitude and between 8° and 17° in latitude using 26 training images. Considering these view point changes the obtained recognition rates are very good.

It is interesting that the number of training images affects the acceptance rates. The fewer images are used the higher are the acceptance rates. A reason for this is that the descriptor space is saturated with the number of trained image neighborhoods. Increasing the dimensionality of the descriptor space would make the neighborhoods in the space more sparse, but this would increase the storage and computation costs.

The above recognition rates are obtained evaluating the hypotheses of each test point standing alone. Combining the hypotheses list from previous recognition processes using an intelligent vote or prediction-verification algorithm would significantly improve the recognition rates.

6 Conclusions

The results presented in this article are incremental and primarily experimental. We have experimentally investigated the extension of the technique of [4] to the problem

of real time observation of the physical icons for computer human interaction. Certain characteristics of real world objects, such as specularity, transparency or low structure, variable background and changing camera positions make the identification of objects difficult. An approach is developed that increases significantly the description ability of receptive fields and produces such good results that the approach can be extended to the recognition of objects invariant of the view point.

The recognition technique evaluated in this article employs local orientation normalisation to provide invariance to image plane rotations. Robustness to scale changes is provided by local normalisation using automatic scale selection. The technique can be implemented to operate in real time by recursively computing separable Gaussian filters. Such filters are steered to the local orientation using the steerability property of Gaussian derivatives. Training was performed for the grey scale technique in 237s on a Pentium II 333 MHz. The techniques using colour needed both 278s for 16 training images of average size of 39 212 pixels.

Grey scale receptive fields are applied to object recognition. It can clearly be stated that this technique works well for textured objects. The object classification of uniform objects fails. The method is extended by the addition of chrominance information. An chrominance descriptor space is presented that can describe colour images and does not increase the dimensionality greatly in comparison to the starting point technique. Problems with high dimensional spaces are avoided. A system is obtained that preserves the advantages of the pure luminance approach and is capable of classifying a much wider range of objects. It is not significantly more expensive in computation and storage. The experimental section validates that objects with difficult features can be recognised, even on cluttered background. It also indicates that chrominance is more important to recognition than higher order derivatives.

The approach is applied to object classification under changing view points. Considering the strong changes in view point and the lighting changes caused by the view point change especially for specular objects, the obtained results are highly satisfactory. In the experiments the hypotheses of each test point standing alone are evaluated. The classification rates can be increased by combining the results of other test points of the same image using intelligent algorithms such as prediction-verification.

We are currently working to extend the approach to view-variant pose estimation. Once the object and its rough position and pose is recognized a second algorithm based on the same principle is used to estimate the precise pose. This algorithm is trained with images of one single object under many different view points. Naturally there will be many ambiguities be present, wich makes a precise pose estimation difficult.

References

[1] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions.* MIT Press, 1965.

[2] E.H. Adelson and J.R. Bergen. The plenoptic function and the elements of early vision. *Computational Models of Visual Processing*, pages 3–20, 1991.

[3] V. Colin de Verdière. *Représentation et Reconnaissance d'Objets par Champs Réceptifs*. PhD thesis, Institut National Polytechnique de Grenoble, France, 1999.

[4] V. Colin de Verdière and J.L. Crowley. A prediction-verification strategy for object recognition using local appearance. Technical report, PRIMA Group, GRAVIR Lab, Grenoble, France, 1999. available at ftp://ftp.inrialpes.fr/pub/prima/.

[5] G. Fitzmaurice, H. Ishii, and W. Buxton. Bricks: laying the foundations for graspable user interfaces. In *Proceedings of Computer Human Interaction (CHI '95)*, pages 442–449. ACM Press, 1995.

[6] W.T. Freeman and E.H. Adelson. The design and use of steerable filters. *Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, September 1991.

[7] D.H. Hubel. *Eye, Brain, And Vision*. Scientific American Library, New York, USA, 1988.

[8] J.J. Koenderink and A.J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, pages 367–375, 1987.

[9] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[10] B. Schiele. *Object Recognition using Multidimensional Receptive Field Histograms*. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, July 1997.

[11] C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997.

[12] M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[13] B. Ullmer, H. Ishii, and D. Glas. mediablocks: Physical containers, transports, and controls for online media. In *Special Interest Group on Computer Graphics (SIGGRAPH '98)*, Orlando, USA, July 1998.

[14] I.T. Young and L.J. van Vliet. Recursive implementation of the gaussian filter. *Signal Processing*, pages 139–151, 1995.