Object Recognition Using Composed Receptive Field Histograms of Higher Dimensionality[*]

Oskar Linde and Tony Lindeberg

Computational Vision and Active Perception Laboratory (CVAP)
Department of Numerical Analysis and Computer Science
KTH, SE-100 44 Stockholm, Sweden

Abstract

Recent work has shown that effective methods for recognising objects or spatio-temporal events can be constructed based on receptive field responses summarised into histograms or other histogram-like image descriptors. This paper presents a set of composed histogram features of higher dimensionality, which give significantly better recognition performance compared to the histogram descriptors of lower dimensionality that were used in the original papers by Swain & Ballard (1991) or Schiele & Crowley (2000). The use of histograms of higher dimensionality is made possible by a sparse representation for efficient computation and handling of higher-dimensional histograms.

Results of extensive experiments are reported, showing how the performance of histogram-based recognition schemes depend upon different combinations of cues, in terms of Gaussian derivatives or differential invariants applied to either intensity information, chromatic information or both. It is shown that there exist composed higher-dimensional histogram descriptors with much better performance for recognising known objects than previously used histogram features. Experiments are also reported of classifying unknown objects into visual categories.

1. Introduction

During recent years, the use of view-based representations in terms of receptive field responses (Koenderink & van Doorn 1992) has emerged as a highly promising paradigm for visual recognition. Two main groups of approaches have been developed, where receptive field responses are computed either globally or regionally in terms of histograms (Swain & Ballard 1991, Schiele & Crowley 2000, Schneiderman & Kanade 2000, Zelnik-Manor & Irani 2001), or locally at sparse sets of interest points (Lowe 1999, Hall et al. 2000, Mikolajczyk & Schmid 2001, Mikolajczyk & Schmid 2002, Laptev & Lindeberg 2004b).

The goal of this paper is to address the area of histogram-based recognition and to show that there are other composed histogram features of higher dimensionality, which capture more of the informations contents in the local image structure and give significantly better performance compared to previously used histogram features. Previous work has mainly been based on histograms in two, three or six dimensions, obtained by accumulating the statistics of either colour channels (Swain & Ballard 1991), first-order derivatives alternatively lower-order differential invariants (Schiele & Crowley 2000), or wavelet coefficients (Schneiderman & Kanade 2000).

In this work, we shall first present an efficient method for sparse handling of higher-dimensional histograms, and then use this method for comparing histogram-based recognition schemes that combine Gaussian derivative operators and differential invariants computed from either grey-level information, chromatic information or both. It will be shown how different combinations of differential invariants with chromatic cues allow for significantly better performance compared to the original histogram features used by (Swain & Ballard 1991) or (Schiele & Crowley 2000).

2. Related Work

In early work, Swain & Ballard (1991) showed how reasonable performance of an object recognition scheme could be obtained by comparing RGB colour histograms. Schiele & Crowley (2000) generalised this idea to histograms of receptive fields (Koenderink & van Doorn 1992), and computed histograms of either first-order Gaussian derivative operators or the gradient magnitude and the Laplacian operator at three scales. Schneiderman & Kanade (2000) showed that efficient recognition of faces and cars could be performed from histograms of wavelet coefficients.

With regard to approaches based on interest points, Mikolajczyk & Schmid (2001, 2002) combined the ideas of scale selection (Lindeberg 1998) and affine normalisation (Lindeberg & Gårding 1997) to construct scale invariant and affine invariant Harris operators for object recognition. Hall et al. (2000) performed object recognition based on a combination of Gaussian derivative operators and chromatic derivatives computed at scale adapted interest points. Tuytelaars & Van Gool (2000) developed a method for affine normalisation of interest points based on local edge information. Lowe (1999) combined ideas of feature based and histogram based image descriptors, and defined a scale invariant feature transform, SIFT, which integrates the accu-

[*] The support from the Swedish Research Council and from the Royal Swedish Academy of Sciences as well as the Knut and Alice Wallenberg Foundation is gratefully acknowledged. We would also like to thank Barbara Caputo for valuable discussions and for granting access to her modified version of libSVM.

mulation of statistics of gradient vectors in local neighbourhoods of scale adapted interest points with summarising information about the spatial layout. Other closely related probabilistic methods have been presented by Weber et al. (2000) and Fergus et al. (2003). A recent evaluation and comparison of several of these spatial recognition methods is given in (Mikolajczyk & Schmid 2003).

Histogram-based methods for recognising spatio-temporal events have been developed by Zelnik-Manor & Irani (2001) and Laptev & Lindeberg (2004a), with an extension to local spatio-temporal features in (Laptev & Lindeberg 2004b). Histogram-like descriptors that combine spatial moments with colour information have been proposed by Slater & Healey (1995) and an early study of the sensitivity of histogram descriptors to view variations was done by Schiele & Crowley (1996). A theory for modelling multi-scale histograms has been presented by Koenderink & van Doorn (1999) and a method for performing scale selection in multi-scale histograms has been developed by Hadjidemetriou et al. (2002).

3. Higher-Dimensional Histograms

When using histograms as a basis for recognition, computational problems can easily occur if the dimensionality of the histogram is too high. For example, a 16-dimensional histogram with 15 quantisation levels per dimension contains $15^{16} \approx 10^{19}$ cells in total, while most of the cells will be empty in practice.

For a D-dimensional histogram, with $r_1, \ldots, r_D$ quantisation levels for the D different dimensions, we enumerate the cells from 0 to $\left(\prod^D r_i\right) - 1$, such that a quantised local measurement $M = (m_1, \ldots, m_D)$, $0 \leq m_i < r_i$, will correspond to the cell with index $c = \sum_{i=1}^{D} \left(m_i \prod_{j=1}^{i-1} r_j \right)$. A histogram, H, with n non-zero cells $c_1 \ldots c_n$, with values $v_1 \ldots v_n$, respectively, will then be stored as an interleaved array of size $2n$, $H = (c_1, v_1, c_2, v_2, \ldots, c_n, v_n)$, sorted in such a way that $c_1 < c_2 < \ldots < c_n$.

The sorted property of the stored histograms makes it possible to define efficient operations on them. When accumulating a histogram we first create an array with one entry for each pixel in the image, and use efficient sorting to simultaneously solve the indexing and updating steps when accumulating the histogram. In this way, a sixteen-dimensional histogram of a 256×256 image can be computed in about 0.1 s on a 1 GHz Sun Fire.

4. Image Descriptors

In the following, we will evaluate the performance of histogram based object recognition using different types of image descriptors on different image databases. In brief, we will consider the following basic types of image descriptors, as well as various combinations of these:

- Normalised Gaussian derivatives, obtained by computing partial derivatives $(L_x, L_y, L_{xx}, L_{xy}, L_{yy})$ from the scale-space representation $L(\cdot, \cdot;\ t) = g(\cdot, \cdot;\ t) * f$ obtained by smoothing the original image f with a Gaussian kernel $g(\cdot, \cdot;\ t)$, and multiplying the regular partial derivatives by the standard deviation $\sigma = \sqrt{t}$ raised to the order of differentiation (Lindeberg 1994).

- Differential invariants, invariant to rotations in the image plane, mainly the normalised gradient magnitude $|\nabla_{\text{norm}} L| = \sqrt{t(L_x^2 + L_y^2)}$, the normalised Laplacian $\nabla^2_{\text{norm}} L = t(L_{xx} + L_{yy})$, the normalised determinant of the Hessian $\det(\mathcal{H}_{\text{norm}} L) = t^2(L_{xx} L_{yy} - L_{xy}^2)$.

- Chromatic cues obtained from RGB-images according to $C_1 = (R - G)/2$ and $C_2 = (R + G)/2 - B$.

Unless otherwise mentioned, all image primitives are computed at scale levels $\sigma \in \{1, 2, 4\}$. For the databases studied in this work, this choice is reasonable, since there are no major scale variations in these databases.

Measures. The histograms h_1 and h_2 for any pair of images will be compared using the χ^2 measure: $\chi^2(h_1, h_2) = \sum_i (h_1(i) - h_2(i))^2 / (h_1(i) + h_2(i))$.

Prior to the accumulation of the histograms, the maximum and minimum values of the image primitives for the entire database were computed, and 15 quantisation levels were uniformly distributed between the extremum values, with the bins arranged symmetrically around the origin.

Image databases. For evaluation, we have mainly used the COIL-100 and ETH-80 databases. The COIL-100 database (Nene et al. 1996) contains colour images of 100 objects, each image of size 128×128. There are 72 images per object, corresponding to the viewpoints obtained by rotating the objects through 360° at 5° increments.

The ETH-80 database (Leibe & Schiele 2003) contains images of 80 objects grouped into the eight categories “apples”, “pears”, “tomatoes”, “cars”, “cups”, “cows”, “horses” and “dogs”, with a view-sphere sampling of 22.5° along the equator. Here, we used images of size 256×256.

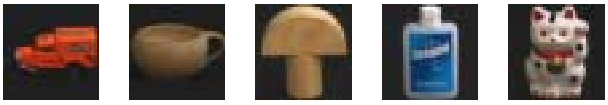

Figure 1: *Sample images from the COIL-100 database.*

Figure 2: *Sample images from the ETH-80 database.*

Classification criteria. Using nearest-neighbour classification, a match will be regarded as correct if the lowest χ^2 score is obtained for any view of the correct object. When using SVM, we follow the approach by (Chapelle et al. 1999, Belongie et al. 2002) with a kernel of the form $K(h_1, h_2) = \exp(-\gamma\chi^2(h_1, h_2))$, where $\chi^2(h_1, h_2)$ denotes the regular χ^2 measure for comparing two histograms, h_1 and h_2, and we set $\gamma = 1$ from initial experiments on the COIL-100 database. The histograms were normalised to unit mass prior to SVM-based training and recognition. The actual implementation of the SVM was done based on the libSVM software (Chang & Lin 2001).

5. Single-View Object Representations

Due to the dense sampling of the viewing directions in the COIL-100 and ETH-80 databases, we can use the images in these databases for evaluating the robustness of different types of image descriptors under variations in the viewing direction. Initially, we restrict ourselves to the problem of object recognition from a single view; thus using a single view of each object for training and all other views in the database for testing. A match on the COIL-100 database is regarded as correct if the lowest χ^2 score is obtained for the same object. Since the ETH-80 database, however, contains many similarly looking object instances, such as individual apples, pears and tomatoes, we will initially regard a match on the ETH-80 database as correct if the best match is obtained for an object within the same category.

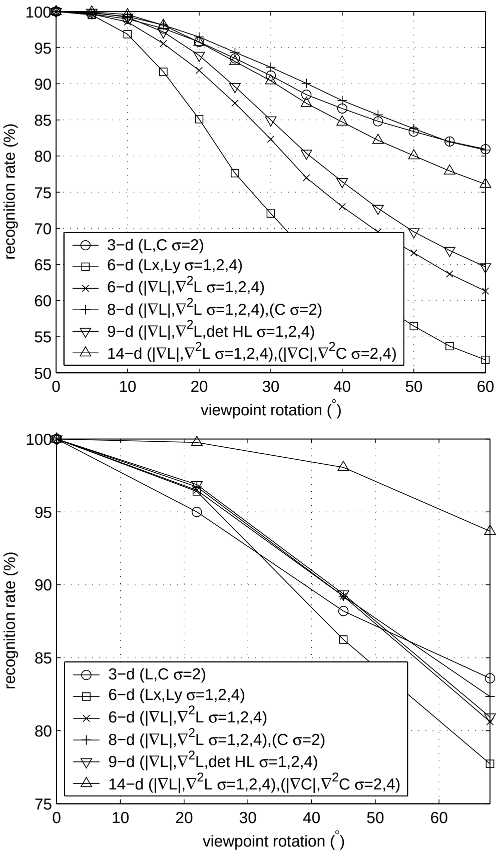

Figure 3: *Comparison of how the recognition rate varies with the difference in viewing angle for different types of histogram features applied to single-view recognition on the COIL-100 (top) and ETH-80 (bottom) databases.*

The experiment will be repeated for all combinations of reference and test views in the database, and for each one of the following histogram features: (i) 3-D colour histograms of (L, C_1, C_2) at one scale, (ii) 6-D shape histograms of first-order partial derivatives (L_ξ, L_η) at three scales, (iii) 6-D shape histograms of the gradient magnitude and the Laplacian $(|\nabla_{\text{norm}} L|, \nabla^2_{\text{norm}} L)$ at three scales, (iv) 8-D combined histograms of chromatic information (C_1, C_2) at one scale and the gradient magnitude and the Laplacian $(|\nabla_{\text{norm}} L|, \nabla^2_{\text{norm}} L)$ at three scales (v) 9-D shape histograms of the gradient magnitude, the Laplacian as well as the determinant of the Hessian $(|\nabla_{\text{norm}} L|, \nabla^2_{\text{norm}} L, \det(\mathcal{H}_{\text{norm}} L))$ at three scales, (vi) 14-D combined histograms of the gradient magnitude and the Laplacian computed from grey-level cues $(|\nabla_{\text{norm}} L|, \nabla^2_{\text{norm}} L)$ at three scales and chromatic cues $(|\nabla_{\text{norm}} C|, \nabla^2_{\text{norm}} C)$ of (C_1, C_2) at two scales.

Figure 3 summarises the results, by showing how the average recognition rate decreases with the angle between the training view and the test view. As can be seen, the performance of the higher-dimensional histogram features, in particular the 9-D shape descriptor and the 14-D combined shape and colour descriptor, is substantially better than for the more commonly used 3-D or 6-D descriptors.

6. Multi-View Object Representations

Let us next turn to internal object representations in terms of multi-view object representations, where we initially focus on the COIL-100 database due to its denser sampling of the viewing directions on the equator of the viewsphere. The methodology we will follow is to use views with a regular spacing of $10°/20°/30°/45°/60°/90°$ for training and the other views for testing. Classification will be performed either using nearest-neighbour comparisons (NN) on the χ^2 measure or using a support vector machine (SVM).

Table 1 shows the recognition performance obtained in this way for the following histogram descriptors: (i) 3-D colour histograms of (L, C_1, C_2) at one scale, (ii) 6-D shape histograms of the gradient magnitude and the Laplacian $(|\nabla_{\text{norm}} L|, \nabla^2_{\text{norm}} L)$ at three scales, (iii) 9-D histograms of the gradient magnitude, the Laplacian and the determinant of the Hessian $(|\nabla_{\text{norm}} L|, \nabla^2_{\text{norm}} L, \det(\mathcal{H}_{\text{norm}} L))$ at three scales, (iv) 14-D histograms of the gradient magnitude and the Laplacian of grey-level information $(|\nabla_{\text{norm}} L|, \nabla^2_{\text{norm}} L)$ at three scales and the gradient magnitude and the Laplacian of chromatic information $(|\nabla_{\text{norm}} C|, \nabla^2_{\text{norm}} C)$ at two scales.

number of training views	36	18	12	8	6	4
angle between views	10°	20°	30°	45°	60°	90°
3-D colour NN	0.03 %	0.20 %	0.62 %	1.45 %	2.65 %	4.50 %
6-D shape NN	0.03 %	0.26 %	1.15 %	3.27 %	5.53 %	12.62 %
9-D shape NN	0.03 %	0.09 %	0.47 %	2.25 %	4.61 %	11.43 %
14-D combined NN	**0 %**	**0.07 %**	**0.23 %**	1.23 %	2.82 %	5.68 %
3-D colour SVM	**0 %**	**0.02 %**	0.22 %	0.77 %	1.73 %	4.13 %
6-D shape SVM	0.03 %	0.26 %	0.57 %	2.03 %	4.73 %	8.26 %
9-D shape SVM	0.03 %	0.06 %	0.18 %	0.69 %	3.53 %	5.26 %
14-D combined SVM	**0 %**	**0.04 %**	**0.12 %**	0.63 %	2.01 %	2.87 %

Table 1: *Error rates for multi-view recognition on the equator of the viewsphere for the COIL-100 database.*

As can be seen from a comparison with figure 3, the performance is improved considerably, almost by an order of magnitude, by using a view-sphere representation compared to recognition from single object views. Moreover, using a support vector machine for the classification step results in significantly better performance than using nearest-neighbour classification. For narrow (10°) or rather sparse (60–90°) view-sphere samplings, the 6-D shape descriptor is almost as good as the 9-D shape descriptor. In the intermediate range (20–45°), however, the 9-D shape descriptor gives a significantly better recognition rate than the 6-D descriptor. The pure 3-D colour descriptor performs very well on this database, which can be understood from the fact that the objects in this database have rather distinctive colours. The 14-D combined shape and colour descriptor with SVM, however, performs better than all the other descriptors. For view-sphere samplings up to 30°, the recognition rate for the 14-D descriptor is 99.9 % or higher, and no errors are made for a 10° view-sphere sampling.

7. Object Instances vs. Categories

In the COIL-100 database, most of the objects have rather different appearance. To investigate the ability of the histogram features to distinguish between more similar objects, as well as their ability to group similarly looking objects, we will next perform three different types of object recognition experiments on the ETH-80 database:

Classifying previously seen objects into categories. Let us first consider an experiment, where we regard a match as correct if the best match is obtained for an object within the same category. The motivation for carrying out this type of experiment is that many of the objects in the same category are quite similar. For example, for the categories “apples” and “pears”, the requirement of recognising any specific object instance implies that we have to recognise a specific apple or a specific pear. For training, we here use every second or every fourth image on the equator of the viewsphere, while all other images on the equator are used for testing.

Table 2 shows the result of this experiment. Here, the 6-D shape descriptor performs better than the 9-D shape descriptor. The performance of the 14-D combined shape and colour descriptor is, however, much better than for any of the other histogram features. Notably, with nearest-neighbour matching and a view-sphere sampling of 45°, the recognition rate is 100 % on this dataset.

	Nearest Neighbour	
view sampling	45°	90°
3-D colour	1.41 %	4.06 %
6-D shape	0.94 %	3.65 %
9-D shape	2.66 %	4.69 %
14-D combined	**0 %**	0.73 %

Table 2: *Category matching of known objects on ETH-80.*

Recognising specific object instances. To investigate the specificity of the image descriptors, we also performed an experiment where the recognition performance was evaluated by requiring a match to the correct object instance, using every second image on the equator for training and the other images on the equator for testing.

Table 3 shows result of this experiment. As we can see, the performance of the 14-D combined shape and colour descriptor is superior to the other histogram features. Considering that the system is faced with tasks such as recognising individual tomatoes, apples and pears, a performance of 98.75 % with nearest-neighbour matching or 99.84 % with SVM must be regarded as extremely good. The situation is, however, simplified due to the reasonably dense sampling of the view-sphere by increments of 22.5° and image acquisition using similar backgrounds and illumination conditions.

	NN	SVM
view sampling	45°	45°
3-D colour	5.00 %	1.09 %
6-D shape	5.00 %	1.09 %
9-D shape	6.14 %	1.25 %
14-D combined	**1.25 %**	**0.16 %**

Table 3: *Object instance matching on ETH-80.*

It should be noted, however, that these results, as well as all previous results, are not optimised. With complementary tuning of the bin size, the performance of the 14-D descriptor on this problem increases to 100 % using SVM.

Classifying previously unseen objects into categories. To investigate the generalising property of the descriptors, we also made an experiment on the ETH-80 database, where the entire viewsphere of the test object was removed from the database prior to recognition. Contrary to the previous experiments, we here use the entire viewsphere of each object for training and testing. The performance in this setup was hence measured by requiring a match to the appropriate category, while disabling the possibility of matching any view of the correct object.

	NN	r	SVM	r
3-D colour	26.2 %	23	18.8 %	23
6-D shape	24.2 %	55	10.4 %	55
9-D shape	23.8 %	45	7.7 %	45
14-D combined	19.9 %	45	8.4 %	45
6-D partial-derivatives	19.9 %	55	6.9 %	95
15-D partial-derivatives	11.0 %	2	12.9 %	2

Table 4: *Classifying previously unseen objects on ETH-80.*

Table 4 shows the results for the rotationally invariant 3-D, 6-D, 9-D and 14-D descriptors used previously, as well two more rotationally variant descriptors; a 6-D histogram of the first-order partial derivatives (L_x, L_y) at three scales and a 15-D descriptor with first- and second-order derivatives $(L_x, L_y, L_{xx}, L_{xy}, L_{yy})$ at three scales.

For this category classification task, we also performed a coarse search over the number of bins, r, for each dimension in the histogram. As can be seen from the results, for this experiment, which does not involve variations over viewing directions, the rotationally variant partial derivative descriptors give better results than the previous rotationally invariant descriptors. Using nearest-neighbour classification, the 15-D partial derivative descriptor (89.0 %) gives much better results than any other of these descriptors (second best 80.1 %). Somewhat surprisingly, however, the classification performance of the support vector machine applied to the 15-D descriptor is lower than for nearest neighbours. Hence, with SVM classification, the 6-D histogram of (L_x, L_y) at three scales gives the best results (93.1 %). Among the rotationally invariant histogram descriptors, which can be expected to have higher robustness to viewing variations, the 9-D shape descriptor with $(|\nabla L|, \nabla^2 L, \det \mathcal{H} L)$ has the best performance (92.3 %).

We can compare these results to (Leibe & Schiele 2003), where the best single-cue performance was 86.4 %, and then using a multi-cue classification scheme based on shape descriptors, colour histograms, PCA, texture histograms and assuming a perfect segmentation of the objects, they arrived at a 93.0 % classification rate.

Table 5 shows the confusion matrix using the 9-D shape descriptor. From an inspection of the errors, the majority of the mistakes are due to mixing up cows, horses and dogs. If we group the cows, horses and the dogs into a joint category “plastic toy animals”, the performance on the simpler category classification problem is 98.4 %.

	1	2	3	4	5	6	7	8
1. apple	409	0	1	0	0	0	0	0
2. pear	0	409	0	0	0	0	0	1
3. tomato	1	0	409	0	0	0	0	0
4. car	0	0	0	405	0	0	3	2
5. cup	0	0	0	5	372	32	0	1
6. cow	0	0	0	2	0	347	39	22
7. horse	0	0	0	0	0	40	325	45
8. dog	0	0	0	4	0	16	38	352

Table 5: *Confusion matrix for category classification on the view-sphere of the ETH-80 database using the 9-D shape descriptor.*

Classifying previously unseen objects in cluttered scenes. To test the generalisation property of the image descriptors on another dataset, we also made experiments on the CalTech/Oxford database (Fergus et al. 2003), which contains 800 images of objects for each one of the three categories “motor-bikes”, “airplanes” and “car rears” as well as 435 images of “faces”.

Figure 4: *Sample images from the CalTech/Oxford database.*

This dataset was divided into one training set with 400 images for each one of the categories motor-bikes, airplanes and car rears as well as 218 images of faces. The test set consisted of all the other images in the database.

	NN	SVM
3-D colour	8.5 %	5.4 %
6-D shape	6.6 %	1.6 %
9-D shape	12.9 %	**0.8 %**
14-D combined	6.1 %	**1.3 %**

Table 6: *Classification errors for the CalTech/Oxford database.*

Table 6 shows the result of these experiments. Using SVM on the 9-D shape histogram, we obtain a classification rate of 99.2 %. Qualitatively, we can compare this result with the results in (Fergus et al. 2003), who report equal error rates from an ROC curve within the range 90–96 % depending on the object class. These error rates are, however, not directly comparable, since in our experiment we consider a multi-category problem, while (Fergus et al. 2003) consider a category detection problem, with explicit modelling of the background class.

	1	2	3	4
1. airplane	396	0	1	3
2. car	0	400	0	0
3. face	0	0	213	4
4. motorbike	3	0	0	397

Table 7: *Confusion matrix for category classification on the CalTech/Oxford database using the 9-D shape descriptor.*

There is, however, also potential for improving our global experiments by computing histograms of image patches and/or using more refined statistical models of the visual properties that are common for the objects within each category. From an inspection of the classification errors, it appears as if a main cause of errors in this experiment is that the matching was sometimes made on properties of the background instead of properties of the objects.

8. Summary and Conclusions

We have presented an evaluation of using different types of histogram features for view-based object recognition. Specifically, we have explored a set of composed higher-dimensional histogram features and evaluated them on different types of recognition problems and image databases.

The proposed histogram descriptors were formulated in terms of different combinations of partial derivatives and differential invariants, computed from either grey-level information, colour information or both. The motivation for studying such composed descriptors of higher dimensionality is to capture more of the information contents in the local image structure than is reflected in previously used histogram features, and we have shown that there exist composed image descriptors that lead to significant improvements in recognition performance compared to previously used histograms descriptors. In order to be able to use higher-dimensional histograms for recognition, which may easily lead to computational problems unless special precautions are taken, we have also presented a sparse method for efficient computation, handling and comparison of higher-dimensional histograms.

An overall conclusion from this study is that compared to previously used histogram features, there exist histogram descriptors that are more powerful for capturing the visual appearance of the objects in the training set and histogram descriptors that allow for higher degree of robustness under variations in the viewing direction. If the task is to recognise specific instances of known objects, even a simple classifier, such as nearest-neighbour comparisons on the χ^2 measure, may in some cases give a high recognition rate based on the proposed composed histogram features. The performance is, however, usually much better using a support vector machine for the recognition step.

Concerning limitations of this work, we have throughout this investigation computed global histograms of all images. For the image databases studied here, such an approach is reasonable, since the objects to be recognised usually cover a major part of the image. To apply histogram-based recognition in a more general setting, where the objects to be recognised only occupy a minor part of the image, these image descriptors should obviously be complemented by some mechanism to perform comparisons based on regional as opposed to global image features. Concerning other extensions, the effectiveness of composed higher-dimensional image cues revealed by this study should be of interest also for view-based recognition schemes based on local features, as well as for related recognition tasks, such as texture analysis or interpretation of spatio-temporal events.

To conclude, the overall performance of the proposed histogram features must be regarded as highly competitive, and there is an interesting potential in exploring other types of composed image descriptors (Linde & Lindeberg 2004).

References

Belongie, S., Fowlkes, C., Chung, F. & Malik, J. (2002). Partitioning with indefinite kernels using the Nyström extension, *ECCV'02*, Springer LNCS 2352, III:531–542.

Chang, C.-C. & Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chapelle, O., Haffner, P. & Vapnik, V. (1999). SVMs for histogram-based image classification, *IEEE-NN*, **10**(5): 1055–1065.

Fergus, R., Perona, P. & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning, *CVPR'03*, 264–271.

Hadjidemetriou, E., Grossberg, M. & Nayar, S. (2002). Resolution selection using generalized entropies of multiresolution histograms, *ECCV'02*, Springer LNCS 2350, I:220–235.

Hall, D., de Verdiere, V. & Crowley, J. (2000). Object recognition using coloured receptive fields, *ECCV'00*, LNCS 1842, I:164–177.

Koenderink, J. J. & van Doorn, A. J. (1992). Generic neighborhood operators, *IEEE-PAMI* **14**(6): 597–605.

Koenderink, J. & van Doorn, A. (1999). Blur and disorder, *Proc. Scale-Space'99*, Springer LNCS 1682, 1–9.

Laptev, I. & Lindeberg, T. (2004a). Velocity-adapted spatio-temporal receptive fields for direct recognition of activities, *IVC* **22**(2): 105–116.

Laptev, I. & Lindeberg, T. (2004b). Local descriptors for spatio-temporal recognition, *ECCV'04 Workshop on Spatial Coherence for Visual Motion Analysis*, in press.

Leibe, B. & Schiele, B. (2003). Analyzing appearance and contour based methods for object categorization, *CVPR'03*, II:409–415.

Linde, O. & Lindeberg, T. (2004). in preparation.

Lindeberg, T. (1994). *Scale-Space Theory in Computer Vision*, Kluwer.

Lindeberg, T. (1998). Feature detection with automatic scale selection, *IJCV* **30**(2): 77–116.

Lindeberg, T. & Gårding, J. (1997). Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D structure, *IVC* **15**: 415–434.

Lowe, D. (1999). Object recognition from local scale-invariant features, *ICCV'99*, 1150–1157.

Mikolajczyk, K. & Schmid, C. (2001). Indexing based on scale invariant interest points, *ICCV'01*, I:525–531.

Mikolajczyk, K. & Schmid, C. (2002). An affine invariant interest point detector, *ECCV'02*, Springer LNCS 2350, I:128–142.

Mikolajczyk, K. & Schmid, C. (2003). A performance evaluation of local descriptors, *CVPR'03*, II:257–263.

Nene, S. A., Nayar, S. K. & Murase, H. (1996). Columbia object image library (COIL-100), *TR CUCS-006-96*, CAVE, Columbia University.

Schiele, B. & Crowley, J. (1996). Object recognition using multidimensional receptive field histograms, *ECCV'96*, Springer LNCS 1064, I:610–619.

Schiele, B. & Crowley, J. (2000). Recognition without correspondence using multidimensional receptive field histograms, *IJCV* **36**(1): 31–50.

Schneiderman, H. & Kanade, T. (2000). A statistical method for 3D object detection applied to faces and cars, *CVPR'00*, I:746–751.

Slater, D. & Healey, G. (1995). Combining colour and geometric information for illumination invariant recognition of 3-D objects, *ICCV'95*, 563–568.

Swain, M. & Ballard, D. (1991). Color indexing, *IJCV* **7**(1): 11–32.

Tuytelaars, T. & Van Gool, L. (2000). Wide baseline stereo matching based on local, affinely invariant regions, *BMVC'00*, pp. 412–425.

Weber, M., Welling, M. & Perona, P. (2000). Unsupervised learning of models for visual object class recognition, *ECCV'00*, Springer LNCS 1842, I:18–32.

Zelnik-Manor, L. & Irani, M. (2001). Event-based analysis of video, *CVPR'01*, II:123–130.