Object Recognition using the Invariant Pixel–Set Signature

J. Matas[1,2] J. Buriánek[1,2] J. Kittler[1]

[1]Center for Machine Perception, CTU Prague, Karlovo nám. 13, CZ 121 35
[2]CVSSP, University of Surrey, Guildford, GU2 7XH, UK
{matas,burianek}@cmp.felk.cvut.cz

Abstract

A new object recognition method, the Invariant Pixel Set Signature (IPSS), is introduced. Objects are represented with a probability density on the space of invariants computed from measurements (pixel values) inside convex hulls of n-tuples of interest points. Experimentally the method is tested on COIL–20, a publicly available database of 72 views of 20 natural object rotating on a turntable. With a model built from *a single view*, recognition performance measured by the average match percentile is above 98% for ± 20 degrees and above 96% for ± 30 degrees. For some object, 100% first rank is achieved for all 72 views. Robustness to occlusion is shown using images with one half covered. For a small change of viewpoint (± 10 degrees) recognition of the occluded object is perfect.

1 Introduction

In this paper a novel appearance-based object recognition (ABOR) method is proposed. Appearance-based (or view-based) methods, i.e. approaches that represent objects by descriptors computed from images without building an explicit model of 3D shape, have a number of attractive features. Firstly, object models are built by generalisation from observed views, using statistical learning techniques. Model acquisition is therefore automatic, time-consuming (manual) model-building is avoided. Secondly, appearance depends on the generally complex interaction of object shape, surface reflectance, illumination and observer's viewpoint. Building realistic models of the combined effect, the main focus of computer graphics, is extremely complex and therefore costly. If a sufficiently rich set of input images is presented in the training phase, the model of appearance will represent, albeit approximately, the combined effects. Finally, many real-life objects have an irregular shape and thus cannot be successfully recognised in the 'classical' framework based on geometrical primitives.

It may seem surprising that a collection of 2D views of a complex 3D object holds enough information for view-point and illumination independent recognition of the object. However, both psychophysical evidence [5] and theoretical considerations support the hypothesis [17]. On the experimental side, impressive performance of recognition al-

BMVC 2000 doi:10.5244/C.14.61

gorithms based on the principle supports the view that a 3D model is indeed not necessary for general recognition [19].

No approach to the complex problem of object recognition should be viewed as a panacea and, naturally, there are certain limitations of the applicability of the appearance based method. Since ABOR is based on direct matching of images, resistance to clutter and occlusion is not easily incorporated in the framework and hence prior segmentation of the object of interest is often required. For the same reason global ABOR methods cannot handle substantial amounts of occlusion[1]. Standard ABOR methods require a large training set [18, 19]. If a scaled orthographic transformation is assumed, a set of images taken from points uniformly sampling the view-sphere is sufficient for representing all possible appearances of the object. In similar conditions, it has been shown by Murase and Nayar [11] that real-time recognition in a database of 100 objects is possible. However, if affine or perspective effects are non-negligible, the memory requirement renders direct image-based methods impractical.

Another fundamental aspect of the large training set problem has been neglected in the literature. If the designer of the recognition system is in full control of the objects of interest, e.g. in a conveyer-belt type of applications, automatic acquisition requires only time, suitable hardware and a learning procedure. However, in many applications only a very limited number of images of the object are available, the illumination and other environmental parameters are beyond control (outdoor scenes) or it is not acceptable to submit the object to a time-consuming acquisition procedure (e.g. face recognition).

The proposed *Invariant Pixel Set Signature Method (IPSS)* overcomes two important problems of ABOR, namely robustness to occlusion and the need for a large training set, by merging geometric and appearance-based techniques. Since perspective (and therefore affine and orthographic) projection preserves convex hulls [21, 22], *invariant appearance-based descriptors can be computed on pixel set corresponding to convex hulls of n-tuples of interest points*.

The rest of the paper describes the IPSS method in detail and present a set of initial experiments. In section 3, the invariant pixel-set signature representation of appearance is introduced. An object recognition strategy based on the selection of object with maximum aposteriori probability given an observed signature is proposed in section 3. Recognition experiments presented in section 4 experimentally confirm that the signature is robust w.r.t change of pose and occlusion. Test are carried out on a publicly available database of 72 views of 20 natural objects. The paper is brought to conclusion in section 5.

2 The Invariant Pixel-Set Signature (IPSS) Representation

"Where in the image should appearance representation of the object be computed?", is a central problem of ABOR. "Everywhere", the standard answer of the global approaches, is not robust to occlusion and we do not consider it. Local methods exploiting features computed in neighbourhoods of interest points have been propsed. E.g. Schmid [2] used differential invariants computed from a set of circular neigbourhoods, Lowe[8] proposed a complex multi-scale representation. Such methods work well under orthographic projection since the shape of the neighbourhood where invariants are computed stays the

[1] but c.f. the work of Leonardis and Bischof [7, 6]

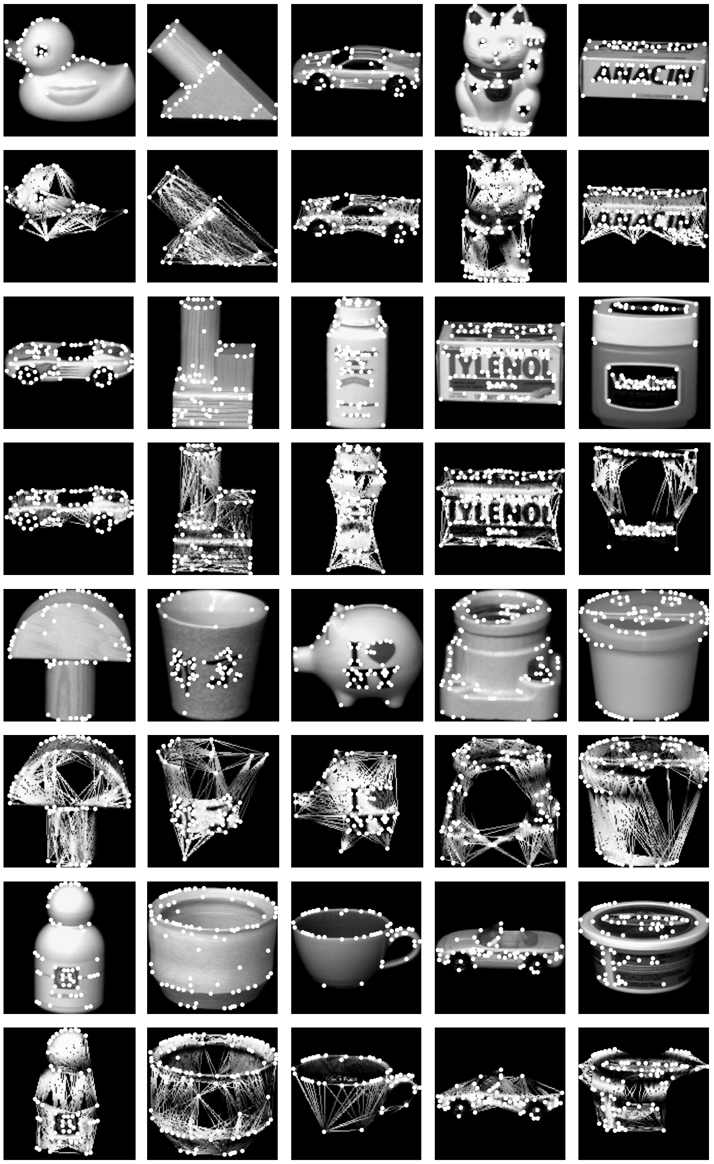


Figure 1: Objects in COIL20 databases with detected intersted points and intensity profiles used for IPSS computation

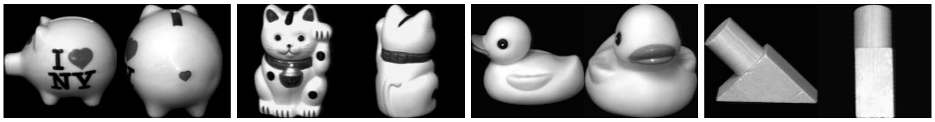


Figure 2: Example of appearance variations of objects in the COIL20 database.

same. However, under affine and perspective transformations, the neighbourhood gets distorted. For example, a circle can become an ellipse. “Invariants” computed from a circular neighbourhood, distorted by the transformation to an extended ellipse, are not invariant at all.

Research into projective invariants has shown that perspective (and therefore affine and orthographic) projection preserves convex hulls [21, 22]. So our answer to the central, “where”, question is: *compute the appearance-based description using pixel values (measurements) from regions defined as convex hulls of the n-tuples of interest points*. The descriptors computed from the pixels inside the convex hulls form the *Invariant Pixel Set Signatures*. Under the assumption of local planarity, the appearance based descriptions is a function of *image data corresponding to the same surface patch in the scene*, which is a necessary pre-requisite for a representation to be invariant. These descriptors could be based on differential invariants, affine-invariant moments, or any of the invariant features proposed in the literature [16, 14, 9].

The computation of the IPSS representation can be summarised as follows:

Algorithm 1: Computing the IPSS

1. Interest point detection (e. g. Harris corner detector).
2. Generation of convex sets (e. g. line segments).
3. Computation of invariants (e. g. normalised intensity profile)

For recognition, the interest points must be stable w.r.t. geometric transformations. We chose the modified Harris corner operator [4] since it outperformed five other detectors evaluated by Schmid et al. [3]. The Harris corner detector is by no means the only possibility; many local operator have been proposed in the literature, e.g. [20, 13]. Examples of interest points detected on models used in experiments described in Section 4 are shown in Fig.1.

The use of n-tuples has a drawback, namely the potential for combinatorial explosion, since if there are N_i interest points in the image, $\binom{N_i}{n}$ n-tuples (regions) can be formed. In order to keep the problem under control, we use only pairs of points (defining line segments) and keep only interest points, that show very high stability with respect to scale and rotation change of the image. From these stable interest points, object descriptor is computed in terms of intensity profiles along a set of line segments. The number of profiles is further reduced by considering only profiles between the interest point and its k nearest neighbourghs[2]. The total number of profiles forming the signature is thus $k * N_p$, where N_p is the number of detected interest points. The selected profiles are shonw in Fig1.

3 MAP Recogniton of IPSS

In the IPSS method, objects are represented by collections (multi–sets) of descriptors $S = \{s_1, s_2, \ldots, s_n\}$, where s_i is a descritor computed from an invariant pixels set i. We

[2] W are aware the the k-nearest neighbour relation is not affine invariant, but, for many configurations it is stable. The issue of selection of an affine invariant subset of profiles is an ongoing research issue.

assume the following probabilstic model of the relationship between model M and the observed signature S.

$$P(S|M) = P(\{s_1, s_2 \ldots, s_n\}|M) = \prod P(s_i|M)P(n|M) \tag{1}$$

The term $P(n|M)$ gives the probability that n invariant sets will form model M. We assume that the components of the signature S are independent and identically distributed. The i.i.d. assumption expressed by eq. (1) is of course a gross simplification and its usefulness must checked by experimentation. In case of no occlusion, $P(n|M)$ provides information about M. However, since $P(n|M)$ it is strongly influenced by occlusion, we chose not to exploit it (equivalent to the assumption that under occlusion $P(n|M_i)$ is a constant).

An maximum aposteriori probability approach to recognition is adopted (another matching strategy for IPSS were explored in [10]). To compute $\arg\max_i P(M_i|S)$, Bayes theorem is used to obtain $P(M_i|s)$. Flat priors are assumed for prior probabilites of observing a given object $P(M_i)$. The probability density functions $P(s_i|M)$ in the intensity profile space are estimated using a kernel-based technique. The estimate of a true multivariate density function $f(\bar{x})$ at a point $\bar{x}_0$ in a d-dimensional data space is given by

$$\hat{f}(\bar{x}_0) = \frac{1}{nh^d} \sum_{i=1}^{n} K_E\left(\frac{\bar{x}_i - \bar{x}_0}{h}\right) \tag{2}$$

where $\bar{x}_i,\ i = 1..n$ are the sample data points and K_E is the Epanechnikov kernel with width h. The kernel was chosen since it has minimum integrated square error [1]. The Epanechnikov kernal is defined as

$$K_E(\bar{x}) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1-\bar{x}^T\bar{x}) & \text{if } \bar{x}^T\bar{x} < 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where c_d is the volume of the unit d-dimensional sphere and $\bar{x}$ are the data points.

The recognition strategy can be summarised as follows:

Algorithm 2: MAP Recongnition using IPSS

1. Compute the IPSS representation for model images S_i.
2. Compute the IPSS representation for test image S^t.
3. Compute $P(M_i|S_t) \approx \prod_j P(M_i|s_j^t)$, $P(M_i|s_j^t)$ are estimated by evaluating eq. (2) at s_j^t using descriptors in S_i.
4. Select the model $M_{i^\star} : i^\star = argmax_i P(M_i|S_t)$.

In practice, negative logarithms of $P(M|s)$ are used and the product is replaced by a sum.

4 Experiments

Recognition performance of the IPSS method is demonstrated in three experiments using data from the COIL–20 database [12]. The medium-size database is publicly available

	Percentage of ranks				Histogram of ranks				
Angle	≤ 1	≤ 2	≤ 3	≤ 4	1	2	3	≥ 4	Average match percentage
+30	85	95	100	100	17	2	1	0	98.95
+20	90	95	95	100	18	1	1	0	99.21
+10	95	100	100	100	19	1	0	0	99.74
-10	95	95	95	100	19	0	0	1	99.21
-20	80	90	95	95	16	2	1	1	98.16
-30	60	90	90	95	12	6	0	2	96.84

Table 1: Ranks for basic experiment with database COIL20.

	Percentage of ranks			
Model #	≤ 1	≤ 2	≤ 3	≤ 4
5	31	39	53	59
9	100	100	100	100
14	100	100	100	100
17	100	100	100	100
20	39	64	80	80

Table 2: The 360 degree expriment. Percentages of ranks of correct models below 1, 2, 3 and 4.

and it has been used in recognition experiments reported in the literature [18, 15]. The database contains images of objects rotated on a turntable taken from a static camera (18 of the 72 views of four selected objects are shown in Fig. 3). Such acquisition arrangement, is more challenging than moving a camera around a static object. As a consequence of the change of the relative position of objects with respect to the light source, relative intensities of different surface patches can change dramatically. Moreover, specularities appear at different positions on the surface. Another reason for choosing COIL–20 is the variability of objects in the database, see Fig. 1. Some of the objects have no surface texture (i.e. its albedo is constant), some are textured (wooden blocks), some have a very complex pattern printed on the surface. The materials are mostly specular. Shape ranges from circularly symmetric with simple geometry (cups and pots) to a complex multi-part, e.g. of the toy cars.

4.1 Recognition in the ± 30 degree range

The first experiment carried out was designed to test the ability of the IPSS method to discriminate between the 20 objects of COIL–20. A model in terms of IPSS was build from a single prototype view (position 0). Images taken at $\pm\ 10, 20, 30$ degrees were used as tests[3] For each test image, the probabilities $P(M_i|S)$ were calculated and the rank of the model corresponding to the test object stored.

Results of the experiment are summarised in Table 1. The histogram of ranks is shown on the right, the cumulative histogram (expressed in percentages) on the left. Unsurprisingly, recognition performance deteriorates as a function of the angular difference between the test and model views. For two reasons, the results are not symmetric w.r.t the

[3] COIL–20 contains 72 views of every object. The turntable was rotated by 5 degree between two consecutive acquisitions.

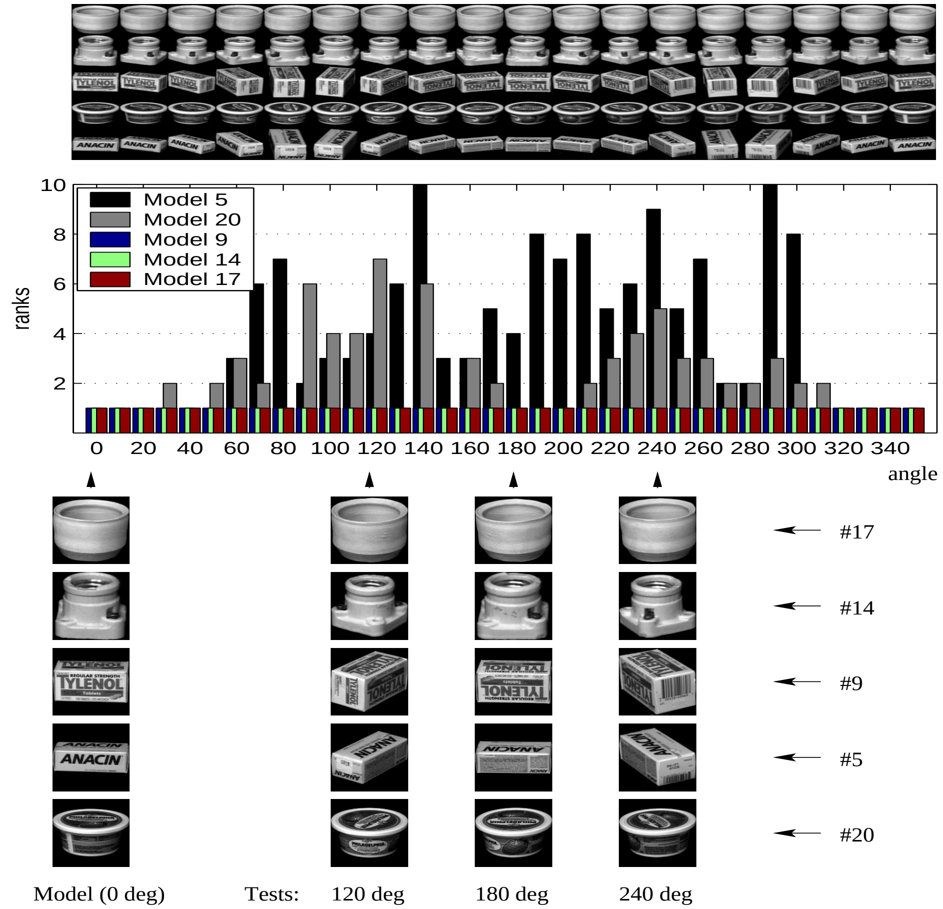


Figure 3: The 360 degree expriment. Test data (top), ranks of the correct model (center) and selected views in higher resolution (bottom).

0 view. Firstly, most objects are not symmetric around the plane passing through the center of the image. Secondly, illumination effects are different for rotations in opposite directions.

Recognition performance, measured by the average match percentile[4] is above 98% in the $(-20, +20)$ interval and above 96% for ± 30 degrees.

In fact, all the mismatches not only for 10 but also for 20 degree views are confined to objects with very similar surface structure, e.g. the three wooden toy blocks (#2, #7 and #11), or the cars (#3, #6 and #19). On the other hand, some object are recognised even if a totally different part of their surface is imaged - see Fig. 2. This is a consequence of the fact that the recognition algorithm does not try to establish correspondence between points (or profiles) which is impossible in this case. However, if the test image profile falls into a region (in the profile space) with high density of measurements from the correct

[4] The average match percentile is defined as $\sum_r \frac{N-r}{N-1} \frac{N(r)}{N}$ where N is the number of models, r is the rank of the model of the test object and $N(r)$ is number of models with rank r.

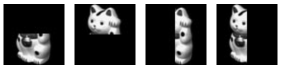
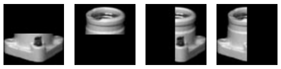
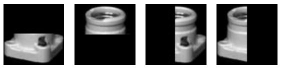
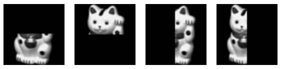
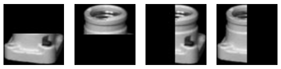
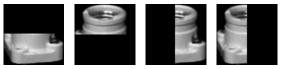
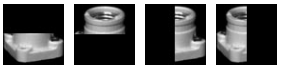
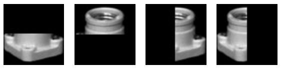
Angle	Model #4	Model #9	Model #14
+30			
+20			
+10			
-10			
-20			
-30			

Table 3: Test images used in the occlusion experiment.

object, it still modifies the probabilities $P(M_i|S)$ in favour of the correct model.

4.2 Recognition of all views from a single model image

Looking again at Fig 2, it is clear that it is unrealistic to expect that all of the objects in the COIL–20 database can be recognised from an arbitrary view using only a model based on a single view. The number of views required for 360 degree recognition depends on various symmetries of the objects. For a selected set of objects, numbers 17, 14, 9, 5 and 20, an experiment was performed where images from all views apart from view 0 were tested (the single view model was built from view 0). Images of rotated objects used in the experiment are shown at the top of Fig. 3, selected views in higher resolution at the bottom.

Results of the experiment are shown in the center of Fig. 3 and summarised in Table 2. For the rotationally symmetric object #17 recognition is perfect. Object #14 has a large rotationally symmetric part and the correct model is also always ranked 1. Object #9 has four almost identical rectangular sides and two squares sides not visible in the model view. The profiles acquired on the model (see Fig. 3, left column of the bottom section) undergo an affine transformation that is not a similarity (highly anisotropic scaling takes place). Recognition performance is still perfect. Object #5 is a similar box, but with different letters printed on its back. Perfect recognition is achieved around 0 and 180 degrees. For the intermediate views, the rank of the correct model is growing to the level close to random guessing. Clearly, at least two views are needed to represent the object. Object #20 has different text written on its side and the lid undergoes a transformation that is not a similarity. Recognition is perfect for a certain interval around 0 degrees $(300, 60)$. This pattern is typical for other asymmetric objects in the database.

4.3 Recognition under occlusion

Four objects were selected for the occlusion experiment. These objects were always recognised correctly in the initial experiments described in section 4.1. Either top, bottom,

	Ranks of models											
Angle	Model #4				Model #9				Model #14			
+30	1	1	2	6	1	1	1	1	1	1	1	3
+20	2	1	1	2	1	1	1	1	1	1	1	1
+10	1	1	1	1	1	1	1	1	1	1	1	1
-10	1	2	1	1	1	1	1	1	1	1	1	1
-20	1	6	1	3	1	1	1	1	1	1	1	1
-30	1	5	2	3	1	1	4	1	1	1	1	1

Table 4: Occlusion experiment. Ranks of correct models.

	Percentage of ranks			
Model #	≤ 1	≤ 2	≤ 3	≤ 4
4	58	79	88	88
9	96	96	96	100
14	96	96	100	100

Table 5: Ranks of correct model in the occlussion experiment.

left or right half of the image was replaced by black in the test image. The test images are shown in Table 3. The model was the same as in the previous experiments, based on a single view at 0 degrees. Ranks of the correct models are shown in Table 4, cumulative percentages summed over all angles are presented in Table 5. For $+10$ degrees, recognition performance is perfect, for -10 it is almost perfect with a single second ranked correct model. As the angular difference grows, performance deteriorates, but for objects #9 and #14 it stay almost perfect. Recognition results for object #4, the toy cat, are worse. However, it is necessary to realise that for the ± 30 degree view a rather small part of the surface visible in view 0 is still visible.

5 Conclusions

A new object recognition method, the Invariant Pixel Set Signature, was introduced. Objects were represented with a probability density on the space of invariants computed from measurements (pixel values) inside convex hulls of n-tuples of interest points. Experimentally the method was tested on COIL-20, a publicly available database of 72 views of 20 natural object rotating on a turntable. With a model built from *a single view*, recognition performance measured by the average match percentile was above 98% in the $(-20, +20)$ interval and above 96% for ±30 degrees. For some object, 100% first rank is achieved for all 72 views. Robustness to occlusion was shown using images with one half covered. For a small change of viewpoint (± 10) recognition of the occluded object is perfect and it deteriorates gracefully with the increase in viewpoint change. Experiments demonstrating recognition performance in scenes containing multiple objects with non-homogeneous background reported in [10] were not included for lack of space.

References

[1] Silverman B.W. *Density Estimation for Statistics and Data Analysis*. New York:Chapman and Hall, 1986.

[2] Cornelia Schmid and Roger Mohr. Matching by local invariants. Technical report, INRIA, 1995.

[3] Cornelia Schmid, Roger Mohr, and Christian Bauckhage. Comparing and evaluating interest points. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'98*, pages 1–7, 1998.

[4] C. Harris and M.J. Stephens. A combined corner and edge detector. In *Alvey88*, pages 147–152, 1988.

[5] Heinrich H. Bulthoff, Shimon Y. Edelman, and Michael J. Tarr. How are three-dimensional objects represented in the brain. Technical report, Massachusetts institute of technology, artificial intelligence laboratory, 1994.

[6] Horst Bischof and Ales Leonardis. Robust recognition of scaled eigenimages through a hierarchical approach. In *Computer Vision and Pattern Recognition*, pages 664–670, 1998.

[7] Ales Leonardis and Horst Bischof. Dealing with occlusions in the eigenspace approach. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pages 453–458. IEEE Compter Society Press, 1996.

[8] David G. Lowe. Object recognition from local scale–invariant features. In *Proceedings of International Conference on Computer Vision ICCV'99*. IEEE, 1999.

[9] A. G. Mamistvalov. n-dimensional moment invariants and conceptual mathematical theory of recognition n-dimensional solids. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:819–831, August 1998.

[10] J.iří Matas, Jan Buriánek, and Václav Hlaváč. Appearance-based object recognition. Research Report CTU-CMP-1999-20, Center for Machine Perception, Czech Technical University, Prague, Czech Republic, November 1999. Confidential. Electronic version is not available.

[11] H. Murase and S. K. Nayar. Visual Learning and Recognition of 3D Objects from Appearance. *International Journal of Computer Vision*, 14:5–24, January 1995.

[12] S. A. Nene, S. K. Nayar, and H. Murase. Columbia Object Image Library (COIL-20). Technical report, Columbia University, 1996.

[13] L. Parida, D. Geiger, and R. Hummel. Junctions: Detection, classification, and reconstruction. *IEEE Transaction on Pattern Analyses and Machine Intelligence*, 20:687–698, July 1998.

[14] Thomas H. Reiss. *Recognizing Planar Objects Using Invariant Image Features*. Springer–Verlag, Berlin Heidelberg, 1993.

[15] Cornelia Schmid and Roger Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, X:Y, to appear 1999.

[16] Dinggang Shen and Horace H.S. Ip. Discriminative wavelet shape descriptors for recognition of 2–D patterns. *Pattern Recognition*, pages 151–165, Feb 1999.

[17] Shimon Ullman and Ronen Basri. Recognition by Linear Combinations of Models. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 13:991–1006, October 1991.

[18] Shree K. Nayar, Hiroshi Murase, and Sameer A. Nene. Parametric Appearance Representation. In S. K. Nayar and T. Poggio, editors, *Early Visual Learning*, pages 131–160, 1996.

[19] Shree K. Nayar and Tomaso Poggio. Early Visual Learning. In S. K. Nayar and T. Poggio, editors, *Early Visual Learning*, pages 1–8, February 1996.

[20] Eero P. Simoncelli and Hany Farid. Steerable wedge filters for local orientation analysis. *IEEE Transactions on Image Processing*, pages 1–15, September 1996.

[21] Tomas Suk and Jan Flusser. Convex layers: A new tool for recognition of projectively deformed point sets. In Franc Solina and Aleš Leonardis, editors, *Computer Analysis of Images and Patterns : 8th Internationa Conference CAIP'99*, number 1689 in Lecture Notes in Computer Science, pages 454–461, Berlin, Germany, September 1999. Springer.

[22] Z. Yang and F. S. Cohen. Image registration and object recognition using affine invariants and convex hulls. *IEEE Transactions on Image Processing*, 8:934–946, July 1999.