In Computer Vision and Image Understanding 116 (2012) 538–5602. DOI: 10.1016/j.cviu.2011.12.003.

Composed Complex-Cue Histograms: An Investigation of the Information Content in Receptive Field Based Image Descriptors for Object Recognition

Oskar Linde and Tony Lindeberg

School of Computer Science and Communication
KTH (Royal Institute of Technology)
SE-100 44 Stockholm, Sweden

Abstract

Recent work has shown that effective methods for recognizing objects and spatio-temporal events can be constructed based on histograms of receptive field like image operations.

This paper presents the results of an extensive study of the performance of different types of receptive field like image descriptors for histogram-based object recognition, based on different combinations of image cues in terms of *Gaussian derivatives* or *differential invariants* applied to either *intensity information*, *colour-opponent channels* or both. A rich set of composed complex-cue image descriptors is introduced and evaluated with respect to the problems of (i) recognizing previously seen object instances from previously unseen views, and (ii) classifying previously unseen objects into visual categories.

It is shown that there exist novel histogram descriptors with significantly better recognition performance compared to previously used histogram features within the same class. Specifically, the experiments show that it is possible to obtain more discriminative features by combining lower-dimensional scale-space features into *composed complex-cue histograms*. Furthermore, different types of image descriptors have different relative advantages with respect to the problems of object instance recognition *vs.* object category classification. These conclusions are obtained from extensive evaluations on two mutually independent data sets.

For the task of *recognizing specific object instances*, *combined histograms of spatial and spatio-chromatic derivatives are highly discriminative*, and several image descriptors in terms *rotationally invariant (intensity and spatio-chromatic) differential invariants* up to order two lead to very high recognition rates.

For *category classification*, primary information is contained in *both first- and second-order derivatives, where second-order partial derivatives constitute the most discriminative cue*.

Dimensionality reduction by principal component analysis and variance normalization prior to training and recognition can in many cases lead to a significant increase in recognition or classification performance. Surprisingly high recognition rates can even be obtained with *binary histograms* that reveal the *polarity of local scale-space features*, and which can be expected to be particularly robust to illumination variations.

An overall conclusion from this study is that compared to previously used lower-dimensional histograms, *the use of composed complex-cue histograms of higher dimensionality reveals the co-variation of multiple cues and enables much better recognition performance*, both with regard to the problems of recognizing previously seen objects from novel views and for classifying previously unseen objects into visual categories.

Keywords: image descriptor, histogram, object recognition, object categorization, Gaussian derivative, spatio-chromatic derivative, differential invariant, spatio-chromatic differential invariant, image feature, colour feature, scale-space, cue combination, multiple cues, multi-scale representation, computer vision.

1. Introduction

During recent years, the use of *view-based representations in terms of receptive field responses* (Koenderink and van Doorn [30]) has emerged as a highly promising paradigm for visual recognition. Three main groups of approaches have been developed, where receptive field responses are computed either (i) *globally or regionally in terms of histograms* (Swain and Ballard [56], Schiele and Crowley [52], Zelnik and Irani [64], Linde and Lindeberg [38], Dalal and Triggs [14], (ii) *locally at sparse sets of interest points* (Lowe [45], Bay *et al.* [3], Csurka *et al.* [13], Laptev and Lindeberg [33]), or (iii) *locally at every point or at a dense grid of points* (Hall *et al.* [25], Jurie and Triggs [27], Lazebnik *et al.* [34], Bosch *et al.* [5]).

The goal of this paper is to address the area of histogram-based recognition and to show that there exist *composed complex-cue image descriptors*, some of these of higher dimensionality, that capture more of the information content in the local image structure and give significantly better performance compared to previously used histogram features of lower dimensionality. Previous work (except our initial work in [38]) has mainly been based on histograms in two, three or six dimensions, obtained by accumulating the statistics of either colour channels (Swain and Ballard [56]), first-order derivatives alternatively lower-order differential invariants (Schiele and Crowley [52]), gradient directions (Lowe [45]; Dalal and Triggs [14]) or wavelet coefficients (Schneiderman and Kanade [53]).

From the view-point of local image measurements, a histogram of first-order partial derivatives captures the full first-order local grey-level structure, a histogram of the gradient magnitude and the Laplacian captures the first-order rotationally invariant structure as well as one component of the second-order rotationally invariant grey-level structure. Similarly, a histogram of local gradient directions captures other partial aspects of the first-order structure, whereas an RGB colour histogram can be seen as capturing the full zero-order colour structure. Hence, each one of these descriptors captures *partial aspects* of the information that is available in the local image structure, while suppressing or ignoring other aspects.

For the purpose of image-based recognition, it is therefore of interest to consider image descriptors that reflect more of the local information content. For example, if we consider local differential structure up to order two, there is additional information to be gained by considering complementary functionally independent differential invariants, as well as by using higher-order derivatives of grey-level information and spatio-chromatic derivates and differential invariants derived from colour cues. A main subject of this article is to explore the applicability of such more complex image descriptors for object recognition, and to *perform a systematic evaluation of the relative performance of different types of histogram-based image descriptors on established image data sets*.

The formulation of such composed image descriptors, however, implies that the histograms may be of higher dimensionality than previously used histogram features, which means that some precautions must be taken. To handle such higher-dimensional histograms, we will present an *efficient sparse method for computing and processing higher-dimensional histograms*. Based on this, we will focus on comparing histogram-based recognition schemes for a rich set of image descriptors that combine Gaussian derivative operators and differential invariants computed from either grey-level images, colour-opponent channels or both. It will be shown that there exist composed complex-cue histogram descriptors of higher dimensionality that have significantly better performance compared to the histogram features previously used by Swain and Ballard [56] and Schiele and Crowley [52], and that highly competitive recognition rates can be achieved. For simplicity of presentation, we shall in this paper restrict ourselves to global or regional histogram features that are accumulated over comparably large support regions. It seems plausible, however, that the general idea of composed complex-cue histograms should apply also to local histogram features computed at sparse interest points or at dense grids and we are investigating this issue in on-going work. With appropriate extensions, a similar formulation of composed histogram features in the spatio-temporal domain should furthermore be expected to be applicable for recognizing actions and events in video data.

Before starting, let us emphasize that the main purpose of this paper is *not* to present a new system for object recognition. Instead, the main purpose is to perform an extensive investigation of what aspects of the information content in the local image structure provide the most discriminative cues for image-based recognition, based on a complete classification of local receptive field like image operations up to order two as can be obtained from scale-space theory [43]. The intention is that these results could then be used as a guide and reference for future research in image-based recognition.

2. Related work

In this survey, we give an overview of work that involves image features and image descriptors defined from *receptive field like image operations at multiple scales*, in the closely related areas of spatial recognition and spatio-temporal recognition.

Swain and Ballard [56] initiated a direction of research on *histogram-based recognition methods* by showing how reasonable performance of an object recognition scheme could be obtained by comparing RGB colour histograms. Schiele and Crowley

[52] generalized this idea to histograms of receptive fields (Koenderink and van Doorn [30]) and computed histograms of either first-order Gaussian derivative operators or the gradient magnitude and the Laplacian operator at three scales, leading to 6-D histograms. Schneiderman and Kanade [53] showed that efficient recognition of faces and cars could be performed from histograms of wavelet coefficients. Linde and Lindeberg [38] presented a set of composed histogram descriptors of higher dimensionality that lead to better recognition performance compared to previously used receptive field like histogram descriptors.

With regard to approaches based on *interest points*, Mikolajczyk and Schmid [46] combined the ideas of scale selection and scale-adapted image features (Lindeberg [42]) with affine normalization (affine shape adaptation) (Lindeberg and Gårding [44]) to construct scale invariant and affine invariant Harris operators for object recognition. Baumberg [2] as well as Tuytelaars and van Gool [58] developed methods for affine normalisation of interest point for image matching based on closely related ideas. Lowe [45] combined the ideas of feature based and histogram based image descriptors, and defined a scale invariant feature transform, SIFT, which integrates the accumulation of statistics of gradient directions in local neighbourhoods of scale adapted interest points with summarizing information about the spatial layout. Bay *et al.* [3] presented an alternative approach with SURF features that are instead expressed in terms of Haar wavelets. Dalal and Triggs [14] extended the local SIFT descriptor to the accumulation of regional histograms of gradient directions over larger support regions. Other closely related probabilistic methods have been presented by Fergus *et al.* [18], and Ke and Suktankar [28]. An evaluation and comparison of several spatial recognition methods has been presented by Mikolajczyk and Schmid [47]. Dense local approaches have in turn been investigated by Jurie and Triggs [27], Lazebnik *et al.* [34], Bosch *et al.* [5] and Agarwal and Triggs [1].

Histogram-based methods for recognizing *spatio-temporal events* have been developed by Chomat *et al.* [12], Zelnik-Manor and Irani [64], Laptev and Lindeberg [33], Shechtman and Irani [54] and Willems *et al.* [61].

Regarding *colour-based object recognition*, Slater and Healey [55] presented histogram-like descriptors that combine spatial moments with colour information. Gevers and Smeulders [22] investigated the sensitivity of different zero-order colour spaces for histogram-based recognition. Geusebroek *et al.* [21] proposed a set of differential colour invariants that are invariant to illumination based on a reflectance model and the Gaussian colour model proposed by Koenderink. Hall *et al.* [25] computed partial derivatives of colour-opponent channels, leading to an N-jet representation up to order one. Linde and Lindeberg [38] extended this idea by showing that highly discriminative image descriptors for object recognition can be obtained from histograms of spatio-chromatic differential invariants up to order two defined from colour-opponent channels. Burghouts and Geusebroek [7] showed that the performance of the SIFT descriptor can be improved by complementing it with a set of colour invariants. More recently, van de Sande *et al.* [59] have presented an evaluation of different colour-based image descriptors for recognition.

An early study of the sensitivity of histogram descriptors to view variations was done by Schiele and Crowley [51]. A theory for modelling multi-scale histograms has been presented by Koenderink and van Doorn [31].

Somewhat related to our approach [38], Broadhurst [6] proposed a texture classification method based on histograms of Gaussian derivative filters of different orders and at different orientations, from which marginal distributions were estimated and assumed to be independent. In our work, we on the other hand explicitly register the co-variation of different image features as obtained from multi-dimensional histograms.

3. Image features and methodology

As a background for defining different types of image descriptors for histogram-based recognition, we will in this section summarize some main results concerning multi-scale image measurements that we will build upon, with emphasis on the types of local image information that can be captured from derivatives up to order two. Then, we will turn to the experimental methodology that will be used for comparing different types of image descriptors.

3.1. Scale-space features

Scale-space theory [62, 29, 30, 41, 40, 19, 57, 43] provides a well-founded framework to formulate local image operations at different scales. Given any image f, its *scale-space representation* L is obtained

by convolution with Gaussian kernels $g(x, y;\ \sigma) = 1/(2\pi\sigma^2)\, e^{-(x^2+y^2)/2\sigma^2}$ of different widths (as measured by their standard deviations σ):

$$L(\cdot, \cdot;\ \sigma) = g(\cdot, \cdot;\ \sigma) * f(\cdot, \cdot). \tag{1}$$

From this representation, *Gaussian derivatives*

$$L_{x^\alpha y^\beta}(\cdot, \cdot;\ \sigma) = \partial_{x^\alpha y^\beta} L(\cdot, \cdot;\ \sigma) \tag{2}$$

are computed and scale normalized derivatives [42] (with $\gamma = 1$) defined according to

$$L_{\xi^\alpha \eta^\beta}(\cdot, \cdot;\ \sigma) = \sigma^{\alpha+\beta}\, L_{x^\alpha y^\beta}(\cdot, \cdot;\ \sigma) \tag{3}$$

where α and β denote the order of differentiation. In this study, we focus on spatial derivatives up to order 2, obtained from the second-order N-jet:

$$\begin{aligned}&(L_\xi(\cdot, \cdot;\ \sigma), L_\eta(\cdot, \cdot;\ \sigma),\\ &\ L_{\xi\xi}(\cdot, \cdot;\ \sigma), L_{\xi\eta}(\cdot, \cdot;\ \sigma), L_{\eta\eta}(\cdot, \cdot;\ \sigma)).\end{aligned} \tag{4}$$

This set of partial derivatives constitutes a *complete basis* for all local differential image operations up to order two at any image point.

The top two rows in figure 1 show the result of computing such Gaussian derivatives from the grey-level information in an image.

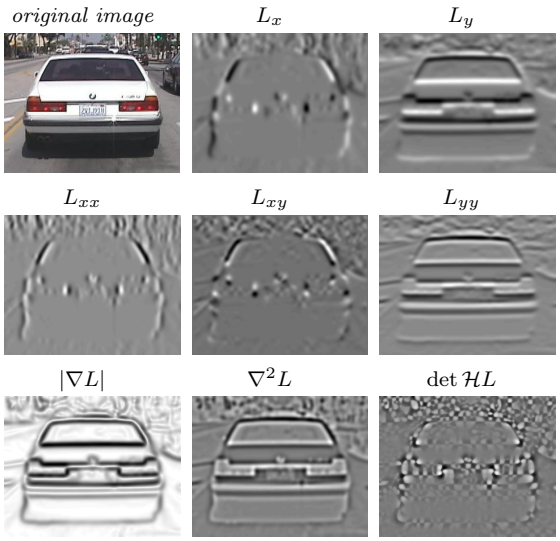

Figure 1: Partial derivatives and differential invariants up to order two computed at scale $\sigma = 4$ from the grey-level information in an image.

3.1.1. Scale-normalized differential invariants

From the above mentioned partial derivatives expressed in a Cartesian reference frame, we consider differential expressions that are invariant to rotations in the image plane. We will mainly use the scale normalized *gradient magnitude* $|\nabla_{norm} L|$, the scale normalized *Laplacian* $\nabla^2_{norm} L$ and the scale normalized *determinant of the Hessian* $\det \mathcal{H}_{norm} L$:

$$\begin{aligned}|\nabla_{norm} L| &= \sqrt{L_\xi^2 + L_\eta^2} = \sigma \sqrt{L_x^2 + L_y^2},\\ \nabla^2_{norm} L &= L_{\xi\xi} + L_{\eta\eta} = \sigma^2 (L_{xx} + L_{yy}),\\ \det \mathcal{H}_{norm} L &= L_{\xi\xi} L_{\eta\eta} - L_{\xi\eta}^2 = \sigma^4 (L_{xx} L_{yy} - L_{xy})^2.\end{aligned} \tag{5}$$

The gradient magnitude is the only rotationally invariant differential expression of first order, while the Laplacian and the determinant of the Hessian together span the space of second-order rotational invariants. These differential expressions therefore span three out of the four possible rotational invariants up to order two [43]. A complete set of differential invariants up to order two can be obtained by including a fourth degree of freedom in terms of the normalized rescaled level curve curvature $\tilde{\kappa}_{norm} = L_x^2 L_{yy} + L_y^2 L_{xx} - 2 L_x L_y L_{xy}$, which is a combination of first- and second-order derivatives and corresponds to the product of the curvature of level curves κ with the gradient magnitude $|\nabla_{norm} L|$ raised to the power of three. In initial experiments, we have, however, not found complementary use of $\tilde{\kappa}_{norm}$ to improve the performance, why we do not include this remaining degree of freedom in the presentation.

The bottom row in figure 1 show the result of computing these differential invariants from grey-level information.

3.1.2. Spatio-chromatic derivatives and differential invariants

For colour images, we also define chromatic cues $(c^{(1)}, c^{(2)})$ from RGB images by red/green and yellow/blue colour-opponent channels according to [25]

$$\begin{pmatrix} f \\ c^{(1)} \\ c^{(2)} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & -1 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \tag{6}$$

from which a corresponding *colour-opponent scale-space representation* $(C^{(1)}, C^{(2)})$ is defined by Gaussian convolution

$$C^{(1)}(\cdot, \cdot;\ t) = g(\cdot, \cdot;\ t) * c^{(1)}(\cdot, \cdot), \tag{7}$$

$$C^{(2)}(\cdot, \cdot;\ t) = g(\cdot, \cdot;\ t) * c^{(2)}(\cdot, \cdot). \tag{8}$$

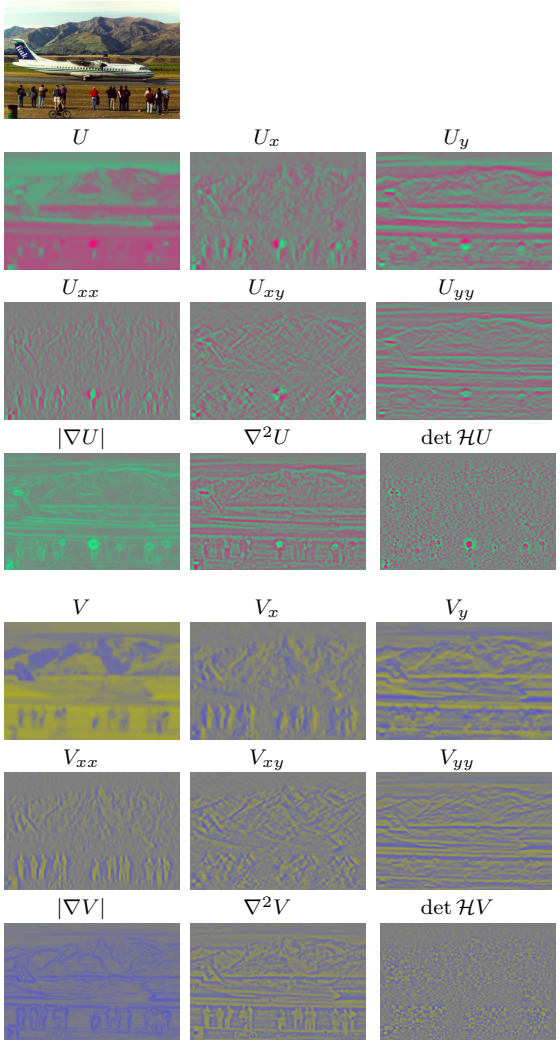

Figure 2: Spatio-chromatic derivatives and differential invariants up to order two computed at $\sigma = 4$ from the colour-opponent channels $U = C^{(1)}$ and $V = C^{(2)}$ of a colour image.

Then, *spatio-chromatic derivatives and differential invariants in colour-opponent scale-space* are defined in a corresponding manner as from grey-levels:

$$\begin{aligned}
(C_\xi, C_\eta) &= (C_\xi^{(1)}, C_\xi^{(2)}, C_\eta^{(1)}, C_\eta^{(2)}) \\
|\nabla_{norm} C| &= (|\nabla_{norm} C^{(1)}|, |\nabla_{norm} C^{(2)}|), \\
(C_{\xi\xi}, C_{\xi\eta}, C_{\eta\eta}) &= (C_{\xi\xi}^{(1)}, C_{\xi\xi}^{(2)}, C_{\xi\eta}^{(1)}, C_{\xi\eta}^{(2)}, C_{\eta\eta}^{(1)}, C_{\eta\eta}^{(2)}) \\
\nabla^2_{norm} C &= (\nabla^2_{norm} C^{(1)}, \nabla^2_{norm} C^{(2)}), \\
\det \mathcal{H}_{norm} C &= (\det \mathcal{H}_{norm} C^{(1)}, \det \mathcal{H}_{norm} C^{(2)}).
\end{aligned} \tag{9}$$

Whereas we do not exclude the possibility of also defining image descriptors based on other types colour features, the motivations for using these spatio-chromatic derivatives and differential invariants as basic features in this work are that (i) a representation in terms of colour-opponent channels is chosen because of its similarity to biological vision, where a separation into red/green and yellow/blue colour-opponent channels is performed at a very early stage in the visual pathway (see *e.g.* Goldstein [23, pages 197–198]) and (ii) the spatio-chromatic derivatives and differential invariants are computed from the colour-opponent channels in a similar way as grey-level derivatives and differential invariants are computed from intensity information, without need for any illumination model.[1]

Figure 2 shows the result of computing these spatio-chromatic entities from a colour image.

3.1.3. Rotationally invariant vs. rotationally variant image features

In the choice between image features based on partial derivatives *vs.* rotationally invariant differential invariants, the latter imply that the corresponding image descriptors will automatically be invariant to rotations in the image plane. Rotationally variant image features may on the other hand capture more of the local information content in the image data, since information about local image orientations may be lost in the definition of rotationally invariant image features. To investigate the relative advantages of these alternative approaches, we will in this work specifically investigate how different types of rotationally invariant *vs.* rotationally variant image descriptors perform on different types of benchmark problems.

[1] Concerning the choice of colour representation, this colour-opponent model implies that both the grey-level values L and the chromatic channels (U, V) are proportional to luminance. With both the (grey-level and spatio-chromatic) Gaussian derivatives and the differential invariants transformed to be linear in terms of luminance (after the transformation of the determinant of the Hessian in equation (10) performed next), this means that both the image features and the principal components to be computed from these in section 5.1 form a cone under illumination variations.

As will be shown in later sections, for higher-dimensional histogram descriptors it can in many cases be sufficient to use as few as $r = 2$ to $r = 5$ bins per dimension. To a major extent, the corresponding quantized image descriptors will therefore not be very much affected by uniform rescalings with scaling factors close to one, which in combination with automatic exposure control of the camera will imply certain robustness properties with respect to illumination variations.

3.1.4. Dependency on intensity and chromaticity transformations

When choosing image features as the basis for a computer vision system, it is also essential to consider the behaviour of the image features under illumination transformations. Under *uniform rescalings* of the intensity and chromaticity values in the original signal, $f'(x, y) = a\, f(x, y)$ and $c'(x, y) = a\, c(x, y)$, as implied *e.g.* by a uniform rescaling of the RGB values, corresponding Gaussian smoothed intensities L and colour-opponent channels C are transformed by the same scaling factor, as are all Gaussian derivatives L_{x^α} and C_{x^α}, gradient magnitudes $|\nabla L|$ and $|\nabla C|$ as well as the Laplacian responses $\nabla^2 L$ and $\nabla^2 C$. The determinants of the Hessians $\det \mathcal{H}L$ or $\det \mathcal{H}C$ are, however, not transformed in the same way, unless we transform the latter entities to be proportional to the image intensities, which can be achieved by a transformation of the form

$$h(\det \mathcal{H}L) = \mathrm{sign}(\det \mathcal{H}L) \sqrt{|\det \mathcal{H}L|}. \qquad (10)$$

Under *additive intensity transformations* of the original signal, $f'(x, y) = a_f + f(x, y)$ and $c'(x, y) = a_c + c(x, y)$, the values of the explicit zero-order image intensities L and chromaticity channels C are transformed by the same offset, whereas the value of any Gaussian derivative L_{x^α} or C_{x^α} will be unaffected, and so will any differential invariant be that is defined in terms of Gaussian derivatives. Second-order differential operators are also preserved under *linear illumination gradients*, $f'(x, y) = f(x, y) + a_f\, x + b_f\, y$ and $c'(x, y) = c(x, y) + a_c\, x + b_c\, y$. In these respects, higher-order differential features can be expected to be more robust to illumination variations than zero-order cues.

3.2. Histogram descriptors

For every image point (x, y), a set of N different scalar image features will be computed, resulting in an N-dimensional histogram. Each feature dimension is quantized into r bins, determined from the maximum and minimum values of the feature values over the data set. The Gaussian derivatives $L_{x^\alpha y^\beta}$ as well as the differential invariants $|\nabla L|$ and $\nabla^2 L$ are quantized on a linear scale, while the determinant of the Hessian $\det \mathcal{H}L$ is transformed to a parameterization proportional to image intensities (10) prior to quantization. Concerning the quantization step, one could conceive using different numbers of bins r for different types of scale-space features. Since we will perform extensive comparisons for different values of r, however, we will here use the same number of bins for all the image features in a histogram, to delimit the search space.

3.3. Efficient handling of higher-dimensional histograms

A histogram of an N-dimensional set of image features that are quantized using r bins for each feature dimension may contain up to r^N cells. With N set to the maximum dimensionality of our image descriptors $N = 35$ and $r = 10$, this means that an N-dimensional histogram may contain up to the order of 10^{35} cells. Usually, however, a vast majority of these cells will be empty. Specifically, the number of non-zero entries will always be limited by the number of pixels in the image, and if we consider histograms of the scale-space features in section 3.1 computed from images of man-made or natural environments, the image measurements will usually be clustered to a significantly smaller subset of histogram bins. From tables 2, 6 and 7 to be presented later, it can be seen that for the image descriptors and datasets considered here, the number of non-zero cells n is usually in the range between 300 and 60 000. The receptive field based image measurements can therefore be interpreted as being localized to a submanifold of much, much lower dimensionality than a worst-case estimate would give. Hence, it is natural to handle higher-dimensional histograms using a *sparse representation*.

For an N-dimensional histogram, with $r_1, \ldots, r_N$ quantization levels for the N different dimensions, we enumerate the cells from 0 to $\left(\prod^D r_i\right) - 1$, such that a quantized local measurement $M = (m_1, \ldots, m_N)$, $0 \leq m_i < r_i$, will correspond to the cell with index

$$c = \sum_{i=1}^{N} \left(m_i \prod_{j=1}^{i-1} r_j \right). \qquad (11)$$

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, it takes about 1.5 ms for a single-threaded process to accumulate a sixteen-dimensional histogram of a 256×256 image on a 2.66 GHz Intel Core i7$^{\text{TM}}$processor.

3.4. Dissimilarity measures and classification

To measure the similarity between histograms h_1 and h_2 of images from the training and test data, respectively, we use the χ^2-measure:

$$\chi^2(h_1, h_2) = \sum_i \frac{(h_1(i) - h_2(i))^2}{h_1(i) + h_2(i)} \qquad (12)$$

(where terms with $h_1(i)+h_2(i) = 0$ are set to zero). Based on such χ^2-comparisons, we use a support vector machine[2] for comparing higher-dimensional histogram descriptors, following the approach by [11, 4, 9] with a kernel of the form

$$K(h_1, h_2) = \exp(-\gamma \chi^2(h_1, h_2)). \qquad (13)$$

From initial experiments, we set $\gamma = 1$. The penalty constant C on over-training was set to $C = 100$. No other optimization was made of the SVM. The histograms were normalized to unit mass prior to SVM-based training and recognition. The actual implementation of the SVM was done based on the libSVM software [10].

To handle $k > 2$ classes, we use the one-against-one method [20, 32], which constructs $k(k-1)/2$ classifiers, with each one trained to separate two different classes. When deciding which of the k classes an unknown sample belongs to, it is tested on all $k(k-1)/2$ classifiers, and each classifier votes for one of two classes. The class with the greatest vote is regarded as the predicted class.

3.5. Benchmark problems: Object recognition and object classification

In this work, we shall investigate the performance of image descriptors for two types of recognition problems: (i) recognition of known objects from previously unseen views of a set of known objects and (ii) classification of previously unseen objects, given images of similar objects from the same category. For these purposes, we will make use of the COIL-100, ETH-80 and CalTech-4 data sets.

[2]In our pre-study [38] we showed that histogram classification with a kernel-based support vector machine performs significantly better than nearest-neighbour classification.

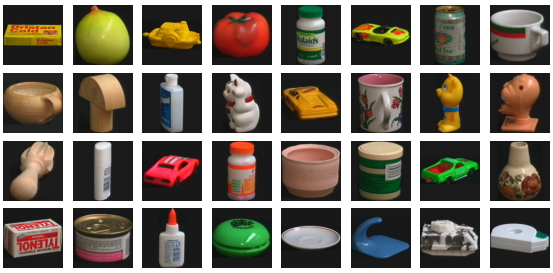

Figure 3: Sample images from the COIL-100 dataset.

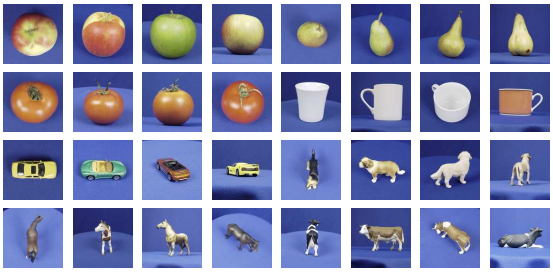

Figure 4: Sample images from the ETH-80 dataset.

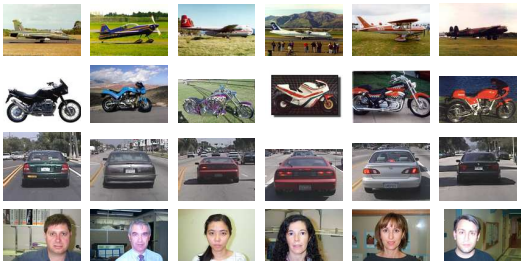

Figure 5: Sample images from the CalTech-4 dataset.

The Columbia Object Image Library, COIL-100 [48] consists of colour images of 100 objects with 72 images per object, corresponding to viewpoints obtained by rotating the objects through 360° at 5° increments (see figure 3). For the task of object instance recognition, we will use training data with differences in viewing directions of either 30° or 60°. Then, the intermediate views middle in between will be used in the recognition step, which means that the differences in viewing direction between the training and testing data will be 15° or 30°, respectively. These experiments are then repeated with different offsets in the viewing direction, *i.e.*, 5 more times for a 30° viewsphere sampling and 11 more times with a 60° viewsphere sampling.

The ETH-80 dataset [35] contains images of 80 objects from 8 categories (apples, pears, tomatoes, cups, cars, dogs, horses and cows; see figure 4). For this data set we consider both a recognition task,

which will be regarded as correct if the system recognizes the specific object instance, *i.e.*, a specific car or a specific apple, and a classification problem in which all images of the test object are removed. For the classification task, a result is regarded as correct if the classification leads to the same object category as the test object, with all the views of the test object removed from the training set.

For the object recognition task on the ETH-80 dataset, we will use training data with differences in viewing directions of either 45° or 90°, implying differences in viewing directions between the training data and the test data of 22.5° or 45°, respectively. These experiments are repeated once more with a different offset in the case of a 45° viewsphere sampling and for three different offsets with 90° differences. For the object classification task, however, the experiment does not explicitly test for differences in viewing direction, since the training set contains images of other objects from the object category with similar viewing directions.

For the images in the ETH-80 dataset, the background pixels were set to zero, motivated by an initial companion study for local histogram descriptors [36, 37], which shows that there are background cues in the images in this dataset that may simplify the task of classifying objects into the predefined categories. By enforcing a black background, we suppress the influence of such background cues.

The CalTech-4 dataset is a subset of the CalTech dataset [18], which contains images from 4 object categories: 800 motorcycles, 800 air-planes, 435 faces and 800 cars (rear) (see figure 5). The images have varying resolution and varying complex backgrounds in real environments. This dataset is used for object categorization, where it is divided into a training set and a testing set of equal size.

The reasons why we have chosen to use these datasets for evaluating image descriptors in this work are that the COIL-100 and ETH-80 datasets contain images of objects taken with a rather dense set of viewing directions and that the CalTech-4 dataset contains a large number of objects from each object category, with only one object in each image. This means that the importance of low-level image cues can be expected comparably higher than for *e.g.* the CalTech-101 [17] or the PASCAL [16] datasets. Specifically, we avoid the localization/segmentation issue, which is important for this study, where the focus is on evaluating and comparing the information content in different types of image descriptors.

4. Primitive histogram descriptors

The scale-space features listed in section 3.1 constitute a complete basis for the types of local image operations that can be defined at any image point from either image intensities or colour-opponent channels. Specifically, we can classify the different types of image features depending upon (i) whether they are based on either pure grey-level cues or spatio-chromatic cues, (ii) whether they are rotationally invariant or rotationally variant, and (iii) the maximum order of differentiation; see table 1 for a classification of the types of image features that we will consider.

Rotationally invariant features		
Order	Grey-level	Spatio-chromatic
0	L	C
1	$\|\nabla L\|$	$\|\nabla C\|$
2	$\nabla^2 L$, $\det \mathcal{H} L$	$\nabla^2 C$, $\det \mathcal{H} C$

Rotationally variant features		
Order	Grey-level	Spatio-chromatic
1	(L_x, L_y)	(C_x, C_y)
2	(L_{xx}, L_{xy}, L_{yy})	(C_{xx}, C_{xy}, C_{yy})

Table 1: Basic types of scale-space features used for defining histogram descriptors.

In this section, we shall investigate the information content in such scale-space features, by computing the image features at a set of scales in scale-space, and then defining multi-dimensional histograms of image features over these scales. For simplicity, we shall restrict ourselves to a set of fixed scale levels and all image primitives will be computed at some subset of the scale levels $\sigma \in \{1, 2, 4, 8, 16\}$. For the COIL-100 and ETH-80 datasets, this choice is reasonable, since there are no major scale variations in these datasets. For the CalTech-4 dataset, we will also make use of a fixed set of scale levels, although the presence of scale variations in this dataset could warrant the use of an explicit scale selection mechanism. The situation with the CalTech-4 dataset is, however, improved by the fact that both the training set and the test set contain similar ranges of scale variations. In this way, we can avoid the scale selection problem and focus on evaluating the information content in different types of image descriptors.

4.1. Evaluation of primitive histogram descriptors

Table 2 show the result of computing the performance of two different types of primitive histogram

Image features		30°	COIL-100		45°	ETH-80		ETH	Class.	
		60°	r	n	90°	r	n	CalTech	r	n
3-D	$(L_{xx}, L_{xy}, L_{yy}$;	95.6	55	465	96.3	36	760	93.5	80	2644
	$\sigma = 1)$	84.5	55	460	87.0	60	1798	*99.5*	140	19247
3-D	$(L_{xx}, L_{xy}, L_{yy}$;	95.0	36	374	97.7	36	1099	92.1	16	213
	$\sigma = 2)$	85.8	55	798	89.5	50	1990	99.1	40	2406
6-D	$(L_{xx}, L_{xy}, L_{yy}$;	96.6	36	1787	97.2	19	1766	96.0	30	4188
	$\sigma = 1, 2)$	89.5	25	774	90.0	29	3756	**99.8**	24	11117
6-D	$(L_{xx}, L_{xy}, L_{yy}$;	*97.7*	25	1000	**98.6**	15	1383	96.3	30	5813
	$\sigma = 1, 4)$	89.8	21	671	89.5	30	5996	*99.5*	24	14078
6-D	$(L_{xx}, L_{xy}, L_{yy}$;	96.8	20	923	**98.5**	17	2039	95.9	14	1448
	$\sigma = 2, 4)$	88.8	15	404	89.5	15	1558	*99.4*	20	7616
9-D	$(L_{xx}, L_{xy}, L_{yy}$;	**97.8**	15	756	*98.4*	15	3133	96.1	14	4516
	$\sigma = 1, 2, 4)$	**90.5**	17	989	**91.4**	17	4012	*99.5*	10	6805
12-D	$(L_{xx}, L_{xy}, L_{yy}$;	95.7	12	4648	*98.3*	15	5584	95.9	2	5245
	$\sigma = 1, 2, 4, 8)$	89.8	19	2388	89.6	21	9103	*99.5*	6	11055
15-D	$(L_{xx}, L_{xy}, L_{yy}$;	95.0	2	5731	97.0	7	1579	**96.9**	2	11884
	$\sigma = 1, 2, 4, 8, 16)$	87.3	13	1587	86.9	15	7925	99.2	12	86939
2-D	$(\nabla^2 C$;	97.2	195	344	98.6	255	662	78.4	195	438
	$\sigma = 1)$	92.0	125	173	92.3	195	440	91.7	255	598
2-D	$(\nabla^2 C$;	98.7	155	390	98.5	255	731	77.9	100	166
	$\sigma = 2)$	94.6	125	275	93.8	255	737	93.4	255	1336
4-D	$(\nabla^2 C$;	*99.1*	125	1924	98.5	60	651	**81.7**	30	191
	$\sigma = 1, 2)$	95.1	95	1210	95.3	155	3660	96.9	195	7346
4-D	$(\nabla^2 C$;	**99.3**	100	2042	**99.6**	195	6414	**81.8**	75	1102
	$\sigma = 1, 4)$	*95.9*	75	1209	*96.2*	125	3099	96.3	155	8957
4-D	$(\nabla^2 C$;	**99.2**	80	1563	98.7	60	724	*81.6*	95	1707
	$\sigma = 2, 4)$	**96.2**	100	2169	**96.6**	195	6290	96.9	255	27712
6-D	$(\nabla^2 C$;	*99.1*	95	3974	99.1	60	2720	75.5	2	128
	$\sigma = 1, 2, 4)$	*96.0*	75	2786	95.1	75	3313	*97.0*	195	39846
8-D	$(\nabla^2 C$;	**99.2**	36	2586	98.8	30	2195	76.2	80	10231
	$\sigma = 1, 2, 4, 8)$	*96.0*	40	2945	93.9	75	6967	**97.5**	75	21070

Table 2: The performance of two examples of *primitive histogram descriptors* for the 6 different benchmark problems described in section 3.5 and using different combinations of scale levels for computing the scale-space features. For each experiment, the number of bins r that gave the best performance is shown, as is the average number n of non-zero bins n for the corresponding histograms. The first two columns show results for object instance recognition, while the third column shows results for object category classification. Corresponding results for a larger set of primitive histogram descriptors are presented in [39].

descriptors obtained by combining the responses of single grey-level and spatio-chromatic cues at multiple scales for the image features (L_{xx}, L_{xy}, L_{yy}) and $\nabla^2 C$. Corresponding results for a larger set of grey-level features L, (L_x, L_y), $|\nabla L|$, (L_{xx}, L_{xy}, L_{yy}), $\nabla^2 L$ and $\det \mathcal{H} L$ as well as the colour-opponent features C, (C_x, C_y), $|\nabla C|$, (C_{xx}, C_{xy}, C_{yy}), $\nabla^2 C$ and $\det \mathcal{H} C$ are presented in [39].

Results from six types of recognition problems are reported: (i) recognition of individual objects in the COIL-100 dataset using 30° and 60° differences in the viewing direction, (ii) recognition of individual objects in the ETH-80 dataset using 45° and 90° differences in the viewing direction, and (iii) classification of previously unseen objects into object categories in the ETH-80 and CalTech-4 datasets (without additional variations in the viewing direction). To investigate the influence of the choice of scale levels on the performance, we show results for different subsets of the scale levels $\sigma \in \{1, 2, 4, 8, 16\}$ for each type of scale-space feature. In addition to the actual performance values p, we also show the number of histogram bins r that leads to the best result under variations of the bin size as well as the average number of non-zero bins n for the corresponding histograms over all images in the dataset. To emphasize which descriptor performs best within each group of similar descriptors, the best result is marked in boldface, while a few results just below the best are indicated with slanted font.

As can be seen from table 2 as well as the more extensive results in [39], the combination of scale levels $\sigma \in \{1, 2, 4\}$ lead to results that are generally good for the histogram descriptors that are based on pure grey-level features. For chromatic features, good results are often obtained at the scale level

$\sigma = 2$ as well as the pairs of scale levels $\sigma \in \{1, 2\}$ and $\sigma \in \{1, 4\}$.[3]

Image features	COIL-100		ETH-80		Classification	
	V	Rate	V	Rate		Rate
L	30°	98.4	45°	99.0	ETH	91.3
	60°	89.9	90°	92.4	CT	97.5
L_x, L_y	30°	98.5	45°	*99.1*	ETH	*93.9*
	60°	92.2	90°	92.1	CT	*99.6*
$\|\nabla L\|$	30°	99.1	45°	97.0	ETH	90.9
	60°	92.0	90°	90.8	CT	98.0
L_{xx}, L_{xy}, L_{yy}	30°	97.8	45°	98.6	ETH	**96.9**
	60°	90.5	90°	91.4	CT	**99.8**
$\nabla^2 L$	30°	98.9	45°	97.4	ETH	*94.0*
	60°	92.5	90°	90.4	CT	98.9
$\det \mathcal{H}L$	30°	98.8	45°	94.9	ETH	92.0
	60°	91.9	90°	85.2	CT	*99.2*
C	30°	**99.7**	45°	**99.8**	ETH	76.4
	60°	**97.8**	90°	**98.7**	CT	96.5
C_x, C_y	30°	99.0	45°	98.5	ETH	85.5
	60°	95.9	90°	94.5	CT	98.6
$\|\nabla C\|$	30°	*99.2*	45°	98.8	ETH	80.5
	60°	*96.4*	90°	*95.7*	CT	96.5
C_{xx}, C_{xy}, C_{yy}	30°	98.1	45°	98.2	ETH	87.9
	60°	94.0	90°	94.0	CT	98.5
$\nabla^2 C$	30°	*99.3*	45°	*99.6*	ETH	81.8
	60°	*96.2*	90°	96.6	CT	97.5
$\det \mathcal{H}C$	30°	98.4	45°	97.7	ETH	86.6
	60°	94.9	90°	92.2	CT	97.0

Table 3: Summary of the best recognition scores for each type of primitive histogram descriptor in table 2 as well as a more extensive evaluation in [39] under variations of the sets of scale levels used for computing the histograms.

Table 3 gives a more compact summary of the results from this initial evaluation, by showing the best results for each type of primitive histogram descriptor under variations of the scale levels. From this table we can see that discriminative cues for object recognition can be obtained from a variety of local scale-space features. Notably, for these datasets the strongest single histogram cues for recognizing object instances are obtained from chromatic and spatio-chromatic information, specifically the colour-opponent channels C, the colour-opponent gradient magnitudes $|\nabla C|$ and the colour-opponent Laplacians $\nabla^2 C$. For classifying previously unseen objects into categories, however, the strongest single histogram cues are obtained from pure grey-level information, in particular the second-order partial derivatives (L_{xx}, L_{xy}, L_{yy}), the first-order partial derivatives (L_x, L_y), and the second-order differential invariants $\nabla^2 L$ and $\det \mathcal{H}L$.

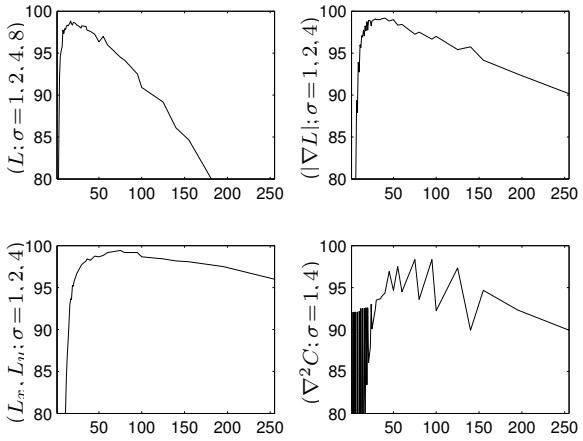

Figure 6: Dependency of the recognition rate on the number of bins per dimension, r, when recognizing object instances using different types of primitive histogram descriptors applied to reduced subsets of the COIL-100 data set. Each graph has been computed for the combination of scale levels that gave the best performance according to table 2 and the complementary results in [39].

4.2. Dependency on the number of histogram bins

Figure 6 gives a more detailed study of how the recognition performance depends upon the number of bins, r, in the histogram for a subset of lower-dimensional histograms, for the problem of recognizing specific object instances on reduced subsets of the COIL-100 dataset. As can be seen from the graphs, the recognition rates can be strongly dependent on the number of bins used for accumulating the histograms. The choice of bin width is therefore an important parameter for histogram-based recognition methods. To avoid having the comparisons between image descriptors biased by a specific choice of r, we have therefore decided to perform experiments for a wide range of bin sizes and then comparing the performance of different types of image descriptors using the bin size that gives the best result for each image descriptor.

A particular observation that can be made here is that some of the graphs have a saw-toothed shape. Such phenomena occur for polarity dependent image descriptors, such as partial derivatives, Laplacians and determinants of Hessians. The reason for this is that depending on whether the number of bins is odd or even, there is a qualitative difference

[3] These scale levels have been selected from experiments with the COIL-100 and ETH-80 data sets. Depending on the conditions under which the image data are acquired, one could conceive that other combinations of scale levels could be better for other data sets. In a general setting, it is therefore natural to consider using *e.g.* automatic scale selection [42]. The restriction to these scale levels is made here solely to reduce the number of degree of freedom in the study.

in the handling of values close to zero. With an even number of bins, values close to zero fall in different bins depending on the sign of the differential expression. If we use an odd number of bins, all values close to zero fall into the same bin.

5. Composed complex-cue histograms

In previous section, we investigated a set of primitive image descriptors formulated in terms of histograms of receptive field like responses defined from partial derivatives and differential invariants up to order two, computed from either grey-level images or colour-opponent channels. The results showed that important cues for object recognition can be obtained from a variety of different low-level cues. Specifically, we can expect different types of performance depending on the order of the image descriptors, whether the image descriptors are rotationally invariant or not, and if the recognition is based on either pure shape information or combinations between shape and colour information.

A simplifying assumption that is commonly used in vision applications, is to consider image measurements as independent, thereby simplifying subsequent statistical analysis. With regard to histogram-based recognition, such an assumption corresponds to the computation of separated lower-dimensional histograms for each type of image descriptor. In this work, we are interested in investigating composed image descriptors of higher dimensionality, that explicitly reflect the co-variation of different types of primitive image measurements, so as to obtain more distinctive image measurements for recognition. To study this problem, we shall therefore complement the above mentioned general theory for multi-scale receptive field based image measurements with an empirical study of the performance of different non-separable combinations of lower-dimensional image primitives. Specifically, we shall in this section define a rich set of complex-cue histogram descriptors, usually of higher dimensionality, that are formed from combinations of lower-dimensional image measurements and then evaluate the performance of these image descriptors on our benchmark problems.

When defining such composed image descriptors, where some may be of rather high dimensionality (often between 10 and 15 dimensions, but sometimes as high as 25 or 35), this raises a general issue concerning how much image data are needed to produce sufficient statistics for comparing histograms. Specifically, one may ask how many of the histogram dimensions carry essential information, which would then reveal the *effective dimensionality of the histogram*. Hence, it is natural to complement the higher-dimensional histograms by some mechanism for dimensionality reduction.

The use of complementary dimensionality reduction can also be motivated from a machine learning perspective. Since the kernel-based support vector machine [11] we use for classifying the histograms is not guaranteed to preserve the theoretical optimality results that hold for a regular support vector machine [60], one may ask if the overall performance of object recognition or object classification can be improved by reducing the dimensionality of the histograms prior to learning and classification.

5.1. PCA-reduced histograms

Since the scale-space features, from which the histograms are defined, may assume values in different ranges, we start with an optional initial *variance normalization* of the image descriptors. The mean $m_{\mathcal{D}L}$ and the standard deviation $\sigma_{\mathcal{D}L}$ for each normalized scale-space feature $\mathcal{D}L$ are computed over a subset of the image data.[4] Then, variance normalized image features $\mathcal{D}_W L$ are defined as

$$\mathcal{D}_W L = \frac{\mathcal{D}L - m_{\mathcal{D}L}}{\sigma_{\mathcal{D}L}}. \tag{14}$$

Given a set of N normalized image features $D = \{(\mathcal{D}_W L)_n : n = 1..N\}$, the covariance matrix Σ of the normalized data is computed over the data set. Then, a singular value decomposition is computed of the covariance matrix

$$\Sigma = U \Lambda U^T \tag{15}$$

where U is orthogonal, $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$. Finally, a transformed feature subset F is defined by

$$F = U^T D, \tag{16}$$

the first $M \leq N$ components are extracted and an M-dimensional histogram is computed. Quantization is then performed over a range of 5 standard deviations in the transformed data and values beyond this range are truncated.

[4] For polarity dependent image features for which the expected value would be zero, such as the partial derivatives $(L_x, L_y, L_{xx}, L_{xy}, L_{yy})$ and $(C_x, C_y, C_{xx}, C_{xy}, C_{yy})$ or the differential invariants $(\nabla^2 L, \det \mathcal{H}L)$ and $(\nabla^2 C, \det \mathcal{H}C)$, the mean m is *set* to zero.

Image features		COIL-100				ETH-80			
		V	Rate	r	n	V	Rate	r	n
13-D	$(\lvert\nabla L\rvert, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4)$,	60°	97.6	12	2409	45°	99.2	8	2780
13-PC	$(\lvert\nabla C\rvert, \nabla^2 C; \sigma = 2)$	60°	98.5	3	2314	45°	**99.8**	2	5953
12-PC		60°	98.8	3	1903	45°	**99.8**	2	4015
10-PC		60°	**99.0**	3	1189	45°	99.5	2	1658
8-PC		60°	98.8	3	650	45°	99.3	2	510
6-PC		60°	98.4	3	306	45°	98.9	2	165
5-PC		60°	97.5	3	140	45°	98.8	3	383
4-PC		60°	96.8	3	84	45°	99.3	4	351
3-PC		60°	95.4	5	118	45°	98.8	7	511
2-PC		60°	91.3	15	213	45°	97.3	21	731
1-PC		60°	69.5	25	31	45°	82.7	155	239
7-D	$(\nabla^2 L; \sigma = 1, 2, 4, 8, 16)$,	60°	94.0	12	1145	45°	99.1	10	1915
7-PC	$(\nabla^2 C; \sigma = 2)$	60°	95.1	5	1421	45°	99.5	4	3228
6-PC		60°	95.5	5	804	45°	99.5	3	760
5-PC		60°	**95.7**	5	411	45°	**99.6**	4	822
4-PC		60°	95.2	5	224	45°	**99.6**	4	349
3-PC		60°	95.1	5	87	45°	99.0	5	228
2-PC		60°	90.4	13	127	45°	97.0	18	508
1-PC		60°	76.8	50	72	45°	90.5	125	274

Table 4: Dependency of the recognition performance on the number of principal components in PCA-reduced histograms of a non-linear and a linear histogram descriptor. The results show the recognition rates obtained for two of the object recognition benchmark problems, *i.e.*, COIL-100 with 60° viewsphere sampling and ETH-80 with 45° viewsphere sampling. The top row in each box shows results without variance normalization and without PCA. All PCA-reduced histograms have been computed with variance normalization. The number prior to PC in the leftmost column indicates the number of principal components.

Dependency on the number of principal components. Table 4 shows recognition rates obtained by performing object recognition using PCA-reduced histograms of (i) a 13-D non-linear image descriptor with $(|\nabla L|, \nabla^2 L, \det \mathcal{H} L)$ at three scales and $(|\nabla C|, \nabla^2 C)$ at one scale, and (ii) a 7-D linear image descriptor with $\nabla^2 L$ at five scales and $\nabla^2 C$ at one scale. For each descriptor the performance on the COIL-100 60° and ETH-80 45° object recognition problems have been computed for all numbers[5] $M \in \{1, 2, 3, 4, 5, 6, 8, 10, 12, 15, 20, 25, N\}$ of principal components between 1 and the dimensionality N of the descriptor. The table also shows the influence of variance normalization, by comparing the results of the N-dimensional variance normalized descriptor to the performance without using PCA. As can be seen from the tables, the use of variance normalization by itself can increase the performance of histogram-based recognition, although this normalization is not guaranteed to always have a positive effect. We can also clearly see that we obtain higher recognition performance for values of M lower than N. Notably, the optimized number of bins r is often close to a minimum for the number of principal components that give rise to the best performance. Thus, principal component analysis should also reduce the risk for over-training. These examples show that the performance of histogram-based recognition methods can be significantly improved by the use of dimensionality reduction prior to using a kernel-based support vector machine for recognition and learning.

Interpretation of the principal components. When computing PCA-reduced histograms, it is interesting to interpret the principal components F in equation (16) in terms of equivalent image operators. Such an interpretation is particularly illuminating for image features constructed from linear operations, such as the partial derivatives $(L_x, L_y, L_{xx}, L_{xy}, L_{yy})$ and $(C_x, C_y, C_{xx}, C_{xy}, C_{yy})$ and the differential invariants $\nabla^2 L$ and $\nabla^2 C$.

Table 5 shows the principal components of the 5-D histogram of the grey-level Laplacians $\nabla^2 L$ at five scales. As can be seen from the data, the principal components with indeces $m > 1$ correspond to differences between corresponding scale-space descriptors at different scales and can hence be interpreted as reflecting approximations of *derivatives with respect to scale*[6], $\delta_t \approx \partial_t$, with $t = \sigma^2$. The principal component u_1 is essentially the average

[5] In the experiments to be presented later, we investigate $M \in \{1 \ldots 12, 15, 20, 25, N\}$ principal components.

[6] In this respect, a multi-scale representation with linear features of order N at S adjacent scale levels can be seen as implicitly encoding information up to order $N + 2(S - 1)$.

Rate 30°	Rate 60°	λ_m
98.2	97.3	2.71
98.1	97.4	1.46
98.2	**97.9**	0.57
97.7	96.8	0.19
91.7	88.5	0.07

u_1	u_2	u_3	u_4	u_5
0.37	0.54	0.55	-0.44	0.28
0.49	0.43	-0.06	0.51	-0.55
0.53	-0.01	-0.59	0.07	0.60
0.47	-0.46	-0.11	-0.58	-0.47
0.33	-0.56	0.57	0.44	0.21
$\approx \int_t$	$\approx \partial_t$	$\approx \partial_{tt}$	$\approx \partial_{t^3}$	$\approx \partial_{t^4}$

Table 5: The principal components u_m of the 5-D grey-level descriptor with $\nabla^2 L$ at five scales $\sigma = 1, 2, 4, 8, 16$ for the COIL-100 data set together with the principal values λ_m and the recognition scores obtained with PCA-reduced histograms using $M = 5...1$ principal components with viewsphere samplings of 30° and 60°.

of the different channels over scale, the principal component u_2 corresponds to a first-order difference over scales and the principal component u_3 can be interpreted as a second-order difference, etc.

For this 5-D grey-level descriptor with Laplacians $\nabla^2 L$ at five scales, the best performance is obtained using $M = 3$ principal components. In terms of corresponding principal components, an interpretation of this is that scale differences up to order two are useful for improving the performance of object recognition using this 5-D grey-level descriptor.

From the fact that the scale-space derivatives satisfy the diffusion equation

$$\partial_t L_{x^\alpha y^\beta} = \frac{1}{2} \nabla^2 L_{x^\alpha y^\beta} \tag{17}$$

it follows that a first-order derivative with respect to scale t can be transformed into second-order derivative with respect to space x and y using the Laplacian operator. The summation of filter responses over scales can, however, be interpreted as an integration over scales, and will therefore reduce the effective order of differentiation.

5.2. Composed histograms of pure grey-level cues

In our investigation of composed complex-cue histograms of scale-space features up to order two, let us first restrict ourselves to image descriptors formed from grey-level cues only. Of order zero, we have the image intensity L; of order one the image gradient (L_x, L_y) and the gradient magnitude $|\nabla L|$; and of order two the components of the Hessian matrix (L_{xx}, L_{xy}, L_{yy}) as well as the Laplacian $\nabla^2 L$ and the determinant of the Hessian $\det \mathcal{H} L$.

By computing these descriptors at scale levels $\sigma \in \{1, 2, 4\}$, motivated by the observation in section 4.1 that this combination of scale levels gives results that are among the best for primitive histograms formed from scale-space features defined from single grey-level cues, we obtain a set of primitive image descriptors as well as a set of combined descriptors as listed in table 6. These composed image descriptors have been constructed by combining primitive scale-space features that are either rotationally invariant or rotationally variant for different orders of differentiation. This table also shows performance values for object recognition and object classification on our six benchmark problems, including the best number of principal components in PCA-reduced histograms, the number of histogram bins r and the number of non-zero cells n. We will analyze these results in section 5.5.

5.3. Composed histograms from chromatic channels

Given the above mentioned definitions of grey-level descriptors, we can also define corresponding descriptors from the colour-opponent channels $(C^{(1)}, C^{(2)})$. Motivated by the observation in section 4.1 that the single scale $\sigma = 2$ gave results that are either among or near the best for the chromatic image descriptors, we will evaluate composed chromatic image descriptors at this scale only, to reduce the degrees of freedom in the experiments.

In [39] we present a detailed evaluation of such purely chromatic histogram descriptors, where a 6-D histogram of (C, C_x, C_y) at $\sigma = 2$ and 2-D histogram of the chromatic channels C at $\sigma = 2$ have the best performance for object instance recognition and a 10-D histogram of $(C_x, C_y, C_{xx}, C_{xy}, C_{yy})$ has the best performance for category classification. Some chromatic descriptors will later be included in summaries of the best performing descriptors.

5.4. Composed histograms of combined grey-level and spatio-chromatic cues

By the combination of grey-level features with spatio-chromatic colour-opponent cues, we can also define a set of composed descriptors that capture the joint distribution of grey-level cues and chromatic cues as shown in table 7. This set of composed image features basically follows the organization of the above mentioned descriptors for pure grey-level cues or chromatic cues, with the minor

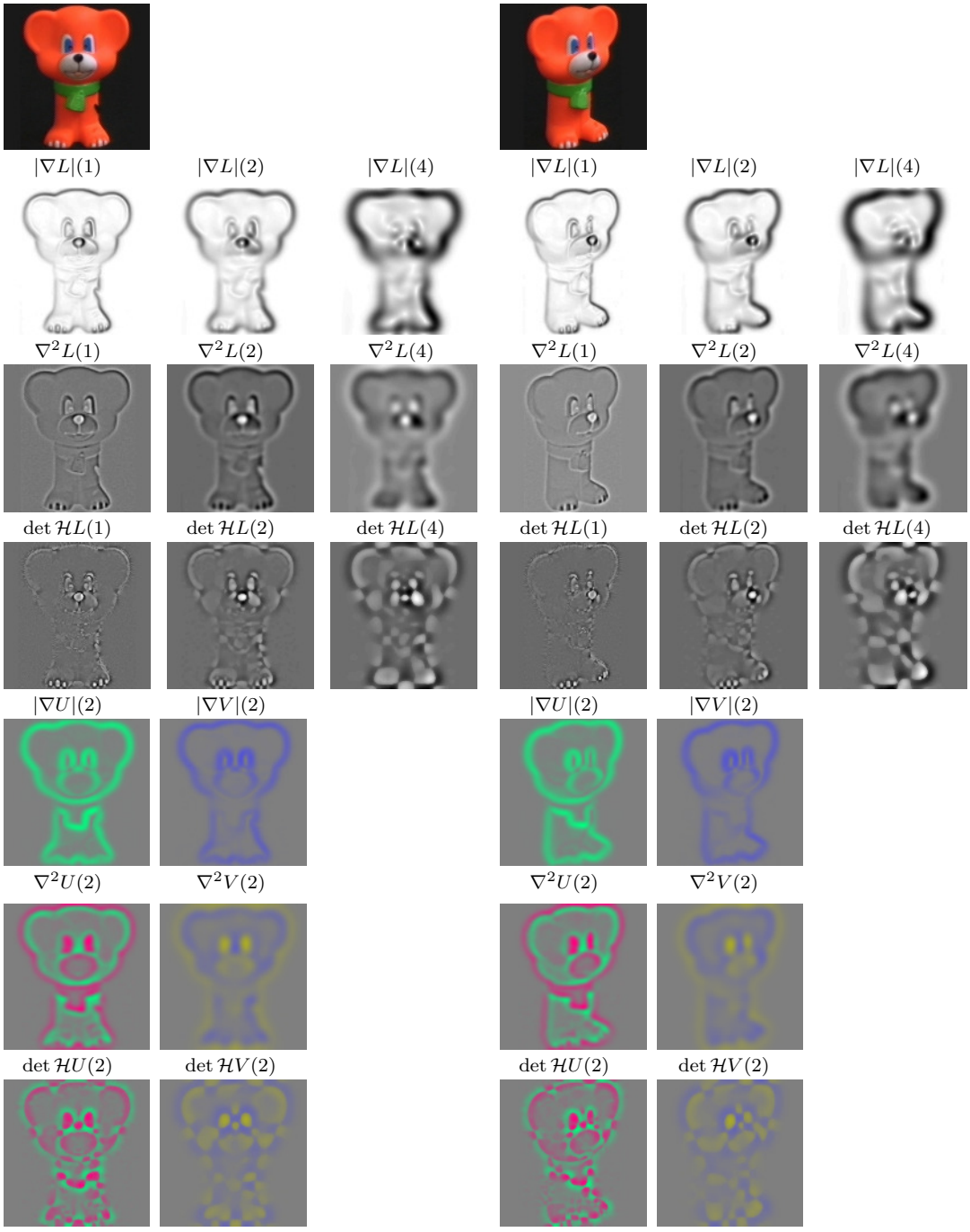

Figure 7: Illustration of some the underlying image features in the view variation experiments, where composed multi-cue histograms of local scale-space features are computed for a subset of views on the view-sphere and then matched to previously unseen views. This figure shows the rotationally invariant differential invariants for grey-levels ($|\nabla L|, \nabla^2 L, \det \mathcal{H}L$) at scale levels $\sigma = 1, 2, 4$ and the corresponding rotationally invariant spatio-chromatic differential invariants for the colour-opponent channels ($|\nabla C|, \nabla^2 C, \det \mathcal{H}C$) at $\sigma = 2$ computed for two different views of a a three-dimensional object rotated by 30° in the COIL-100 data set.

Image features	30°	COIL-100			45°	ETH-80			ETH	Class.		
	60°	r	PC	n	90°	r	PC	n	CalTech	r	PC	n
3-D $(L; \sigma = 1, 2, 4)$	*99.4*	18	3	1118	*99.0*	11	3	857	91.9	25	3	4235
	93.1	14	3	681	94.1	19	3	2667	98.0	13	3	1497
6-D $(L_x, L_y; \sigma = 1, 2, 4)$	99.1	45	$\tilde{6}$	1755	**99.3**	5	6	4399	94.8	6	6	6003
	93.8	7	5	1517	93.1	5	5	2478	**99.6**	10	6	49605
3-D $(\lvert\nabla L\rvert; \sigma = 1, 2, 4)$	*99.4*	10	3	403	98.0	12	3	1329	92.2	195	$\tilde{3}$	3121
	94.3	100	$\tilde{3}$	435	91.6	155	$\tilde{3}$	2123	98.4	195	$\tilde{3}$	5681
9-D $(L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2, 4)$	97.9	100	$\tilde{3}$	1288	*99.0*	4	4	593	96.1	14	–	4516
	91.8	95	$\tilde{3}$	1171	92.5	45	$\tilde{3}$	1199	**99.6**	10	$\tilde{9}$	7908
3-D $(\nabla^2 L; \sigma = 1, 2, 4)$	98.9	40	–	909	97.4	13	3	2903	92.7	75	$\tilde{3}$	1601
	93.7	95	$\tilde{3}$	751	90.9	13	3	2924	98.9	30	–	3245
3-D $(\det \mathcal{H} L; \sigma = 1, 2, 4)$	99.0	7	3	445	95.2	20	$\tilde{3}$	1379	93.0	10	$\tilde{3}$	386
	91.6	9	3	733	85.8	19	$\tilde{3}$	1274	99.2	50	–	2393
6-D $(\nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4)$	*99.5*	5	5	1014	*99.0*	6	5	4384	95.7	6	–	320
	95.8	7	4	933	92.5	17	–	2440	*99.4*	25	$\tilde{6}$	6459
9-D $(L, L_x, L_y; \sigma = 1, 2, 4)$	99.2	3	9	1098	**99.4**	8	4	1787	94.7	10	$\tilde{9}$	1340
	94.2	7	7	2953	94.1	5	6	3223	*99.6*	4	8	14328
6-D $(L, \lvert\nabla L\rvert; \sigma = 1, 2, 4)$	**99.7**	7	5	920	*99.1*	12	–	1431	93.3	7	6	4541
	95.4	11	5	2266	*95.1*	11	4	2496	98.8	155	$\tilde{6}$	13185
15-D $(L_x, L_y, L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2, 4)$	99.0	45	$\tilde{5}$	1064	**99.4**	5	6	4504	**96.6**	8	$\tilde{10}$	3200
	94.6	55	$\tilde{6}$	2429	93.8	45	$\tilde{5}$	4703	**99.7**	6	$\tilde{12}$	10377
6-D $(\lvert\nabla L\rvert, \nabla^2 L; \sigma = 1, 2, 4)$	*99.7*	8	4	826	98.7	6	6	4896	94.2	4	5	1003
	95.7	75	$\tilde{5}$	1808	94.4	4	6	2076	*99.2*	24	–	22358
6-D $(\lvert\nabla L\rvert, \det \mathcal{H} L; \sigma = 1, 2, 4)$	*99.7*	6	5	1427	98.3	5	6	4802	94.0	5	5	2395
	96.0	5	6	1647	93.6	4	6	2609	99.1	8	5	14164
9-D $(\lvert\nabla L\rvert, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4)$	*99.7*	4	5	457	98.6	4	8	6017	**96.6**	14	–	4664
	96.2	5	5	761	94.6	4	7	4195	*99.5*	12	–	12307
9-D $(L, \lvert\nabla L\rvert, \nabla^2 L; \sigma = 1, 2, 4)$	**99.7**	5	6	921	**99.3**	4	6	1502	93.9	60	$\tilde{7}$	5565
	95.9	5	6	907	**95.4**	8	4	1324	*99.2*	95	$\tilde{7}$	32582
12-D $(L, \lvert\nabla L\rvert, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4)$	**99.8**	5	7	1905	*99.1*	3	10	5215	94.8	95	$\tilde{5}$	3996
	96.9	3	8	862	*95.2*	4	8	5324	*99.3*	55	$\tilde{8}$	34419

Table 6: The performance of *PCA-reduced composed complex-cue histogram descriptors* formed from *pure grey-level structure* for the 6 different benchmark problems described in section 3.5 and using different combinations of scale levels for computing the scale-space features. For each experiment, the number of principal components PC and the number of bins r that gave the best performance are shown, as is the average number n of non-zero bins for the corresponding histograms. A tilde over the number of principal components PC means that the best result was obtained without variance normalization prior to PCA. A dash "–" in the PC column denotes that the best result was obtained without PCA or variance normalization. The first two columns show results for object instance recognition, whereas the third column shows results for object category classification.

difference that we have here also included few additional descriptors that correspond to the suppression of certain cues. In next section, we will analyse the results of these composed histogram descriptors and compare them to the earlier defined histogram descriptors in terms of either grey-level cues or chromatic cues.

Figure 7 shows an illustration of computing the underlying grey-level and colour features from two different views of an object in the COIL-100 dataset concerning the rotationally invariant scale-space features ∇L, $\nabla^2 L$, $\det \mathcal{H} L$, ∇C, $\nabla^2 C$ and $\det \mathcal{H} C$.

5.5. Comparisons between different descriptors

The above mentioned experimental setup provides an exhaustive catalogue of the different types of image descriptors that are natural to construct from the differential image structure up to order two. Taken together, we have defined 6+6 = 12 types of primitive histogram descriptors and 15 +15 +19 = 49 composed histogram descriptors; see also [39] for further experimental results and a more detailed analysis. Table 8 and table 9 show summarizing rankings of the 5 or 10 best image descriptors for each one of the six benchmark task.

As can be seen from table 8, for the COIL-100 data set with 30° viewsphere sampling, as many as 12 histogram descriptors reached a performance of 100.0 % and 20 histogram descriptors (not all shown) were above 99.9 %. Notably, a clear majority of these image descriptors are based on different combinations of image intensities L, colour-opponent channels C, grey-level differential invariants $|\nabla L|$, $\nabla^2 L$, $\det \mathcal{H} L$ and the

Image features		30°	COIL-100			45°	ETH-80			ETH	Class.		
		60°	r	PC	n	90°	r	PC	n	CalTech	r	PC	n
5-D	$(L; \sigma=1,2,4), (C; \sigma=2)$	**99.9**	13	4	297	*99.8*	6	–	188	87.3	11	–	513
		98.4	30	3	241	*99.1*	255	$\tilde{5}$	750	98.7	8	5	1964
10-D	$(L_x, L_y; \sigma=1,2,4)$,	*99.8*	3	10	900	99.4	4	8	4214	94.0	2	10	2188
	$(C_x, C_y; \sigma=2)$	97.6	21	$\tilde{8}$	732	97.1	9	6	6659	**99.9**	10	5	18107
5-D	$(\|\nabla L\|; \sigma=1,2,4)$,	**100.0**	6	5	490	**99.9**	15	4	4038	91.3	140	–	11915
	$(\|\nabla C\|; \sigma=2)$	*98.6*	5	5	312	98.3	9	5	3888	98.7	10	5	10644
15-D	$(L_{xx}, L_{xy}, L_{yy}; \sigma=1,2,4)$,	99.3	12	$\tilde{9}$	1926	*99.8*	3	12	6782	94.8	24	$\tilde{5}$	3270
	$(C_{xx}, C_{xy}, C_{yy}; \sigma=2)$	95.3	13	$\tilde{9}$	1104	96.1	4	8	5876	**99.9**	4	8	14749
5-D	$(\nabla^2 L; \sigma=1,2,4)$,	**100.0**	11	4	1399	*99.8*	13	$\tilde{5}$	988	91.2	95	$\tilde{4}$	8602
	$(\nabla^2 C; \sigma=2)$	97.9	29	$\tilde{5}$	501	98.5	7	4	1905	99.4	36	$\tilde{5}$	8146
5-D	$(\det \mathcal{H}L; \sigma=1,2,4)$,	*99.7*	7	5	2079	99.1	6	$\tilde{5}$	2711	92.7	8	$\tilde{3}$	255
	$(\det \mathcal{H}C; \sigma=2)$	97.2	9	$\tilde{5}$	577	95.0	5	5	3334	99.3	40	–	4629
10-D	$(\nabla^2 L, \det \mathcal{H}L; \sigma=1,2,4)$,	*99.9*	21	$\tilde{7}$	1271	*99.7*	4	8	6522	93.8	30	2	1954
	$(\nabla^2 C, \det \mathcal{H}C; \sigma=2)$	*98.6*	3	9	1157	97.3	3	9	4901	*99.6*	10	$\tilde{9}$	12909
8-D	$(\nabla^2 L, \det \mathcal{H}L; \sigma=1,2,4)$,	**100.0**	5	7	2139	*99.8*	4	6	2506	93.5	10	$\tilde{7}$	2335
	$(\nabla^2 C; \sigma=2)$	98.4	3	7	625	98.0	5	6	4453	*99.7*	5	8	26669
15-D	$(L, L_x, L_y; \sigma=1,2,4)$,	**99.9**	2	–	2390	*99.8*	2	11	1680	93.8	35	$\tilde{8}$	1230
	$(C, C_x, C_y; \sigma=2)$	98.4	6	–	3962	97.7	6	10	8034	*99.8*	4	10	17049
10-D	$(L, \|\nabla L\|; \sigma=1,2,4)$,	**100.0**	7	8	1510	**99.9**	2	$\tilde{10}$	518	87.8	3	10	1824
	$(C, \|\nabla C\|; \sigma=2)$	98.3	7	7	731	**99.4**	19	4	2193	99.1	6	$\tilde{10}$	1834
25-D	$(L_x, L_y, L_{xx}, L_{xy}, L_{yy}; \sigma=1,2,4)$,	99.4	9	–	1172	*99.5*	3	9	3127	**95.9**	14	$\tilde{9}$	4277
	$(C_x, C_y, C_{xx}, C_{xy}, C_{yy}; \sigma=2)$	96.0	9	–	1140	97.3	3	12	5905	**99.9**	4	9	20548
10-D	$(\|\nabla L\|, \nabla^2 L; \sigma=1,2,4)$,	**100.0**	2	10	1160	*99.8*	4	6	1388	92.3	95	$\tilde{3}$	1144
	$(\|\nabla C\|, \nabla^2 C; \sigma=2)$	*98.8*	5	8	1476	98.8	4	8	3560	99.3	6	$\tilde{10}$	4292
8-D	$(\|\nabla L\|, \nabla^2 L; \sigma=1,2,4)$,	**100.0**	4	8	1606	**100.0**	4	5	822	94.2	60	$\tilde{5}$	3880
	$(\nabla^2 C; \sigma=2)$	98.2	35	$\tilde{6}$	779	98.3	4	6	1662	*99.4*	24	–	37221
8-D	$(\|\nabla L\|, \nabla^2 L; \sigma=1,2,4)$,	**100.0**	8	5	1082	**99.9**	4	7	2574	94.1	8	8	10290
	$(\|\nabla C\|; \sigma=2)$	**98.9**	5	6	674	98.3	8	5	3435	99.4	4	7	5288
10-D	$(\|\nabla L\|, \det \mathcal{H}L; \sigma=1,2,4)$,	**100.0**	20	$\tilde{7}$	1631	*99.6*	2	$\tilde{10}$	2113	93.1	24	$\tilde{7}$	7348
	$(\|\nabla C\|, \det \mathcal{H}C; \sigma=2)$	*98.8*	3	9	1064	97.0	6	$\tilde{9}$	2953	99.3	11	$\tilde{10}$	16783
15-D	$(\|\nabla L\|, \nabla^2 L, \det \mathcal{H}L; \sigma=1,2,4)$,	**100.0**	2	$\tilde{11}$	1552	*99.8*	4	7	3656	94.8	75	$\tilde{4}$	2806
	$(\|\nabla C\|, \nabla^2 C, \det \mathcal{H}C; \sigma=2)$	*98.8*	3	10	1257	97.6	2	$\tilde{15}$	11170	*99.6*	5	9	30440
13-D	$(\|\nabla L\|, \nabla^2 L, \det \mathcal{H}L; \sigma=1,2,4)$,	**100.0**	2	$\tilde{12}$	2380	*99.8*	2	$\tilde{12}$	3178	94.9	75	$\tilde{4}$	2793
	$(\|\nabla C\|, \nabla^2 C; \sigma=2)$	**99.0**	3	11	1520	98.3	4	9	6327	*99.7*	2	$\tilde{13}$	12982
15-D	$(L, \|\nabla L\|, \nabla^2 L; \sigma=1,2,4)$,	**100.0**	2	$\tilde{15}$	2371	**99.9**	2	15	5305	92.7	100	$\tilde{5}$	1634
	$(C, \|\nabla C\|, \nabla^2 C; \sigma=2)$	*98.6*	29	$\tilde{11}$	1023	98.7	8	7	6166	*99.7*	2	$\tilde{12}$	5284
20-D	$(L, \|\nabla L\|, \nabla^2 L, \det \mathcal{H}L; \sigma=1,2,4)$,	**100.0**	2	$\tilde{10}$	561	*99.8*	2	15	6998	93.2	35	$\tilde{8}$	2279
	$(C, \|\nabla C\|, \nabla^2 C, \det \mathcal{H}C; \sigma=2)$	*98.9*	3	10	1014	98.0	20	$\tilde{11}$	5837	*99.6*	2	$\tilde{15}$	26029

Table 7: The performance of *PCA-reduced composed complex-cue histogram descriptors* formed from *combinations of grey-level and chromatic structures* for the 6 different benchmark problems described in section 3.5 and using different combinations of scale levels for computing the scale-space features. (See caption of figure 6 for details regarding the notation.)

spatio-chromatic differential invariants $|\nabla C|$, $\nabla^2 C$ and $\det \mathcal{H}C$. With 60° viewsphere sampling, the best results were obtained for histogram descriptors involving different combinations of the grey-level differential invariants $|\nabla L|$, $\nabla^2 L$, $\det \mathcal{H}L$ and the spatio-chromatic differential invariants $|\nabla C|$ and $\nabla^2 C$, for one descriptor also including raw image intensities L and colour-opponent channels C.

For object recognition on the ETH-80 data set with 45° viewsphere sampling, we get 100.0 % performance for two histogram descriptors and 8 histogram descriptors reach a performance above 99.9 %. Many of the best image descriptors are based on image intensities L, colour-opponent channels C, grey-level differential invariants $|\nabla L|$ and $\nabla^2 L$ and the spatio-chromatic differential invariants $|\nabla C|$ and $\nabla^2 C$. A difference compared to the results for the COIL-100 data set, however, is that the determinants of the Hessians $\det \mathcal{H}L$ and $\det \mathcal{H}C$ are no longer included among the best image descriptors. Moreover, we also get very good performance with a PCA-reduced 22-dimensional histograms of the partial derivatives L_x, L_y, C_x and C_y reduced to 8 dimensions. With 90° viewsphere sampling on the ETH-80 data set, the best performance is obtained with a 10-D histogram of image intensities L, chromatic channels C and image gradients $|\nabla L|$ and $|\nabla C|$ reduced to 4 dimensions.

Object instance recognition COIL-100 30°

Image features		PC	r	Raw	Norm.	Best
5-D	$(\|\nabla L\|; \sigma = 1, 2, 4), (\|\nabla C\|; \sigma = 2)$	5	6	99.9	100.0	100.0
10-D	$(L, \|\nabla L\|; \sigma = 1, 2, 4), (C, \|\nabla C\|; \sigma = 2)$	8	7	99.8	100.0	100.0
10-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	10	2	99.7	100.0	100.0
15-D	$(L, \|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C; \sigma = 2)$	$\tilde{15}$	2	99.8	100.0	100.0
20-D	$(L, \|\nabla L\|, \nabla^2 L, \det \mathcal{H}L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C, \det \mathcal{H}C; \sigma = 2)$	$\tilde{10}$	2	99.8	99.7	100.0
10-D	$(\|\nabla L\|, \det \mathcal{H}L; \sigma = 1, 2, 4), (\|\nabla C\|, \det \mathcal{H}C; \sigma = 2)$	$\tilde{7}$	20	99.7	99.9	100.0
15-D	$(\|\nabla L\|, \nabla^2 L, \det \mathcal{H}L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C, \det \mathcal{H}C; \sigma = 2)$	$\tilde{11}$	2	99.8	99.9	100.0
13-D	$(\|\nabla L\|, \nabla^2 L, \det \mathcal{H}L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	$\tilde{13}$	2	99.8	100.0	100.0
5-D	$(\nabla^2 L; \sigma = 1, 2, 4), (\nabla^2 C; \sigma = 2)$	4	11	99.8	99.9	100.0
8-D	$(\nabla^2 L, \det \mathcal{H}L; \sigma = 1, 2, 4), (\nabla^2 C; \sigma = 2)$	7	5	99.6	99.9	100.0
8-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\nabla^2 C; \sigma = 2)$	8	4	99.7	100.0	100.0
8-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\|\nabla C\|; \sigma = 2)$	5	8	99.8	99.9	100.0

Object instance recognition COIL-100 60°

Image features		PC	r	Raw	Norm.	Best
13-D	$(\|\nabla L\|, \nabla^2 L, \det \mathcal{H}L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	11	3	97.6	98.5	99.0
8-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\|\nabla C\|; \sigma = 2)$	6	5	97.0	98.6	98.9
20-D	$(L, \|\nabla L\|, \nabla^2 L, \det \mathcal{H}L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C, \det \mathcal{H}C; \sigma = 2)$	10	3	97.3	97.1	98.9
10-D	$(\|\nabla L\|, \det \mathcal{H}L; \sigma = 1, 2, 4), (\|\nabla C\|, \det \mathcal{H}C; \sigma = 2)$	9	3	97.6	98.7	98.8
15-D	$(\|\nabla L\|, \nabla^2 L, \det \mathcal{H}L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C, \det \mathcal{H}C; \sigma = 2)$	10	3	96.7	98.2	98.8
10-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	8	5	98.1	98.1	98.8

Object instance recognition ETH-80 45°

Image features		PC	r	Raw	Norm.	Best
8-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\nabla^2 C; \sigma = 2)$	5	4	99.3	99.8	100.0
5-D	$(\|\nabla L\|; \sigma = 1, 2, 4), (\|\nabla C\|; \sigma = 2)$	4	15	99.4	99.8	99.9
10-D	$(L, \|\nabla L\|; \sigma = 1, 2, 4), (C, \|\nabla C\|; \sigma = 2)$	9	4	99.8	99.9	99.9
8-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\|\nabla C\|; \sigma = 2)$	7	4	98.6	99.6	99.9
15-D	$(L, \|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C; \sigma = 2)$	15	2	99.8	99.9	99.9
14-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4, 8, 16), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	6	5	98.8	99.9	99.9
22-D	$(L_x, L_y; \sigma = 1, 2, 4, 8, 16), (C_x, C_y; \sigma = 1, 2, 4)$	8	5	98.8	99.0	99.9
4-D	$(C; \sigma = 1, 2)$	$\tilde{4}$	255	99.7	99.9	99.9

Object instance recognition ETH-80 90°

Image features		PC	r	Raw	Norm.	Best
10-D	$(L, \|\nabla L\|; \sigma = 1, 2, 4), (C, \|\nabla C\|; \sigma = 2)$	4	19	98.2	99.1	99.4
5-D	$(L; \sigma = 1, 2, 4), (C; \sigma = 2)$	$\tilde{5}$	255	97.4	99.1	99.1
10-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	8	4	97.2	98.3	98.8
15-D	$(L, \|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C; \sigma = 2)$	7	8	98.0	97.8	98.7
4-D	$(C; \sigma = 1, 2)$	–	75	98.7	98.1	98.7

Table 8: The 5 to 10 best PCA-reduced histogram descriptors when performing *object instance recognition* on the COIL-100 and ETH-80 data sets. The column labelled "Best" shows the result for the best number of principal components shown in the column labelled "PC". A tilde ˜ over the number of principal components PC indicates that the result without variance normalization was better than the result obtained with variance normalization. A dash "–" in the PC column indicates that PCA did not lead to any improvement in performance.

Classification of object categories. Table 9 shows a corresponding ranking of the 5 best histogram descriptors for the category classification tasks. For object classification on the ETH-80 data set, we can see that several of the image descriptors are improved by variance normalization and that some of the histogram descriptors are improved by PCA, but not all. The best results are obtained from partial derivatives of grey-levels (L_{xx}, L_{xy}, L_{yy}) or the second-order differential invariants of grey-levels $(\nabla^2 L, \det \mathcal{H}L)$ complemented by the colour Laplacian $\nabla^2 C$. Notably, the best descriptor reaches a performance of 97.7 % and the four following descriptors a performance of 96.6–96.9 %. Hence, one may ask if there are inherent difficulties in the data.

Table 10 show the confusion matrix for one of the second best image descriptors on this benchmark problem, *i.e.*, the 12-D histogram with

Object category classification ETH-80

Image features		PC	r	Raw	Norm.	Best
25-D	$(L_x, L_y, L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2, 4, 8, 16)$	$\tilde{20}$	2	95.3	97.1	97.7
15-D	$(L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2, 4, 8, 16)$	–	2	96.9	96.7	96.9
12-D	$(\nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4, 8, 16), (\nabla^2 C; \sigma = 2)$	12	2	94.8	96.9	96.9
10-D	$(L_x, L_y; \sigma = 1, 2, 4, 8, 16)$	10	2	93.9	96.6	96.6
9-D	$(\lvert\nabla L\rvert, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4)$	–	14	96.6	94.7	96.6

Object category classification CalTech-4

Image features		PC	r	Raw	Norm.	Best
10-D	$(L_x, L_y; \sigma = 1, 2, 4), (C_x, C_y; \sigma = 2)$	5	10	99.6	99.6	99.9
15-D	$(L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2, 4), (C_{xx}, C_{xy}, C_{yy}; \sigma = 2)$	8	4	99.6	99.8	99.9
25-D	$(L_x, L_y, L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2, 4), (C_x, C_y, C_{xx}, C_{xy}, C_{yy}; \sigma = 2)$	9	4	99.5	99.3	99.9
15-D	$(L, L_x, L_y; \sigma = 1, 2, 4), (C, C_x, C_y; \sigma = 2)$	10	4	98.6	99.6	99.8
6-D	$(L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2)$	4	17	99.8	99.6	99.8

Table 9: The 5 best PCA-reduced histogram descriptors when performing *object category classification* on the ETH-80 and CalTech-4 data sets. The column labelled "Best" shows the result for the best number of principal components shown in the column labelled "PC". A tilde ˜ over the number of principal components PC indicates that the result without variance normalization was better than the result obtained with variance normalization. A dash "–" in the PC column indicates that PCA did not lead to any improvement in performance.

	apple	pear	tomato	car	cup	cow	horse	dog
apple	160	0	0	0	0	0	0	0
pear	0	160	0	0	0	0	0	0
tomato	0	0	160	0	0	0	0	0
car	0	0	0	160	0	0	0	0
cup	0	0	0	0	160	0	0	0
cow	0	0	0	0	0	155	4	1
horse	0	0	0	0	0	11	139	10
dog	0	0	0	0	0	9	5	146

Table 10: Confusion matrix for the ETH-80 dataset using the 12-D descriptor with $(\nabla^2 L, \det \mathcal{H} L)$ at $\sigma = 1, 2, 4, 8, 16)$ and $\nabla^2 C$ at $\sigma = 2$ transformed using PCA with 12 principal components. The overall performance is 96.9 %. With the cows, horses and dogs grouped into the super-category "plastic toy animals", the performance on the simplified category classification task is 100.0 %.

	airplane	car	face	motorbike
airplane	398	0	0	2
car	0	400	0	0
face	0	0	217	0
motorbike	0	0	0	400

Table 11: Confusion matrix for the CalTech-4 dataset using the 15-D descriptor with (L_{xx}, L_{xy}, L_{yy}) at $\sigma = 1, 2, 4$ and (C_{xx}, C_{xy}, C_{yy}) at $\sigma = 2$ transformed using PCA with 8 principal components. The overall performance is 99.9 %.

$(\nabla^2 L, \det \mathcal{H} L)$ at five scales and $\nabla^2 C$ at one scale. From an inspection of the errors, it can be seen that the majority of the mistakes are due to mixing up the different plastic toy animals "cows", "horses" and "dogs". From inspection of the corresponding images (see the last three rows in figure 4), we can easily understand this by noting that these toy animals have been manufactured in a similar way and been painted with similar patterns. If we group the cows, horses and dogs into a joint category "plastic toy animals", then the performance on the simpler category classification problem reaches 100.0 % for this 12-D histogram with the second-order grey-level differential invariants $(\nabla^2 L, \det \mathcal{H} L)$ at five scales, and also for the 25-D grey-level histogram with the first- and second-order partial derivatives $(L_x, L_y, L_{xx}, L_{xy}, L_{yy})$ at five scales.

For object categorization on the CalTech-4 dataset, we get significant increase in performance by dimensionality reduction by PCA. The (i) 10-D histogram of (L_x, L_y) at three scales and (C_x, C_y) at one scale, the (ii) 15-D histogram of (L_{xx}, L_{xy}, L_{yy}) at three scales and (C_{xx}, C_{xy}, C_{yy}) at one scale as well as the (iii) 25-D histogram of the spatial 2-jet $(L_x, L_y, L_{xx}.L_{xy}, L_{yy})$ and the 2-jet of the colour-opponent channels $(C_x, C_y, C_{xx}.C_{xy}, C_{yy})$ at one scale do all reach 99.9 % when reduced to 5, 8 or 9 dimensions respectively. Table 11 shows the confusion matrix for one of the best descriptors. As can be seen, only 2 out of 1417 images are misclassified and the mistakes correspond to classifying two airplane images as motorbike images. Very good results are also obtained for a 6-D histogram of (L_{xx}, L_{xy}, L_{yy}) at three scales reduced to four dimensions and for a 15-D histogram of the zero- and first-order cues (L, L_x, L_y) at three scales and (C, C_x, C_y) at one scale, when reduced to 10 dimensions. Compared to the ETH-80 data set most of the best categorization results are obtained with pure grey-level cues, most of the best category clas-

sification results for the CalTech-4 data set are obtained with combinations of grey-level and colour.

For both the ETH-80 and the CalTech-4 data sets, we can conclude that also histogram-based methods for visual object categorization can be improved significantly by variance normalization and/or dimensionality reduction by PCA.

Concerning the number of bins, we can also note that many of the best image descriptors for object categorization have been computed using as few as 2 or 4 bins per dimension. For object instance recognition with a rather dense sampling of the view-sphere as in the COIL-100 30° and ETH-80 45° benchmarks, many of the best image descriptors were also obtained with as few as 2 or 4 bins. With a less dense sampling of the view-sphere as in the COIL-100 60° benchmark, the best results were obtained with 3 or 5 bins. From these observations, we can conclude that a rather low number of bins r is often sufficient to obtain good recognition or classification performance. By using such low numbers of bins, the histograms can therefore in combination with automatic exposure control be expected to possess certain robustness properties to illumination variations.

Combined ranking of the performance on the object recognition tasks. For the four different types of object recognition problems we have considered here, it is clear that the results depend on the types of objects in the dataset as well as the density of the viewsphere sampling and thus the amount of image deformations between the images in the training and the test sets. From this observation we can conclude that which descriptor performs best on a given task may in general depend on how well the variability in the training data spans the variability in the test data. To allow for a joint interpretation of these results, table 12 shows the result of a summarizing ranking the data on a single (logarithmic) performance measure of the form

$$W(D) = -\frac{1}{K} \sum_{k=1}^{K} \log_2(1 - p_k(D) + \varepsilon) \qquad (18)$$

for each image descriptor D, where k is an index for each one of the $K = 4$ object recognition tasks, $p_k(D)$ denotes the performance of the image descriptor D in that experiment and $\varepsilon \approx 0.25$ % is a constant to avoid computing logarithms of values close or equal to zero when the performance approaches 100.0 %.

As can be seen from the table, all the best image descriptors on this ranking contain explicit spatio-chromatic colour-opponent cues, where the best result is obtained from a 10-D histogram that combines grey-level values L, gradient magnitudes $|\nabla L|$, colour-opponent channels C and colour-opponent gradient magnitudes $|\nabla C|$. Somewhat more surprisingly, however, zero-order cues in terms of image intensities L and/or colour-opponent channels C are included in the two best histogram descriptors. Even a 5-D histogram of grey-level gradient magnitudes $|\nabla L|$ and colour-opponent gradient magnitudes $|\nabla C|$ leads to very good performance.

Table 13 shows a corresponding ranking where all image descriptors involving explicit zero-order cues have been suppressed. Such an exclusion can, for example, be motivated by a requirement of having the image descriptors being invariant to additive transformations of image intensities and chromatic channels. As can be seen from the results, all of the best solely derivative based image descriptors are based on combinations or subsets of the differential invariants $(|\nabla L|, \nabla^2 L)$ and $(|\nabla C|, \nabla^2 C)$.

From both these summaries it is apparent that with regard to the problem of recognizing previously seen objects from novel views, much better results are obtained from rotationally invariant image descriptors defined from differential invariants compared to rotationally variant descriptors defined from partial derivatives.

Combined ranking of the performance on the object classification tasks. Table 14 shows the result of a corresponding joint ranking of the performance on the object classification tasks. Here, second-order grey-level cues in terms of the partial derivatives (L_{xx}, L_{xy}, L_{yy}) are included in the three best image descriptors. Specifically, the best result is obtained for a 25-D rotationally variant histogram based on the 2-jet of image intensities $(L_x, L_y, L_{xx}, L_{xy}, L_{yy})$ at three scales and the spatio-chromatic colour-opponent 2- jet $(C_x, C_y, C_{xx}, C_{xy}, C_{yy})$ at one scale. The best rotationally invariant descriptor turns out to be a 9-D histogram of the independent rotationally invariant differential invariants $(|\nabla L|, \nabla^2 L, \det \mathcal{H} L)$ computed from intensity information only, with its combined performance value $W = 4.1$ just below the five best rotationally variant image descriptors.

Combined score for the object recognition tasks

Image features		D	r	W
10-D	$(L, \|\nabla L\|; \sigma = 1, 2, 4), (C, \|\nabla C\|; \sigma = 2)$	4–9	4–19	5.1
15-D	$(L, \|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C; \sigma = 2)$	7–15	2–29	5.0
10-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	6–10	2–5	5.0
8-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\|\nabla C\|; \sigma = 2)$	5–7	4–8	5.0
5-D	$(L; \sigma = 1, 2, 4), (C; \sigma = 2)$	3–5	6–255	4.9
5-D	$(\|\nabla L\|; \sigma = 1, 2, 4), (\|\nabla C\|; \sigma = 2)$	4,5	5–15	4.9
8-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\nabla^2 C; \sigma = 2)$	5–8	4,35	4.9
13-D	$(\|\nabla L\|, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	9–13	2–4	4.9
20-D	$(L, \|\nabla L\|, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C, \det \mathcal{H} C; \sigma = 2)$	10–15	2–20	4.9
15-D	$(\|\nabla L\|, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C, \det \mathcal{H} C; \sigma = 2)$	7–15	2–4	4.8

Table 12: The 10 best histogram descriptors according to the combined logarithmic performance measure W in equation (18) for the 4 *object instance recognition tasks* for the COIL-100 and ETH-80 datasets.

Combined score for the object recognition tasks excluding 0-order cues

Image features		D	r	W
10-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	6–10	2–5	5.0
8-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\|\nabla C\|; \sigma = 2)$	5–7	4–8	5.0
5-D	$(\|\nabla L\|; \sigma = 1, 2, 4), (\|\nabla C\|; \sigma = 2)$	4,5	5–15	4.9
8-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\nabla^2 C; \sigma = 2)$	5–8	4,35	4.9
13-D	$(\|\nabla L\|, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	9–13	2–4	4.9

Table 13: The 5 best histogram descriptors according to the combined logarithmic performance measure W in (18) for the 4 *object instance recognition tasks* for the COIL-100 and ETH-80 datasets. This table shows similar results as table 12, with the difference that *image descriptors involving explicit zero-order cues have been excluded.*

Combined score for the category classification tasks

Image features		D	r	W
25-D	$(L_x, L_y, L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2, 4), (C_x, C_y, C_{xx}, C_{xy}, C_{yy}; \sigma = 2)$	9	4,14	4.3
15-D	$(L_x, L_y, L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2, 4)$	10,12	6,8	4.3
6-D	$(L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2)$	4,6	17,30	4.3
10-D	$(L_x, L_y; \sigma = 1, 2, 4), (C_x, C_y; \sigma = 2)$	5,10	2,10	4.3
21-D	$(L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2, 4, 8, 16), (C_{xx}, C_{xy}, C_{yy}; \sigma = 2)$	8,21	2,4	4.2
10-D	$(L_x, L_y; \sigma = 1, 2, 4, 8, 16)$	5,10	2,13	4.2

Table 14: The 6 best histogram descriptors according to the combined logarithmic performance measure W in equation (18) for the 2 *object category classification tasks* for the ETH-80 and CalTech-4 datasets.

6. Histograms of local binary information

An observation that can be made regarding higher-dimensional histogram descriptors is that in many cases a quite reasonable performance can be obtained by using as few as $r = 2$ bins per dimension. For image descriptors that reflect the *local polarity* of the signal, such as the partial derivatives $(L_x, L_y, L_{xx}, L_{xy}, L_{yy})$ and $(C_x, C_y, C_{xx}, C_{xy}, C_{yy})$ or the differential invariants $(\nabla^2 L, \det \mathcal{H} L)$ and $(\nabla^2 C, \det \mathcal{H} C)$, this means that object recognition or category classification is performed using only the *sign* of the differential expression. Therefore, the classification will automatically be *invariant to uniform rescalings of the intensity values and the chromaticity channels*, $f'(x, y) = a_f \, f(x, y)$ and $c'(x, y) = a_c \, c(x, y)$. Since differential scale-space features are also invariant under additive intensity and chromaticity transformations, we may therefore expect a particularly good robustness to illumination variations. In particular, the histogram descriptors will be *invariant under affine intensity and chromaticity transformations*.

Interpretation. Using $r = 2$ bins for image intensities L means that a distinction is only made between bright or dark image point, and for the colour-opponent channels that distinctions are only made between red or green and between yellow or blue. For the gradient magnitudes $|\nabla L|$ and $|\nabla C|$, a corresponding use of $r = 2$ bins means that classifications are only made between edge or non-edge image structures in grey-level information and colour-opponent information. Concerning the Laplacian

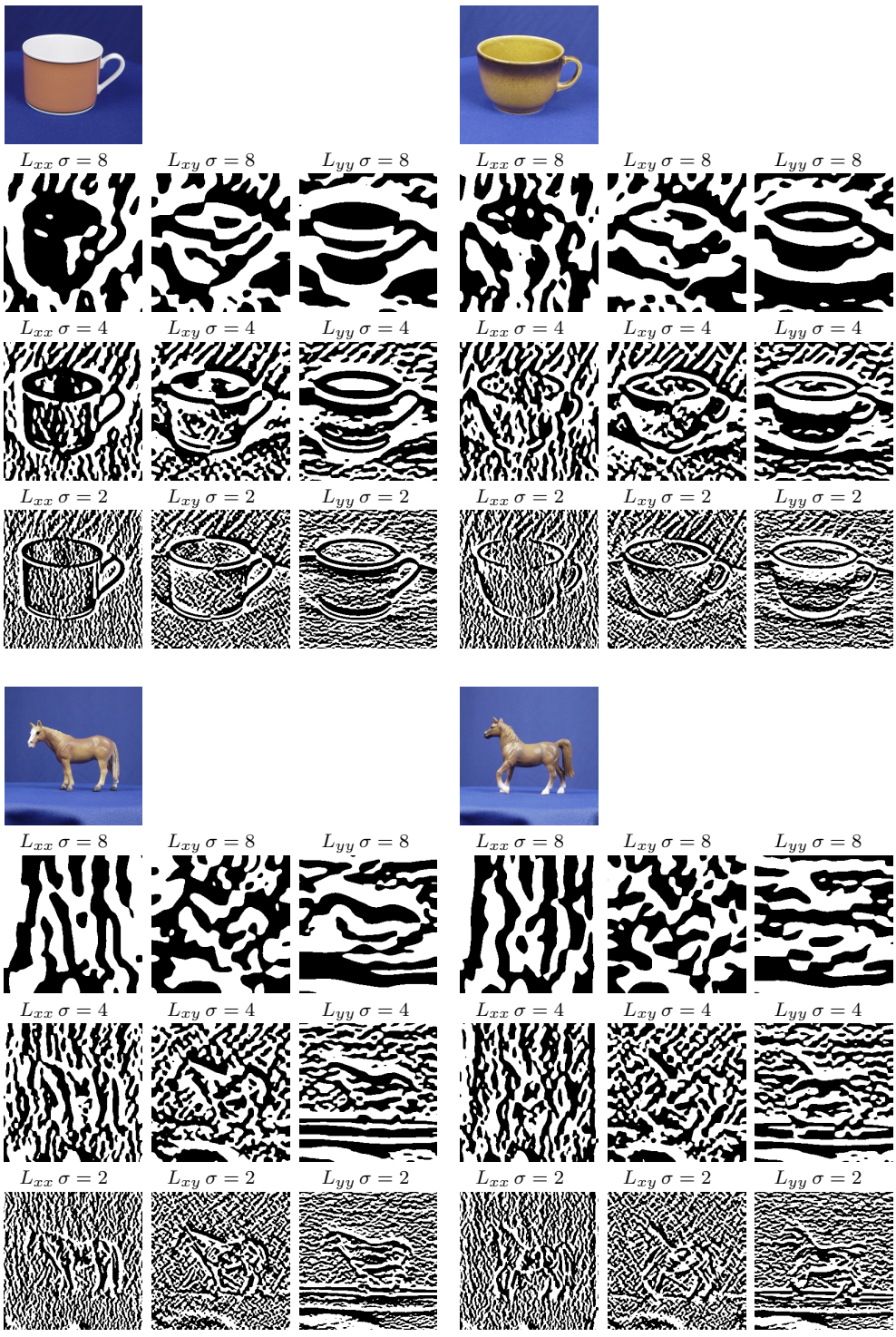

Figure 8: Illustration of the *quantized scale-space features* that underlie the computation of *binary histograms*. This figure shows the signs of the second-order partial derivatives (L_{xx}, L_{xy}, L_{yy}) computed at scale levels $\sigma = 2, 4, 8$ for two images of cups and horses, respectively, from the ETH-80 data set. These quantized differential entities occur in the 15-D binary histogram with (L_{xx}, L_{xy}, L_{yy}) at $\sigma = 1, 2, 4, 8, 16$, which leads to the best category classification results for the ETH-80 data set.

Object instance recognition COIL-100 30°

Image features		PC	r	Raw	Norm.	Best
10-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	10	2	94.2	100.0	100.0
15-D	$(L, \|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C; \sigma = 2)$	$\tilde{15}$	2	99.1	100.0	100.0
20-D	$(L, \|\nabla L\|, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C, \det \mathcal{H} C; \sigma = 2)$	$\tilde{10}$	2	99.2	99.6	100.0
14-D	$(\nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4, 8, 16), (\nabla^2 C, \det \mathcal{H} C; \sigma = 2)$	$\tilde{11}$	2	99.0	99.5	100.0
15-D	$(\|\nabla L\|, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C, \det \mathcal{H} C; \sigma = 2)$	$\tilde{11}$	2	99.0	99.8	100.0
13-D	$(\|\nabla L\|, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	$\tilde{13}$	2	98.7	100.0	100.0
8-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\nabla^2 C; \sigma = 2)$	8	2	94.2	100.0	100.0
10-D	$(L, \|\nabla L\|; \sigma = 1, 2, 4), (C, \|\nabla C\|; \sigma = 2)$	$\tilde{10}$	2	90.8	100.0	100.0
8-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\|\nabla C\|; \sigma = 2)$	$\tilde{8}$	2	55.6	99.9	99.9
15-D	$(L, L_x, L_y; \sigma = 1, 2, 4), (C, C_x, C_y; \sigma = 2)$	–	2	99.9	99.7	99.9

Object instance recognition COIL-100 60°

Image features		PC	r	Raw	Norm.	Best
15-D	$(L, \|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C; \sigma = 2)$	$\tilde{15}$	2	95.1	98.3	98.3
20-D	$(L, \|\nabla L\|, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C, \det \mathcal{H} C; \sigma = 2)$	$\tilde{15}$	2	96.5	96.8	98.1
10-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	$\tilde{10}$	2	85.5	98.1	98.1
13-D	$(\|\nabla L\|, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	$\tilde{11}$	2	93.2	97.5	97.9
15-D	$(\|\nabla L\|, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C, \det \mathcal{H} C; \sigma = 2)$	$\tilde{11}$	2	95.0	96.8	97.9

Object instance recognition ETH-80 45°

Image features		PC	r	Raw	Norm.	Best
10-D	$(L, \|\nabla L\|; \sigma = 1, 2, 4), (C, \|\nabla C\|; \sigma = 2)$	$\tilde{10}$	2	92.2	99.9	99.9
15-D	$(L, \|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C; \sigma = 2)$	15	2	98.7	99.9	99.9
14-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4, 8, 16), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	$\tilde{14}$	2	96.3	99.9	99.9
15-D	$(L, L_x, L_y; \sigma = 1, 2, 4), (C, C_x, C_y; \sigma = 2)$	11	2	99.7	99.8	99.8
20-D	$(L, \|\nabla L\|, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C, \det \mathcal{H} C; \sigma = 2)$	15	2	99.3	99.5	99.8

Object instance recognition ETH-80 90°

Image features		PC	r	Raw	Norm.	Best
15-D	$(L, \|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C; \sigma = 2)$	$\tilde{15}$	2	95.8	98.4	98.4
10-D	$(\|\nabla L\|, \nabla^2 L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	$\tilde{10}$	2	80.4	98.1	98.1
10-D	$(L, \|\nabla L\|; \sigma = 1, 2, 4), (C, \|\nabla C\|; \sigma = 2)$	$\tilde{10}$	2	82.1	98.0	98.0
13-D	$(\|\nabla L\|, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (\|\nabla C\|, \nabla^2 C; \sigma = 2)$	$\tilde{13}$	2	88.7	97.9	97.9
20-D	$(L, \|\nabla L\|, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (C, \|\nabla C\|, \nabla^2 C, \det \mathcal{H} C; \sigma = 2)$	15	2	96.2	96.9	97.8

Table 15: The 5 or 10 best PCA-reduced histogram descriptors when performing *object instance recognition* using *binary histograms* on the COIL-100 and ETH-80 data sets. The column labelled "Best" shows the result for the best number of principal components shown in the column labelled PC. A tilde ˜ over the number of principal components PC indicates that the result without variance normalization was better than the result with variance normalization. A dash "–" in the PC column indicates that PCA did not lead to any improvement in performance.

operators $\nabla^2 L$ and $\nabla^2 C$, the sign of the grey-level Laplacian reflects whether the local image structure should be regarded as a bright or dark blob, while the sign of the colour-opponent Laplacian determines whether the local image pattern can be regarded as a red blob on a green background, green blob on a red background, yellow blob on a blue background or a blue blob on a yellow background, respectively. For the determinant of the Hessian $\det \mathcal{H} L$ applied to grey-level information, the sign determines whether the local image pattern is elliptic (the Hessian matrix is positive or negative definite) or hyperbolic (the Hessian matrix is indefinite). The sign of the Laplacian together with the sign of the determinant of the Hessian do hence characterize the qualitative type of local second-order differential geometric structure. The sign of the determinant of the Hessian $\det \mathcal{H} C$ applied to chromatic information makes corresponding judgements for the red/green and yellow/blue colour-opponent channels. Hence, binary classification of such data, as implied of using $r = 2$ bins per dimension, means that *a detailed qualitative judgement is made of the local differential image structure at each image point*, optionally complemented by a characterization in terms of bright/dark, red/green, yel-

Object category classification ETH-80

Image features		PC	r	Raw	Norm.	Best
25-D	$(L_x, L_y, L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2, 4, 8, 16)$	$\tilde{20}$	2	95.3	97.1	97.7
15-D	$(L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2, 4, 8, 16)$	–	2	96.9	96.7	96.9
12-D	$(\nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4, 8, 16), (\nabla^2 C; \sigma = 2)$	12	2	90.8	96.9	96.9
10-D	$(L_x, L_y; \sigma = 1, 2, 4, 8, 16)$	10	2	92.2	96.6	96.6
14-D	$(L_x, L_y; \sigma = 1, 2, 4, 8, 16), (C_x, C_y; \sigma = 2)$	14	2	90.5	96.2	96.2

Object category classification CalTech-4

Image features		PC	r	Raw	Norm.	Best
15-D	$(L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2, 4), (C_{xx}, C_{xy}, C_{yy}; \sigma = 2)$	$\tilde{15}$	2	99.5	99.8	99.8
25-D	$(L_x, L_y, L_{xx}, L_{xy}, L_{yy}; \sigma = 1, 2, 4), (C_x, C_y, C_{xx}, C_{xy}, C_{yy}; \sigma = 2)$	12	2	97.2	98.1	99.8
13-D	$(\lvert\nabla L\rvert, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (\lvert\nabla C\rvert, \nabla^2 C; \sigma = 2)$	$\tilde{13}$	2	96.5	99.7	99.7
15-D	$(L, \lvert\nabla L\rvert, \nabla^2 L; \sigma = 1, 2, 4), (C, \lvert\nabla C\rvert, \nabla^2 C; \sigma = 2)$	$\tilde{12}$	2	97.4	99.6	99.7
10-D	$(L_x, L_y; \sigma = 1, 2, 4), (C_x, C_y; \sigma = 2)$	10	2	97.5	99.6	99.6
15-D	$(L, L_x, L_y; \sigma = 1, 2, 4), (C, C_x, C_y; \sigma = 2)$	15	2	98.1	99.6	99.6
20-D	$(L, \lvert\nabla L\rvert, \nabla^2 L, \det \mathcal{H} L; \sigma = 1, 2, 4), (C, \lvert\nabla C\rvert, \nabla^2 C, \det \mathcal{H} C; \sigma = 2)$	$\tilde{15}$	2	98.4	99.3	99.6

Table 16: The 5 best PCA-reduced histogram descriptors when performing *object category classification* using *binary histograms* on the ETH-80 and CalTech-4 data sets. The column labelled "Best" shows the result for the best number of principal components shown in the column labelled PC. A tilde ˜ over the number of principal components PC indicates that the result without variance normalization was better than the result obtained with variance normalization. A dash "–" in the PC column indicates that PCA did not lead to any improvement in performance.

low/blue and edge/non-edge in grey-level information and/or the chromatic channels.

Figure 8 shows an example of computing such binary features from the second-order grey-level structure of two images from the ETH-80 data set. Please, note how well the sign of the second-order derivatives reflect the qualitative shape of a cup. For the horse images, we get fragmentary representative shapes of different parts of a horse, where the body of the horse is reflected at $\sigma = 8$, while the individual legs become visible at $\sigma = 4$ and $\sigma = 2$.

Experimental results. Table 15 gives a summary of performance values obtained using $r = 2$ bins for the problem of recognizing previously seen objects from novel views. Table 16 shows corresponding results for the two category classification benchmarks.

Table 17 gives a summary of the best performance values for each one of the six different benchmark problems using either (i) regular composed histograms without PCA, (ii) PCA-reduced histograms as proposed in section 5.1, or (iii) binary histograms. As can be seen from the table, with dimensionality reduction by PCA even binary histograms do often give a similar or even better performance than regular histograms without dimensionality reduction. The only exception is the object recognition task on the ETH-80 90° data set, where the error rate is about 25 % greater.

From a brief inspection, one may at first regard the recognition performance of the binary histograms as surprisingly high. Plain binary thresholding at a single image point can neither be regarded as robust nor discriminative. In combination with higher-dimensional histograms of multi-scale differential image features, however, these experiments show that such information can be highly discriminative. If we use a larger number of bins $r > 2$ for accumulating the histograms, this simply means that a more fine-grained classification is made at each image point prior to the computation of the histograms.

In connection with these binary histograms it is also worth noting that with regard to object instance recognition, very good performance can also be obtained with *trinary histograms* using $r = 3$ bins per dimension. A trinary histogram also implies a qualitative judgement of the local image structure in the sense that a polarity dependent image feature is locally classified as either (i) clearly positive, (ii) around zero or (iii) clearly negative.

In contrast to the binary histograms, the trinary histograms are, however, not parameter free. They depend on a threshold for distinguishing values that are to be regarded as close to zero from values that are either clearly positive or clearly negative. There is a possibility of optimizing the level of this threshold, although no such effort has been made in this treatment. The trinary histograms will, however, usually be more compact than binary histograms, since values classified as around zero will fall into the same bin (and such values are very common).

		Raw	PCA-reduced	binary PCA-reduced	trinary PCA-reduced
COIL-100	30°	99.9 %	100.0 %	100.0 %	100.0 %
COIL-100	60°	98.4 %	99.0 %	98.3 %	99.0 %
ETH-80	45°	99.8 %	100.0 %	99.9 %	99.9 %
ETH-80	90°	98.7 %	99.4 %	98.4 %	98.9 %
ETH-80	Class	96.9 %	97.7 %	97.7 %	95.9 %
CalTech-4	Class	99.8 %	99.9 %	99.8 %	99.7 %

Table 17: Comparison of the best performance values obtained using the best histogram descriptors within the classes of (i) regular histograms, (ii) PCA-reduced histograms, (iii) binary histograms of PCA-reduced scale-space features and (iv) trinary histograms of PCA-reduced scale-space features.

Beyond binary and trinary histograms, we can also note that if we include the cases with $r = 4$ or $r = 5$ bins per dimension,[7] then we reach performance values that are either among or near the best for all these benchmark problems.

7. Dependency on viewsphere sampling

Due to the dense sampling of the viewing directions in the COIL-100 data set, we can use the image data for evaluating the robustness of different types of image descriptors under variations in the viewing direction. In [39], we present results of such view variation experiments that show that the logarithm of the error increases approximately linearly with the viewsphere sampling $\Delta\phi$. Hence, we get the following approximate error estimate

$$\varepsilon = 1 - p \approx 10^{C + \Delta\phi / \Phi} \tag{19}$$

where a least-squares fitting of a linear model to $\log \varepsilon$ without PCA gives $\Phi \approx 49°$ for the grey-level descriptors and $\Phi \approx 37°$ for the colour descriptors. For the best grey-level descriptor on this data set we have $C \approx -3.0$ and for the best colour descriptor $C \approx -4.0$. When complemented by dimensionality reduction, the errors become significantly lower.

8. Comparisons with other work

In this section, we compare the performance of our complex-cue histogram descriptors to previously reported results for similar data sets.

[7] A histogram with $r = 4$ bins implies that a qualitative judgement is made at every image point whether the value is (i) strongly positive, (ii) weakly positive, (iii) weakly negative, or (iv) strongly negative. With $r = 5$ bins implies a qualitative judgement as either (i) strongly positive, (ii) clearly positive, (iii) around zero, (iv) clearly negative, or (v) strongly negative.

8.1. COIL-100

Table 18 shows how the performance of one of our PCA-reduced histograms compares to other methods when applied to object instance recognition using a viewsphere representation along the equator for the COIL-100 data set. As representative for our rich set of histogram features, we have chosen the 13-D histogram descriptor with $(|\nabla L|, \nabla^2 L, \det \mathcal{H} L)$ at $\sigma = 1, 2, 4$ and $\nabla^2 C$ at $\sigma = 2$ and computed the recognition performance with viewsphere samplings of 20°, 45° and 90° using *all* the intermediate views for recognition. Thus, the recognition rates represent the *average performance* in contrast to the previously reported results in sections 4–5 that reflect the worst-case performance. This 13-D descriptor was chosen because it is had the best worst-case performance on the COIL-100 data set with 60° viewsphere sampling.

As can be seen from the table, our 13-D histogram descriptor compares very favourably to previously reported results on this data set, such as local affine frames (Obdržálek and Matas [50]), spin-glass Markov random fields (Caputo *et al.* [8]) or sparse networks of linear units based on either edges or intensities (Yang *et al.* [63]). The 13-D histogram descriptor with PCA does also perform significantly better than the 14-D descriptor we reported in [38]. Using a combination of multi-dimensional receptive field histograms and colour histograms that were classified by a support vector machine, Nilsback and Caputo [49] obtained a 98.2 % recognition rate with a viewsphere sampling of 30°. With such a 30° viewsphere sampling, the performance of our 13-D histogram descriptor is 100.0 %. Hence, in all these cases our 13-D histogram descriptor performs better than previously reported results on the COIL-100 data set.

8.2. ETH-80

For the ETH-80 data set, we are not aware of any previously reported results regarding recogni-

	COIL-100		
	20°	45°	90°
13-D ($\|\nabla L\|, \nabla^2 L, \det \mathcal{H}L$) ($\nabla^2 C$) PCA, 10 PC, $r = 3$	100.0 %	99.9 %	98.6 %
14-D ($\|\nabla L\|, \nabla^2 L$) ($\|\nabla C\|, \nabla^2 C$) $r = 15$ (Linde and Lindeberg [38])	99.9 %	99.4 %	97.1 %
Local affine frames (Obdržálek and Matas [50])	99.9 %	99.4 %	94.7 %
Spin-glass MRF (Caputo *et al.* [8])	96.8 %	88.2 %	69.4 %
Nearest neighbour (Yang *et al.* [63])	87.5 %	79.5 %	74.6 %

Table 18: Performance results obtained for *object instance recognition on COIL-100* with viewsphere sampling rates of 20°, 45° and 90°, respectively. The number of training views were 18, 8 and 4 respectively and the number of test images 5400, 6400 and 6800 respectively.

	ETH-80
25-D ($L_x, L_y, L_{xx}, L_{xy}, L_{yy}$) PCA, 20 PC, $r = 2$	97.7 %
Nilsback and Caputo [49] multi-cue decision tree	97.1 %
12-D ($\nabla^2 L, \det \mathcal{H}L$)($\nabla^2 C$) PCA, 12 PC, $r = 2$	96.9 %
Nilsback and Caputo [49] dual cue voting	96.4 %
Nilsback and Caputo [49] single cue	93.9 %
Linde and Lindeberg [38] 6-D (L_x, L_y) $r = 15$	93.1 %
Leibe and Schiele [35] best multi-cue decision tree	93.0 %
Leibe and Schiele [35] best single-cue	86.4 %
Eichhorn and Chapelle [15] interest points + SIFT	85 %
Grauman and Darrell [24] Harris points + PCA-SIFT	83 %

Table 19: Performance results obtained for *category classification on ETH-80*.

	CalTech-4
10-D (L_x, L_y) (C_x, C_y) PCA, 6 PC, $r = 6$	99.9 %
15-D (L_{xx}, L_{xy}, L_{yy}) (C_{xx}, C_{xy}, C_{yy}) PCA, 8 PC, $r = 4$	99.9 %
15-D (L_{xx}, L_{xy}, L_{yy}) (C_{xx}, C_{xy}, C_{yy}) PCA, 15 PC, $r = 2$	99.8 %
8-D ($\nabla^2 L, \det \mathcal{H}L$)($\nabla^2 C$) PCA, 8 PC, $r = 5$	99.7 %
Nilsback and Caputo [49] multi-cue decision tree	99.5 %
Linde and Lindeberg [38] 9-D ($\|\nabla L\|, \nabla^2 L, \det \mathcal{H}L$) $r = 15$	99.2 %
Nilsback and Caputo [49] multi-cue voting	98.5 %
Nilsback and Caputo [49] single cue	96.9 %

Table 20: Performance results obtained for *category classification on CalTech-4*.

tion of specific object instances. Hence, we focus on object category classification in this comparison summarized in table 19.

As representatives for our histogram features, we have chosen the 25-D rotationally variant histogram of ($L_x, L_y, L_{xx}, L_{xy}, L_{yy}$) at $\sigma = 1, 2, 4, 6, 16$ and the 12-D histogram of the rotationally invariant differential invariants ($\nabla^2 L, \det \mathcal{H}L$) at $\sigma = 1, 2, 4, 6, 16$ and $\nabla^2 C$ at $\sigma = 2$. These descriptors reach performance values of 97.7 % and of 96.9 % respectively.

For this data set Leibe and Schiele [35] and Nilsback and Caputo [49] got 86.4 % and 93.9 % performance respectively using their best single cues, 93.0 % using the multi-cue decision tree in [35] and 96.4 % using the dual cue voting scheme in [49]. Notably, our best complex-cue histogram features, which turn out to be binary histograms, give classification rates that are better than those multi-cue schemes. With a multi-cue decision tree, Nilsback and Caputo [49] reached a performance of 97.1 %, *i.e.*, below the performance of our best rotationally variant descriptor but somewhat better than our best rotationally invariant descriptor. Object classification in terms of interest points lead to a performance of about 85 % (Eichhorn and Chapelle [15], Grauman and Darrell [24]), which is not at all comparable to the performance of multi-scale histograms on this data set.

8.3. CalTech-4

Table 20 shows a comparison with regard to category classification on the CalTech-4 data set. Here, we have chosen the rotationally variant 10-D histogram descriptor with (L_x, L_y) at $\sigma = 1, 2, 4$ and (C_x, C_y) at $\sigma = 2$ and the 15-D rotationally variant histogram with (L_{xx}, L_{xy}, L_{yy}) at $\sigma = 1, 2, 4$ and (C_{xx}, C_{xy}, C_{yy}) at $\sigma = 2$. Both reach a performance of 99.9 % when reduced to 6 and 8 dimensions respectively using PCA. The performance of

our best binary histogram on this data set, also with (L_{xx}, L_{xy}, L_{yy}) at $\sigma = 1, 2, 4$ and (C_{xx}, C_{xy}, C_{yy}) at $\sigma = 2$, is 99.8 %. With an 8-D rotationally invariant histogram of the variance normalized scale-space features $(\nabla^2 L, \det \mathcal{H} L)$ at $\sigma = 1, 2, 4$ and $\nabla^2 C$ at $\sigma = 2$ the classification performance is 99.7 %.

Nilsback and Caputo [49] obtained 96.9 % using their best single cue and 98.5 % using a voting scheme on multiple cues. With multi-cue decision tree they reached 99.5 %. Our single complex-cues histogram features are significantly better than this. Our best rotationally invariant histogram and our best binary histogram are also better.

Qualitatively, we can also relate these results to Fergus *et al.* [18], who report equal error rates from an ROC curve in the range 90–96 % depending on the object class. These error rates are, however, not directly comparable, since our experiment concerns a multi-category problem, whereas Fergus *et al.* [18] consider a category detection problem, with explicit modelling of the background class.

9. Summary and conclusions

We have introduced a rich set of histogram descriptors for view-based object recognition and evaluated them with respect to two types of recognition problems using three mutually independent data sets. These histogram descriptors can be defined and computed either globally, regionally or locally. Moreover, they can be computed efficiently and be classified with low computational cost.

The proposed image descriptors are formulated in terms of different combinations of partial derivatives and differential invariants up to order two, computed from either grey-level information, spatio-chromatic colour-opponent channels or both, and capture the statistics of local differential geometric image structure.

Whereas the histograms themselves treat the image measurements at different points completely independently and without any explicit encoding of relations between the image structures at different points within the support region, the formulation of histograms in terms of differential entities defined from Gaussian derivatives implies that such relations between image structures at different points are nevertheless encoded implicitly by the comparisons between neighbouring regions that are performed when computing the underlying derivatives and/or differential invariants.

9.1. Composed complex-cue histograms

Our motivation for studying composed descriptors of higher dimensionality is to capture more of the information content in the local image structure than is reflected in previously used histogram features. From scale-space theory, we obtain a natural classification of local image operations, and we have used this classification of local image features to state an exhaustive catalogue of natural image descriptors up to second-order. The experiments have clearly shown that it is possible to define *composed complex-cue image descriptors* that capture the *co-variation* of different primitive image cues and that these may lead to *significant improvements* in recognition performance compared to previously used histogram descriptors of primitive image cues, which in turn are of lower dimensionality.

9.2. Handling the bin size

The number of bins for accumulating the histograms constitutes a free parameter, and the choice of a suitable number of quantization levels depends on both the distribution and the availability of image measurements. Thus, there is a trade-off between the accuracy of estimating the distribution and the availability as well as the accuracy of the underlying image measurements.

In this work, we have performed a coarse optimization over the number of histogram bins to optimize the performance and also to be able to compare qualitatively different types of histogram descriptors without having the comparison biased by any *ad hoc* choice of bin size. We have also seen that even *binary histograms* may give surprisingly good performance and that such descriptors, which reveal qualitative information in terms of the polarity of image features, can be expected to have good robustness properties under illumination variations.

For the task of object instance recognition, also *trinary histograms*, which imply local classifications in terms of strongly positive, around zero or strongly negative, lead to very good performance. More generally, coarsely quantized histograms with as few as $r = 2$ to $r = 5$ bins per dimension are usually sufficient to obtain recognition or classification performance among or near the best for these datasets.

In the trade-off between a larger number of bins or a larger number of dimensions in the descriptor, it is worth noting that the experimental results often seem to favour higher-dimensional descriptors with a low number of bins.

9.3. General conclusions from the experiments

We have performed an extensive experimental evaluation of a very rich set of image descriptors on primarily six different benchmark problems with regard to the problems of either (i) recognizing previously seen objects from novel views or (ii) classifying previously unseen objects into object categories. For both types of problems, we have presented several histogram descriptors with very good performance. The set of image descriptors that performs best for object recognition does, however, differ from the set of image descriptors that performs best for object categorization.

A general result throughout the experiments is that the recognition or classification performance can be improved by *combining primitive scale-space features* into composed image descriptors that capture the simultaneous co-variation of multiple cues in terms of *composed complex cue histograms,* provided that a sufficient set of representative training data is available for capturing the distributions of the underlying the higher-dimensional histograms.

We have also showed that the performance of histogram-based methods for recognition can be significantly improved by performing *dimensionality reduction* of the underlying image features prior to the accumulation of histograms. For most combinations of scale-space features, the performance can also be improved by *variance normalization* prior to principal component analysis.

9.4. Object instance recognition

When recognizing previously seen objects from novel views, image descriptors that contain combinations of grey-level information and explicit colour cues in terms of *higher-dimensional histograms that combine Gaussian derivatives and spatio-chromatic colour-opponent derivatives* do in general lead to a significant increase in performance compared to more traditional approaches based on lower-dimensional grey-level cues or regular colour histograms. Indeed, all of the best image descriptors for this task contain explicit spatio-chromatic derivatives. Moreover, when there are variations in the viewing direction between the training and the testing data, image descriptors in terms of *rotationally invariant differential invariants* lead to much better performance compared to rotationally variant histogram descriptors. In relation to previous work in this area, we have in particular emphasized the usefulness of *spatio-chromatic differential invariants defined from colour-opponent channels* and *the importance of second-order image cues*.

For the task of recognizing object instances on the COIL-100 and ETH-80 datasets, image descriptors based on zero-order image intensities and chromatic channels performed surprisingly well, where the best results were obtained with (i) a 10-D histogram of the grey-level cues $(L, |\nabla L|)$ at three scales and the spatio-chromatic cues $(C, |\nabla C|)$ at one scale and (ii) a 15-D histogram of the grey-level cues $(L, |\nabla L|, \nabla^2 L)$ at three scales and the spatio-chromatic cues $(C, |\nabla C|, \nabla^2 C)$ at one scale (see table 12). A possible explanation for this may be that these datasets have been acquired under controlled illumination conditions, which means that image intensities and chromaticity values are preserved under large viewing variations, whereas the values of higher-order derivative operators may be more affected by larger image deformations.

Image descriptors that are based on raw image intensities may, however, be very sensitive if the illumination conditions are changed. In such situations, image descriptors that are based on spatial and spatio-chromatic derivatives can be expected to be more robust, since they are invariant under additive illumination variations. For this subset of image descriptors, the best results were obtained with (i) a 10-D histogram of the mixed first- and second-order differential invariants $(|\nabla L|, \nabla^2 L)$ at three scales and $(|\nabla C|, \nabla^2 C)$ at one scale. Very good performance was also obtained for other subsets of these scale-space features (see table 13).

9.5. Object classification

When classifying previously unseen objects into categories, *pure grey-level cues* are much stronger, although spatio-chromatic derivatives may also be valuable depending on the categories involved. A new finding is also that *second-order cues* in terms of partial derivatives and/or differential invariants are highly discriminative, whereas zero-order cues in terms of explicit intensity or chromaticity values are not very useful, at least for these datasets.

When we allow for dimensionaity reduction of the underlying scale-space features by PCA, the dominance of second-order information becomes less manifest. A possible explanation for this may be that the principal components correspond to *differences between scale-space features at different scales*. By an interpretation of derivatives with

respect to scale as spatial Laplacians, which follows from the diffusion equation, such differences between scales can in turn be interpreted in terms of higher order spatial derivatives. Nonwithstanding this, we may nevertheless conclude that *higher-order spatial derivatives provide important cues for category classification*.

Moreover, provided that the orientations of the objects are similar in the training and the testing data, the performance is generally better for *rotationally variant image descriptors* in terms of partial derivatives compared to descriptors based on rotationally invariant scale-space features, where the best results were obtained for different combinations of the partial derivatives (L_x, L_y) and (L_{xx}, L_{xy}, L_{yy}) as well as the spatio-chromatic colour-opponent derivatives (C_x, C_y) and (C_{xx}, C_{xy}, C_{yy}) (see table 14).

There are, however, also rotationally invariant image descriptors that also lead to very good performance, where the best rotationally invariant image descriptor is (i) a 9-D histogram of the independent rotationally invariant differential invariants $(|\nabla L|, \nabla^2 L, \det \mathcal{H} L)$, computed from grey-level information only.

These conclusions have been obtained using three mutually independent data sets, where two mutually exclusive data sets have been used for each type of recognition problem (object recognition *vs.* category classification) and the results are in agreement between the mutually independent data sets.

9.6. Extensions and future work

Concerning possible limitations of the experimental part of work, our experiments in sections 4–6 have all been concerned with global histograms, motivated by our aim to evaluate the information content in different types of image descriptors without having the results influenced by the performance of other visual modules, such as object localization or segmentation. For the image datasets considered here, such an approach is reasonable, since the objects to be recognized usually cover a major part of the image. An extension of sparse multi-cue histograms to a more general setting with regional histograms is outlined in [39].

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 and spatio-temporal recognition. We are addressing these topics in on-going work.

To conclude, the overall performance of the proposed complex-cue histogram descriptors must be regarded as highly competitive.

Acknowledgements

We would like to thank Lars Bretzner for valuable discussions and suggestions as well as the anonymous reviewers for valuable comments that improved the presentation. The support from the Swedish Research Council, Vetenskapsrådet (contracts 2004-4680, 2010-4766), and from the Royal Academy of Sciences as well as the Knut and Alice Wallenberg Foundation is gratefully acknowledged. An earlier version of this work was presented in [38].

Appendix A. Statistical analysis of significance and confidence levels

To analyze the statistical significance of the results for the different benchmark problems considered in this paper, let us make the simplifying assumption that the classifier can be viewed as a function that given a random sample yields an incorrect classification with a probability p_e independent of all other factors. The performance values presented previously can then be regarded as estimates of $1 - p_e$, given the observed outcome of classifying a large number n of samples.

When n is large and p_e is low, the actual number of observed misclassifications u, which by the assumptions follow a binomial distribution, can be approximated by a Poisson distribution

$$f(u, \lambda) = \frac{\lambda^u e^{-\lambda}}{u!}, u = 0, 1, 2, \ldots, \lambda > 0 \qquad \text{(A.1)}$$

with $\lambda = p_e\, n$. Furthermore, if u errors are found in a classification task, the maximum likelihood estimate of λ is $\hat{\lambda} = u$.

Based on a series of observations of the assumed random classification process, there is an inherent uncertainty in the estimated $\hat{\lambda}$, and therefore also in the derived estimated error rate $\hat{p}_e = \hat{\lambda}/n$. This uncertainty can be quantified by the computation of a confidence interval $[a, b]$ that covers the true value of λ being estimated with probability $1 - \alpha$:

$$\Pr\{a \leq \lambda \leq b \,|\, u\} = 1 - \alpha \qquad \text{(A.2)}$$

A central (approximate) confidence interval, *i.e.* an interval where the tail area of the probability distribution outside of either edge of the interval $[a, b]$ is $\alpha/2$,

Benchmark	n	u	p_L (%)	$\hat{p}_e$ (%)	p_U (%)
COIL-100 30°	7 200	0	0	0	0.05
COIL-100 60°	7 200	70	0.76	0.97	1.23
ETH-80 45°	1 280	0	0	0	0.28
ETH-80 90°	1 280	8	0.27	0.63	1.23
ETH-80 Class	1 280	29	1.52	2.27	3.25
CalTech-4 Class	1 417	1	0.00	0.07	0.39

Table A.21: 95 % central confidence intervals $[p_L, p_U]$ covering the maximum-likelihood estimate $\hat{p}_e$ for the error rate of the best descriptor on each benchmark problem. u denotes the actual number of misclassified images and n the total number of images tested.

may then be written as [26, page 96]:

$$0 \leq a = \frac{1}{2}(\chi 2)^{-1}_{2u}\left(\frac{\alpha}{2}\right) \leq \lambda \leq \frac{1}{2}(\chi 2)^{-1}_{2(u+1)}\left(1-\frac{\alpha}{2}\right) = b \tag{A.3}$$

where $(\chi 2)^{-1}_d(v)$ is the v:th quantile of a $\chi 2$-distribution with d degrees of freedom.

Table A.21 shows estimated 95 % confidence intervals ($\alpha = 0.05$) for the best results for each benchmark problem. For COIL-100 the number of tested samples is $n = 7\,200$ and the number of classification errors u with a 30° viewsphere is 0 for the best descriptors. An upper limit p_U for the probability $p_e = \lambda/n$ for the best image descriptor is therefore $p_U \approx 0.05$ %. The results expressed as percentages are therefore statistically significant up to the first decimal digit for performance values close to 100.0 %, given the assumptions above. When the viewsphere sampling is increased to 60°, where the best image descriptors have an ≈ 1.0 % error rate, corresponding calculations give that a 95 % confidence interval for the error rate is given by 0.76 % $\leq p_e \leq$ 1.23 %.

For instance recognition on the ETH-80 dataset, the number of tested samples is $n = 1\,280$ and the corresponding 95 % confidence upper limit for the estimated probability is $p_U \approx 0.28$ % with a 45° viewsphere sampling, which means that the relative ranking for results ≥ 99.7 % will be uncertain. With a 90° viewsphere sampling, the corresponding upper bound is $p_U \approx 1.2$ %.

For category classification on the ETH-80 dataset, the number of samples is 1 280 and a 95 % confidence interval for the estimated probability becomes 1.5 % $\leq p_e \leq$ 3.3 %. For the Caltech-4 dataset, we use 1 417 test samples, and a 95 % upper confidence limit for the estimated probability is $p_U \approx 0.4$ % for values close to 100 %.

References

[1] Agarwal, A. and Triggs, B. [2008], 'Multilevel image coding with hyperfeatures', *Int. J. of Computer Vision* **78**(1), 15–27.

[2] Baumberg, A. [2000], Reliable feature matching across widely separated views, *in* 'Proc. CVPR', Hilton Head, SC, pp. I:1774–1781.

[3] Bay, H., Ess, A., Tuytelaars, T. and van Gool [2008], 'Speeded up robust features (SURF)', *Computer Vision and Image Understanding* pp. 346–359.

[4] Belongie, S., Fowlkes, C., Chung, F. and Malik, J. [2002], Partitioning with indefinite kernels using the Nyström extension, *in* 'Proc. ECCV'02', Vol. 2352 of *Lecture Notes in Computer Science*, Springer, Copenhagen, Denmark, pp. III:531–542.

[5] Bosch, A., Zisserman, A. and Munoz, X. [2007], Image classification using random forests and ferns, *in* 'Proc. 11th International Conference on Computer Vision', Rio de Janeiro, Brazil, pp. 1–8.

[6] Broadhurst, R. E. [2005], Statistical estimation of histogram variation for texture classification, *in* M. Chantler and O. Drbohlav, eds, 'Proc. Fourth Int'l Workshop Texture Analysis and Synthesis', pp. 25–30.

[7] Burghouts, G. J. and Geusebroek, J.-M. [2009], 'Performance evaluation of local colour invariants', *Computer Vision and Image Understanding* **113**(1), 48–62.

[8] Caputo, B., Hornegger, J., Paulus, D. and Niemann, H. [2002], A spin-glass markov random field for 3-D object recognition, Technical Report LME-TR-2001-01, Computer Science Department, University of Erlangen-Nürnberg.

[9] Caputo, B., Wallraven, C. and Nilsback, M. E. [2004], Object categorization via local kernels, *in* '17th International Conference on Pattern Recognition (ICPR'04)', Vol. 2, Cambridge, U.K., pp. 132–135.

[10] Chang, C.-C. and Lin, C.-J. [2001], 'LIBSVM: a library for support vector machines'. Software available from `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[11] Chapelle, O., Haffner, P. and Vapnik, V. [1999], 'SVMs for histogram-based image classification', *IEEE Transactions on Neural Networks* **10**(5), 1055–1065.

[12] Chomat, O., Martin, J. and Crowley, J. [2000], A probabilistic sensor for the perception and recognition of activities, *in* 'Proc. ECCV'00', Vol. 1842 of *Lecture Notes in Computer Science*, Springer-Verlag, Dublin, Ireland, pp. I:487–503.

[13] Csurka, G., Dance, C. R., Fan, L., Willamowski, J. and Bray, C. [2004], Visual categorization with bags of keypoints, *in* 'In ECCV Workshop on Statistical Learning in Computer Vision', Prague, Czech Republic.

[14] Dalal, N. and Triggs, B. [2005], Histograms of oriented gradients for human detection, *in* 'Proc. CVPR', Vol. 1, pp. 886–893.

[15] Eichhorn, J. and Chapelle, O. [2004], Object categorization with SVM: Kernels for local features, Technical Report 137, Max Planck Institute for Biological Cybernetics.

[16] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A. [2009], 'The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results', http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html.

[17] Fei-Fei, L., Fergus, R. and Perona, P. [2004], Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories., *in* 'IEEE CVPR Workshop on Generative-Model Based Vision'. http://www.vision.caltech.edu/Image_Datasets/Caltech101.

[18] Fergus, R., Perona, P. and Zisserman, A. [2003], Object class recognition by unsupervised scale-invariant learning, *in* 'Proc. CVPR', Madison, Wisconsin, pp. 264–271.

[19] Florack, L. M. J. [1997], *Image Structure*, Series in Mathematical Imaging and Vision, Springer.

[20] Friedman, J. H. [1996], Another approach to polychotomous classification, Technical report, Department of Statistics, Stanford University.

[21] Geusebroek, J.-M., van den Boomgaard, R., Smeulders, A. W. M. and Geerts, H. [2001], 'Color invariance', *IEEE Trans. Pattern Analysis and Machine Intell.* **23**(12), 1338–1350.

[22] Gevers, T. and Smeulders, A. W. M. [1999], 'Color-based object recognition', *Pattern Recognition Letters* **32**, 453–464.

[23] Goldstein, E. B. [2002], *Sensation and Perception*, Wadsworth-Thomson. Sixth Edition.

[24] Grauman, K. and Darrell, T. [2005], The pyramid match kernel: Discriminative classification with sets of image features, *in* 'Proc. 10th International Conference on Computer Vision', Vol. 2, Beijing, China, pp. 1458–1465.

[25] Hall, D., de Verdiere, V. and Crowley, J. [2000], Object recognition using coloured receptive fields, *in* 'Proc. ECCV'00', Vol. 1842 of *Lecture Notes in Computer Science*, Springer-Verlag, Dublin, Ireland, pp. I:164–177.

[26] Johnson, N. L. and Kotz, S. [1969], *Discrete Distributions*, Distributions in Statistics, Houghton Mifflin, Boston, Massachusetts.

[27] Jurie, F. and Triggs, B. [2005], Creating efficient codebooks for visual recognition, *in* 'Proc. 10th International Conference on Computer Vision', Vol. 1, Beijing, China, pp. 17–21.

[28] Ke, Y. and Sukthankar, R. [2004], PCA-SIFT: A more distinctive representation for local image descriptors, *in* 'Proc. CVPR', Washington D. C., pp. II: 506–513.

[29] Koenderink, J. J. [1984], 'The structure of images', *Biological Cybernetics* **50**, 363–370.

[30] Koenderink, J. J. and van Doorn, A. J. [1992], 'Generic neighborhood operators', *IEEE Trans. Pattern Analysis and Machine Intell.* **14**(6), 597–605.

[31] Koenderink, J. and van Doorn, A. [1999], 'The structure of locally orderless images', *Int. J. of Computer Vision* **31**(2/3), 159–168.

[32] Kreßel, U. H.-G. [1999], Pairwise classification and support vector machines, *in* B. Schölkopf, C. J. C. Burges and A. J. Smola, eds, 'Advances in kernel methods: Support vector learning', MIT Press, Cambridge, MA, USA, pp. 255–268.

[33] Laptev, I. and Lindeberg, T. [2004], Local descriptors for spatio-temporal recognition, *in* 'Proc. ECCV'04 Workshop on Spatial Coherence for Visual Motion Analysis', Vol. 3667 of *Lecture Notes in Computer Science*, Springer, Prague, Czech Republic, pp. 91–103.

[34] Lazebnik, S., Schmid, C. and Ponce, J. [2006], Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *in* 'Proc. CVPR', Washington, DC, USA, pp. 2169–2178.

[35] Leibe, B. and Schiele, B. [2003], Analyzing appearance and contour based methods for object categorization, *in* 'Proc. CVPR', Madison, Wisconsin, pp. II:409–415.

[36] Linde, O. and Bretzner, L. [2009], Local histogram based descriptors for recognition, *in* 'Proc. of the Fourth International Conference on Computer Vision Theory and Applications VISAPP'09', Lisboa, Portugal, pp. 332–339.

[37] Linde, O., Bretzner, L. and Lindeberg, T. [2009], 'Dense local clustered and contrast normalized histogram descriptors for spatial recognition'. Submitted.

[38] Linde, O. and Lindeberg, T. [2004], Object recognition using composed receptive field histograms of higher dimensionality, *in* 'International Conference on Pattern Recognition', Vol. 2, Cambridge, pp. 1–6.

[39] Linde, O. and Lindeberg, T. [2011], Composed complex-cue histograms: An investigation of the information contents in receptive-field-based image descriptors for object recognition, Technical report, School of Computer Science and Communication, KTH (Royal Institute of Technology), Stockholm, Sweden. In preparation.

[40] Lindeberg, T. [1994*a*], 'Scale-space theory: A basic tool for analysing structures at different scales', *Journal of Applied Statistics* **21**(2), 225–270. Also available from http://www.csc.kth.se/∼tony/abstracts/Lin94-SI-abstract.html.

[41] Lindeberg, T. [1994*b*], *Scale-Space Theory in Computer Vision*, The Kluwer International Series in Engineering and Computer Science, Springer.

[42] Lindeberg, T. [1998], 'Feature detection with automatic scale selection', *Int. J. of Computer Vision* **30**(2), 77–116.

[43] Lindeberg, T. [2008], Scale-space, *in* B. Wah, ed., 'Encyclopedia of Computer Science and Engineering', John Wiley and Sons, Hoboken, New Jersey, pp. 2495–2504. dx.doi.org/10.1002/9780470050118.ecse609 Also available from http://www.nada.kth.se/∼tony/abstracts/Lin08-EncCompSci.html.

[44] Lindeberg, T. and Gårding, J. [1997], 'Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D structure', *Image and Vision Computing* **15**, 415–434.

[45] Lowe, D. [2004], 'Distinctive image features from scale-invariant keypoints', *Int. J. of Computer Vision* **60**(2), 91–110.

[46] Mikolajczyk, K. and Schmid, C. [2004], 'Scale and affine invariant interest point detectors', *Int. J. of Computer Vision* **60**(1), 63–86.

[47] Mikolajczyk, K. and Schmid, C. [2005], 'A performance evaluation of local descriptors', *IEEE Trans. Pattern Analysis and Machine Intell.* **27**(10), 1615–1630.

[48] Nene, S. A., Nayar, S. K. and Murase, H. [1996], Columbia object image library (COIL-100), Technical report CUCS-006-96, CAVE, Columbia University.

[49] Nilsback, M. and Caputo, B. [2004], Cue integration through discriminative accumulation, *in* 'Proc. CVPR', Vol. 2, Los Alamitos, CA, pp. 578–585.

[50] Obdržálek, S. and Matas, J. [2002], Object recognition using local affine frames on distinguished regions, *in* 'Proc. British Machine Vision Conference', Cardiff, U.K., pp. 113–122.

[51] Schiele, B. and Crowley, J. [1996], Object recognition using multidimensional receptive field histograms, *in* 'Proc. ECCV'96', Vol. 1064 of *Lecture Notes in Computer Science*, Springer, Cambridge, UK, pp. 610–619.

[52] Schiele, B. and Crowley, J. [2000], 'Recognition without correspondence using multidimensional receptive field histograms', *Int. J. of Computer Vision* **36**(1), 31–50.

[53] Schneiderman, H. and Kanade, T. [2000], A statistical method for 3D object detection applied to faces and cars, *in* 'Proc. CVPR', Vol. I, Hilton Head, SC, pp. 746–751.

[54] Shechtman, E. and Irani, M. [2007], 'Space-time behavior-based correlation – OR – How to tell if two

underlying motion fields are similar without computing them?', *IEEE Trans. Pattern Analysis and Machine Intell.* **29**(11), 2045–2056.

[55] Slater, D. and Healey, G. [1995], Combining colour and geometric information for illumination invariant recognition of 3-D objects, *in* 'Proc. 5th Int. Conf. on Computer Vision', Cambridge, MA, pp. 563–568.

[56] Swain, M. and Ballard, D. [1991], 'Color indexing', *Int. J. of Computer Vision* **7**(1), 11–32.

[57] ter Haar Romeny, B. [2003], *Front-End Vision and Multi-Scale Image Analysis*, Springer.

[58] Tuytelaars, T. and van Gool, L. [2004], 'Matching widely separated views based on affine invariant regions', *Int. J. of Computer Vision* **59**(1), 61–85.

[59] van de Sande, K. E. A., Gevers, T. and Snoek, C. G. M. [2010], 'Evaluating color descriptors for object and scene recognition', *IEEE Trans. Pattern Analysis and Machine Intell.* **32**(9), 1582–1596.

[60] Vapnik, V. [1998], *Statistical Learning Theory*, John Wiley and Sons, Chichester, England.

[61] Willems, G., Tuytelaars, T. and van Gool, L. [2008], An efficient dense and scale-invariant spatio-temporal interest point detector, *in* 'Proc. ECCV'08', Vol. 5303 of *Lecture Notes in Computer Science*, Springer, Marseille, France, pp. 650–663.

[62] Witkin, A. P. [1983], Scale-space filtering, *in* 'Proc. 8th Int. Joint Conf. Art. Intell.', Karlsruhe, Germany, pp. 1019–1022.

[63] Yang, M.-H., Roth, D. and Ahuja, N. [2000], Learning to recognize 3D objects with SNoW, *in* 'Proc. ECCV'00', Vol. 1842, Springer, Dublin, Ireland, pp. 439–454.

[64] Zelnik-Manor, L. and Irani, M. [2001], Event-based analysis of video, *in* 'Proc. CVPR', Kauai Marriott, Hawaii, pp. II:123–130.