LOCAL HISTOGRAM BASED DESCRIPTORS FOR RECOGNITION

Oskar Linde and Lars Bretzner
CVAP/CSC, KTH, Stockholm, Sweden

Keywords: Image representation, Image descriptor, Histogram, Invariance, Categorization.

Abstract: This paper proposes a set of new image descriptors based on local histograms of basic operators. These descriptors are intended to serve in a first-level stage of an hierarcical representation of image structures. For reasons of efficiency and scalability, we argue that descriptors suitable for this purpose should be able to capture and separate invariant and variant properties. Unsupervised clustering of the image descriptors from training data gives a visual vocabulary, which allow for compact representations. We demonstrate the representational power of the proposed descriptors and vocabularies on image categorization tasks using well-known datasets. We use image representations via statistics in form of global histograms of the underlying visual words, and compare our results to earlier reported work.

1 INTRODUCTION

1.1 Background and Motivation

In this work we explore local image descriptors based on histograms of basic operators, applied on a grid. In previous work (Linde and Lindeberg, 2004), we have shown how global histogram-based descriptors show surprisingly good classification performance on well-known image data sets; here we compare the results with the performance of the proposed local descriptors, giving a much more compact representation.

The representation is based on a dense grid structure applied on the image. Dense grid-based representations have been claimed to better handle recognition and segmentation of a wide range of textures, objects and scenes (Lazebnik et al., 2006; Bosch et al., 2007; Agarwal and Triggs, 2006; Jurie and Triggs, 2005), compared to representations based on sparse local interest points (Schiele and Crowley, 2000; Lowe, 2004; Dorkó and Schmid, 2005; Csurka et al., 2004). One main reason is that methods relying on interest point detectors naturally show poor recognition/classification performance in image areas where such detectors give no or few responses. Representations including local histograms have been proposed by e.g. (Koenderink and Doorn, 1999) and there are many examples of histogram-based image descriptors in the literature, both global (Schiele and Crowley, 2000; Nilsback and Caputo, 2004), and local (Lowe, 2004; Puzicha et al., 1999; Schmid, 2004). One advantage of such descriptors is that they show robustness to small image perturbations, like noise, minor occlusions, translations and distortions.

The proposed descriptors have been tested on object classification problems present in a number of image data sets frequently used for testing image classification frameworks. An unsupervised clustering of the descriptor responses from the training set gives a visual vocabulary. We look upon this vocabulary as a possible building block of higher-lever, semi-local descriptors. In this work however, in order to test the descriptors we use the vocabulary to represent each object class as a global histogram of words. Similar techniques have been used by e.g. (Csurka et al., 2004; Dorkó and Schmid, 2005; Fei-Fei and Perona, 2005). The here proposed multi-scale descriptors are rotationally invariant. We show how a contrast normalization procedure makes the descriptors invariant to contrast changes that could be caused by e.g. varying illumination, and how the normalization increases classification performance. The presented work include comparisons to earlier similar global descriptors for classification problems.

1.2 Invariant Image Descriptors

It is advantageous to separate the image data as far as possible into independent components. This separation should be done at a low level, while still keeping the possibility to model joint probabilities on higher levels. When the data is as separated as possible, the system can at a low level learn structures and features

Linde O. and Bretzner L. (2009).
LOCAL HISTOGRAM BASED DESCRIPTORS FOR RECOGNITION.
In *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, pages 333-339
DOI: 10.5220/0001793103330339

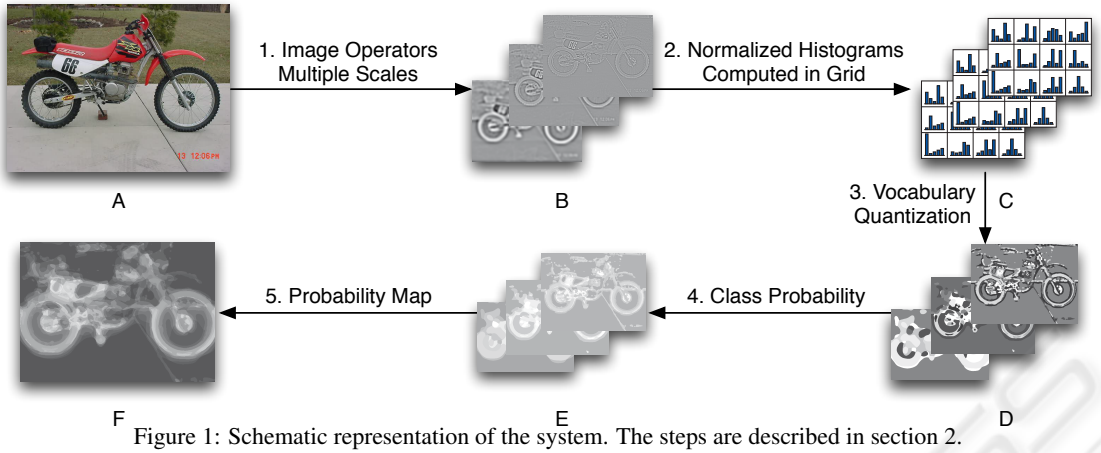

Figure 1: Schematic representation of the system. The steps are described in section 2.

independently of other separated components. This means that a much smaller amount of training data is needed to learn the variations over certain separated feature components. The joint features of the combined components can then be learned at a higher level, at a vastly reduced domain of data.

2 REPRESENTATION AND MATCHING

Figure 1 shows an overview of the processing scheme in which we use and test the proposed descriptors. This section describes the scheme, the main part of the paper focuses on the first three steps.

2.1 Image Operators

The local image descriptors we study in this work are built up by the following components:

- Normalized Gaussian derivatives, obtained by computing partial derivatives $(L_x, L_y, L_{xx}, L_{xy}, L_{yy})$ from the scale-space representation $L(\cdot,\cdot;\ t) = g(\cdot,\cdot;\ t) * f$ obtained by smoothing the original image f with a Gaussian kernel $g(\cdot,\cdot;\ t)$, and multiplying the regular partial derivatives by the standard deviation $\sigma = \sqrt{t}$ raised to the order of differentiation (Lindeberg, 1994).
- Differential invariants, invariant to rotations in the image plane, we use the normalized gradient magnitude $|\nabla_{\text{norm}} L| = \sqrt{t(L_x^2 + L_y^2)}$, and the normalized Laplacian $\nabla^2_{\text{norm}} L = t(L_{xx} + L_{yy})$.
- Chromatic cues from RGB-images according to $C_1 = (R-G)/2$ and $C_2 = (R+G)/2 - B$.

Unless otherwise mentioned, all primitives are computed at scale levels $\sigma \in \{1,2,4\}$. For the data sets studied in this work, this choice is reasonable, as they do not include major scale variations.

Using these image primitives, we build the following rotationally invariant point descriptors:

- 3-D $(\nabla^2 L; \sigma = 1,2,4)$ is the Laplacian applied on the gray level scale space image L at the scales $\sigma \in \{1,2,4\}$.
- 5-D $(\nabla^2 L; \sigma = 1,2,4), (\nabla^2 C; \sigma = 2)$ is the Laplacian applied on L at the scales $\sigma \in \{1,2,4\}$ and on the two chromatic channels C at the scale $\sigma = 2$.
- 6-D $(|\nabla L|, \nabla^2 L; \sigma = 1,2,4)$ is the gradient magnitude $|\nabla \cdot|$ and the Laplacian $\nabla^2 \cdot$ applied to L at the scales $\sigma \in \{1,2,4\}$
- 3-D $(L, C; \sigma = 1)$ is analogous to the classic color histogram descriptor of (Swain and Ballard, 1991) applied to L and C at the scale $\sigma = 1$.

This selection of point descriptors have been shown to perform well as components of global histogram based descriptors in previous work, see (Linde and Lindeberg, 2004).

The different descriptors capture different elements of the image. The 3-D color histogram captures the color distribution of an area and is invariant to any structure within the area. The Laplacian operator, ∇L captures structures of a certain size, corresponding to its scale. The resulting histogram of the Laplacian operator applied at different scales will capture the frequency of structures, such as lines of different widths and blobs of different sizes. The $\nabla^2 C$ operator captures the same structures in the color domain.

2.2 Local Receptive Field Histograms

The image is divided into regularly distributed and partially overlapping local image areas. The N-D de-

scriptors as described in section 2.1 are applied to each point of the area, resulting in a N-dimensional feature vector x, at each point. The local image areas are defined by an accumulation function with a square shape 24×24. (A performance study comparing different window sizes and accumulator functions is presented in (Linde and Bretzner, 2008).) The window size is optimized for the experimental setup in this paper, smaller image regions could be more appropriate when the descriptors are used in a feature hierarchy. For each such local image area, a weighted local contrast normalization is performed. This normalization step is further explained in section 2.3. After normalization, the N-dimensional feature vectors of the local image area are quantized and accumulated in a $q_1 \times q_2 \times \ldots \times q_N$-dimensional histogram.

2.3 Contrast Normalization

Contrast normalization is important for several reasons. It makes the system more robust to contrast changes over the image caused by lighting and shadows and reduces the amount of training data required to learn and categorize the various image structures encountered. The local contrast factor is factored out and kept as an independent value for each subwindow. The contrast factor could be used at a later stage for a richer description of the image.

The normalization is applied to local regions to attain invariance to local contrast variations as follows: From a $b \times b$ sized sub-window, N-dimensional feature vectors $\{x_1, x_2, \ldots, x_{b^2}\}$ are computed. The feature vectors are supplied by a predetermined set of weights, $\{w_1, w_2, \ldots, w_{b^2}\}$.

The different dimensions of the feature vector are assigned to k normalization groups, G, so that each normalization group, G_i, contains a subset of dimension indices, I: $G_i = \{I_{i_1}, I_{i_2}, \ldots\}$, where $1 \leq I_j \leq N$. Each dimension index I_j may belong to at most one normalization group G_i.

Each dimension of the feature vectors has a mean, m_j that is either set to 0 for dimensions corresponding to operators whose responses are expected to be symmetric around 0, such as the Laplacian and normalized gaussian derivatives, or computed for operators, such as the intensity, that has no natural mid-point.

The variance for each dimension is computed:

$$v_j = \sum_{i=1}^{b^2} \frac{(x_{ij} - m_j)^2}{b^2} \tag{1}$$

The variance for each normalization group is the mean variance of its corresponding dimensions:

$$V_i = \sum_{j \in G_i} \frac{v_j}{|G_i|} \tag{2}$$

Each feature vector x is quantized into a vector y:

$$y_{j \in G_i} = \begin{cases} 0 & x_j - m_j < -3\sqrt{V_i} \\ q_j - 1 & 3\sqrt{V_i} \leq x_j - m_j \\ \left\lfloor \frac{q_j (x_j - m_j + 3\sqrt{V_i})}{6\sqrt{V_i}} \right\rfloor & -3\sqrt{V_i} \leq x_j - m_j < 3\sqrt{V_i} \end{cases} \tag{3}$$

In order to avoid amplifying noise in low contrast areas, a threshold is set from the assumed noise level of the images. All areas with a variance below the noise threshold are assumed to be uniform and are therefore normalized by a zero variance, which will result in a local histogram containing only one non-zero bin. The effects of contrast normalization and different levels of the noise threshold are studied in section 3.1.

2.4 Visual Vocabulary

A visual vocabulary is formed by an unsupervised clustering of histograms from random regions of a set of training images. The number of clusters or words, K, is a predefined parameter. The clustering algorithm used is K-means with a limited number of iterations. It is stopped at the first local minimum (which usually happened after about 20–30 iterations for the experiments in this paper). The distance metric used for the clustering is the Bhattacharyya distance, defined as:

$$d(h,t) = \sqrt{1 - \sum_i \sqrt{h(i) \cdot t(i)}} \tag{4}$$

The Bhattacharyya distance is chosen because it has been shown to perform slightly better than the χ^2 measure and because it represents a true distance metric satisfying the triangular inequality which helps when dealing with relative distances to different clusters in the histogram space. See (Linde and Bretzner, 2008) for an experimental comparison of the two distance measures.

In order to study the effect of the number of clusters K, experiments have been performed on the ETH-80 data set (Leibe and Schiele, 2003) in a leave-one-out categorization setting. The number of quantization levels, q, for the different descriptors have been experimentally chosen (between 5–15). See (Linde and Bretzner, 2008) for more information. Although the performance increases with the number of clusters up to at least 1000 clusters, we chose $K = 200$ as a trade-off since we want to keep the vocabulary size limited for several reasons. One reason is efficiency, the computation time increases linearly with the number of clusters. Furthermore, we believe that in a scalable system, the low-level vocabulary should be kept reasonably fixed and limited while more discriminative power should come from higher-level descriptors

built from (combinations of) low-level words. In this work we want to study the representational power of a limited low-level vocabulary, using the proposed descriptors, for categorization tasks in limited domains.

2.5 Categorization

The local histograms from each sub-region of the image is categorized as the closest matching visual word as determined by the Bhattacharyya distance. This results in an image map where each region is assigned a number corresponding to the closest matching visual word. A colored image showing region assignments is shown in image D in figure 1.

The final step of figure 1 shows a probability map computed from the prevalence of each visual word within the two classes foreground and background from 10 training images.

In this work, we focus on showing that the local representation is feature rich and preserves much of the information of the original image. In order to do this, we create a bag-of-words representation from the categorization map (image D in figure 1) and compare the resulting image descriptor with global receptive field histograms of (Linde and Lindeberg, 2004) computed using the same basic image operators. Image classification is done using a SVM classifier with the χ^2-kernel:

$$K(h_1,h_2) = e^{-\gamma\chi^2(h_1,h_2)} \tag{5}$$

with the parameter choice of $\gamma = 1.0$ from (Linde and Lindeberg, 2004). The actual implementation of the SVM was done on a modified variant of the libSVM software (Chang and Lin, 2001).

The bag-of-words descriptor is represented as a histogram with K bins. The resulting image descriptor is remarkably compact. Table 3 compares the average memory footprint of the image descriptors for the local approach used here compared to the global receptive field histograms in (Linde and Lindeberg, 2004).

3 EXPERIMENTAL RESULTS

3.1 Local Contrast Normalization

Variance normalization is performed to attain robustness against local contrast changes. To avoid amplifying noise in uniform image areas, a threshold is introduced. The threshold value will be dependent only on the actual noise level of the image (sensor noise, quantization noise and noise from compression) at a certain scale. The result of introducing a noise threshold are visualized in figure 2. The figures show region classifications, where each of the K code words is assigned a random color. We show how the noise threshold results in larger areas of background being treated as uniform and belonging to the same word.

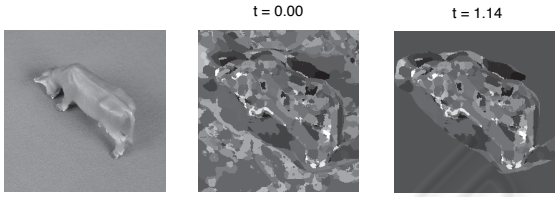

Figure 2: Image region classification from the ETH-80 data set for different values of the noise threshold (t).

Table 1 shows classification performance for the different descriptors on the ETH-80 data set with and without a noise threshold t, see section 3.2 for a description of the setup. A noise threshold t of 1.0 improves the performance slightly compared to no noise thresholding.

Table 1: Classification results on ETH-80 for different noise threshold values.

Descriptor	t = 0.0	t = 1.0	t = 2.0
3-D ($\nabla^2 L$)	10.4 ± 0.2 %	9.1 ± 0.4 %	11.2 ± 0.3 %
5-D ($\nabla^2 L, \nabla^2 C$)	9.2 ± 0.7 %	8.6 ± 0.1 %	10.7 ± 0.2 %
6-D ($\|\nabla L\|, \nabla^2 L$)	10.0 ± 1.3 %	9.5 ± 0.2 %	10.3 ± 0.7 %

A local, as opposed to a global, contrast normalization has advantages in being able to handle images with areas of different contrast. Such areas in images can appear from lighting, such as shadows and varying light intensities over the image, as well as due to atmospheric effects. Figure 3 shows the results of an experiment where testing images from Caltech-4, (Fergus et al., 2003), were subjected to an artificial contrast scaling gradient, see section 3.2 for a description of the setup. The linear artificial contrast scaling gradient was applied vertically over each test image, such that the first row of the image was multiplied by a scaling factor of 1.0, then gradually decreasing so the last (bottom) row was multiplied with 0.75, 0.50 and 0.25 respectively for the three scaling factors (25, 50, 75). Figure 4 shows examples of such artificially altered images. Local contrast normalization clearly reduces the negative effects of the shown contrast variations on the classification results.

Table 2 shows the results for the unaltered data set and different number of clusters. As can be expected, the normalized (e.g. contrast-invariant) descriptors clearly perform better for limited vocabularies. This illustrates one advantage of descriptor invariance properties in a scalable system.

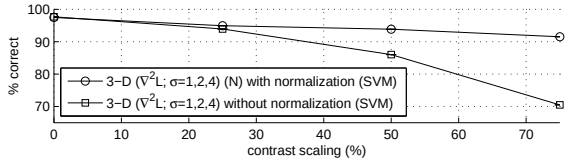

Figure 3: Classification performance on Caltech-4 with different levels of an artificial contrast gradient applied to the testing images.

Figure 4: Example images showing the artificial contrast gradient used for the results shown in figure 3. Images shown are the original (0) and images containing a linear contrast reduction from 0 on the first row down to 50 % and 75 % respectively on the bottom row.

Table 2: Classification error rates from the Caltech-4 data set, comparing normalized and non-normalized descriptors for the 5-D $(\nabla^2 L;\sigma=1,2,4),(\nabla^2 C;\sigma=2)$ descriptor.

Descriptor	K = 20	K = 200
5-D $(\nabla^2 L;\sigma=1,2,4),(\nabla^2 C;\sigma=2)$ normalized	6.99 %	2.12 %
5-D $(\nabla^2 L;\sigma=1,2,4),(\nabla^2 C;\sigma=2)$	8.75 %	2.54 %

3.2 Categorization Results

Earlier work has studied the performance of global histograms of the described point descriptors on classification tasks. In order to evaluate the potential of the proposed local histogram based descriptors, we compare their performance to the results of the global histograms. The proposed approach yields a much more compact image representation, see table 3, as the image is represented by a one-dimensional histogram of the same size as the number of visual words. A motivation for the comparison experiments is to study whether the heavily reduced representation results in much less discriminative power, which would significantly reduce the classification performance, or if the discriminative power is preserved. The two methods are:

- The proposed method presented in section 2. We refer to this as the local (histogram based) descriptor approach.
- Global high-dimensional receptive field histograms computed over the full image (Linde and Lindeberg, 2004), referred to as the global descriptor approach.

Table 3: The average representation size in bytes for the image descriptors used in table 4.

Descriptor	Global	Local
3-D $(\nabla^2 L;\sigma=1,2,4)$	8874 bytes	191 bytes
5-D $(\nabla^2 L;\sigma=1,2,4),(\nabla^2 C;\sigma=2)$	22485 bytes	200 bytes
6-D $(\lvert\nabla L\rvert,\nabla^2 L;\sigma=1,2,4)$	7795 bytes	200 bytes
3-D $(L,C;\sigma=1)$	565 bytes	80 bytes

Table 4: Comparison between global and local image descriptors for classification on the ETH-80 data set. For all cases, the local descriptor performs better.

Descriptor	Global	Local
3-D $(\nabla^2 L;\sigma=1,2,4)$	14.2 %	9.1 %
5-D $(\nabla^2 L;\sigma=1,2,4),(\nabla^2 C;\sigma=2)$	11.5 %	8.6 %
6-D $(\lvert\nabla L\rvert,\nabla^2 L;\sigma=1,2,4)$	13.3 %	9.5 %
3-D $(L,C;\sigma=1)$	17.9 %	17.7 %

Table 5: Classification results on the ETH-80 data set for local descriptors trained on the Caltech-4 data set.

Descriptor	Error rate
3-D $(\nabla^2 L;\sigma=1,2,4)$ (N)	14.33 %
5-D $(\nabla^2 L;\sigma=1,2,4),(\nabla^2 C;\sigma=2)$ (N)	14.42 %
6-D $(\lvert\nabla L\rvert,\nabla^2 L;\sigma=1,2,4)$ (N)	14.42 %
3-D $(L,C;\sigma=1)$	19.91 %

ETH-80. Table 4 shows a comparison in categorization performance between global and local descriptors on the ETH-80 data set. The table shows error rates for an leave-one-out classification problem using a support vector machine classifier. The performance of the local descriptor approach shows better classification results for the tested image operators. A closer look at the results for e.g. 3-D $(\nabla^2 L;\sigma=1,2,4)$ shows that the local histogram based descriptors give a significant reduction of the errors from confusing the categories cows, horses and dogs.

Our experiments on the ETH-80 data set give 8.6% error using color information and 9.1 resp. 9.5% without color. The best results from classification on ETH-80 known to the authors have been presented in (Nilsback and Caputo, 2004), reaching 2.89% error using a large number of different visual cues and a sophisticated decision tree scheme. Without color cues, but with direction sensitive descriptors, they reported 6.07%. (Leibe and Schiele, 2003) reported 6.98% at best using a multi-cue approach while 10.03% was reported without the contour cue but with color and gradient direction. For a rotation invariant descriptor without color they reported 17.77%.

As a somewhat crude test of the generality and scalability of the suggested approach, the ETH-80 experiments were also performed using a visual word vocabulary trained on the Caltech-4 data set. The results are shown in Table 5. Although the error rates are clearly higher than when training and testing are

performed on the same data set, the results indicate that the underlying descriptors allow the vocabulary to be trained on a different image data set as long as it includes sufficient image variations.

The local histogram approach makes it possible to approximately trace how the parts of each test image have contributed to the classification result. In figure 5, the grey scale in the images corresponds to the relative frequency of the corresponding visual word in the object class. The color and texture sensitive descriptor in the third column give higher contribution scores to areas with, for the object class, discriminant color, while the color-blind texture-sensitive descriptor in the second column give higher scores to more texture-rich areas like e.g specularities.

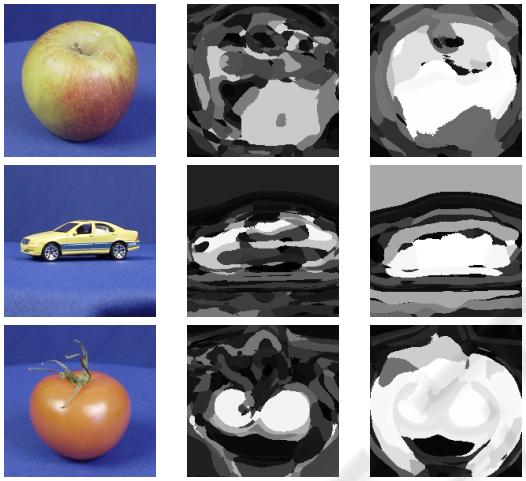

Figure 5: Some examples of back projection images from the ETH-80 data set. The second column corresponds to the 3-D $(\nabla^2 L; \sigma = 1,2,4)$ descriptor and the third column to the 5-D $(\nabla^2 L; \sigma = 1,2,4), (\nabla^2 C; \sigma = 2)$ descriptor.

Caltech-4. The Caltech-4 data set contains 800 images of objects for each one of the three categories "motor-bikes", "airplanes" and "car rears" as well as 435 images of "faces". Our training set was 400 images for each one of the categories motor-bikes, airplanes and car rears,and 218 images of faces. The test set consisted of all the other images in the data set. The results are shown in table 6, we find that the local histogram approach performs better for two of the operators and worse for the other two. The error rate for the local histogram approach is between 2.1 and 2.8 %. For a similar experimental setup using global gradient sensitive descriptors, (Nilsback and Caputo, 2004) reports an error rate of 3.1 %. In a multi-cue setup, with a sophisticated classification scheme combining three different non-invariant descriptors, Nilsback et.al. reports an error rate of 0.50 %.

Table 6: Classification error rates from the Caltech-4 data set, using local and global histograms.

Descriptor	Global	Local
3-D $(\nabla^2 L; \sigma = 1,2,4)$	2.6 %	2.5 %
5-D $(\nabla^2 L; \sigma = 1,2,4), (\nabla^2 C; \sigma = 2)$	1.2 %	2.1 %
3-D $(L, C; \sigma = 1)$	5.4 %	2.7 %
6-D $(\lvert\nabla L\rvert, \nabla^2 L; \sigma = 1,2,4)$	1.6 %	2.8 %

COIL-100. The COIL-100 (Nene et al., 1996) is not an image category data set, but experiments were done to examine the nature of the decrease in object instance recognition performance as the training views of the object get sparse. The recognition performance of the proposed approach was tested for varying constant angles between the training views, and testing was done using all other images. The graph in figure 6 shows that the recognition results degrade in a rather graceful manner with increasing distance between the training views.

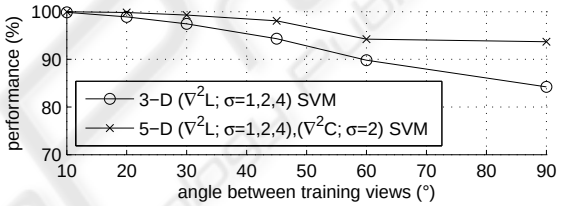

Figure 6: The performance on COIL-100 for the 3-D and 5-D local descriptors.

The recognition results for the 5-D descriptor are 99.7, 98.1 and 93.7% for 20, 45 and 90 degrees. Compared to earlier reported results, the global descriptor approach shows, not surprisingly, better results; for the 5-D descriptors the results are 100, 99.7 and 97.8%. As the generalization properties of the method are less crucial in this experiment, the global descriptor performs better. (Obdržálek and Matas, 2002) compares the performance of a number of different methods and obtains the by far best results for a sparse region-to-region matching using local affine frames; 99.9, 99.4 and 94.7% respectively. Although that method is highly suitable for instance matching under the image transformations present in the data set, our results using the local histogram based descriptors are of similar quality.

4 SUMMARY AND DISCUSSION

We argue that basic, low-level local image descriptors for recognition and classification tasks should include separate parts for invariant and non-invariant image measures. This capability should hold for a number of common image transformations, for example

rotational invariance/direction, scale invariance/scale, contrast invariance/contrast. One motivation is that a learning process, applied to the invariant descriptor parts, in general can be made much more efficient as fewer training examples have to be presented. Such a learning process would result in a reasonably large but limited vocabulary from which higher level descriptors can be formed. The non-invariant measures from the low-level stage could then be used in higher level descriptors, such as hyper-features (Agarwal and Triggs, 2006), to capture semi-local properties. The work presented here should be seen as a first step towards such a framework, by introducing descriptors showing a subset of the desired properties.

We have studied local histogram based image descriptors for representation of image structures. Applying them on a grid, we have tested these descriptors on classification tasks using well-known data sets for object classification in a bag-of-words fashion. The best of the proposed descriptors are based on the responses of Laplace operators applied at different scales, which means that the descriptors capture the distribution of the texture width and relative texture strength in the underlying subregion.

The classification performance has been compared to descriptors presented in earlier work, based on global histograms of the same basic operators. For the classification tasks, the local histogram based descriptors show similar or, in some cases, even better performance than the global histograms while achieving a significantly more compact representation.

The descriptors are of low dimension and suitable for hierarchical representations. For the given data sets, the classification and recognition results are comparable to the best known results from descriptors of similar complexity. We have shown how local contrast normalization improves the performance and is important for limited vocabularies. When contrast variations are applied to the data, there is a substantially increased classification/recognition performance from the proposed contrast normalization step.

We plan to further enhance these first-level descriptors by introducing a direction dependent part, together with a local scale measure and scale selection mechanism for full scale invariance. We will then explore how the scale, contrast and direction information can be incorporated in higher level descriptors.

REFERENCES

Agarwal, A. and Triggs, B. (2006). Hyperfeatures: Multilevel local coding for visual recognition. In *Proc. ECCV*, pages I: 30–43.

Bosch, A., Zisserman, A., and Munoz, X. (2007). Image classification using random forests and ferns. In *Proc. ICCV, Rio de Janeiro, Brazil*.

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

Csurka, G., Bray, C., Dance, C., and Fan, L. (2004). Visual categorization with bags of keypoints. In *Proc. ECCV International Workshop on Statistical Learning in Computer Vision*.

Dorkó, G. and Schmid, C. (2005). Object class recognition using discriminative local features. Rapport de recherche RR-5497, INRIA - Rhone-Alpes.

Fei-Fei, L. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proc. IEEE Conf. CVPR*, pages II: 524–531.

Fergus, R., Perona, P., and Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE Conf. CVPR*, pages II:264–271, Madison, Wisconsin.

Jurie, F. and Triggs, B. (2005). Creating efficient codebooks for visual recognition. In *Proc. ICCV*.

Koenderink, J. J. and Doorn, A. J. V. (1999). The structure of locally orderless images. *Int. J. Comput. Vision*, 31(2-3):159–168.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE CVPR*, pages 2169–2178, Washington, DC, USA. IEEE Computer Society.

Leibe, B. and Schiele, B. (2003). Interleaved object categorization and segmentation. In *Proc. British Machine Vision Conference*, Norwich, GB.

Linde, O. and Bretzner, L. (2008). Local histogram based descriptors for recognition. Technical report, CVAP/CSC/KTH.

Linde, O. and Lindeberg, T. (2004). Object recognition using composed receptive field histograms of higher dimensionality. In *ICPR*, Cambridge, U.K.

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

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. In *IJCV*, vol. 20, pp. 91–110.

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

Nilsback, M. and Caputo, B. (2004). Cue integration through discriminative accumulation. In *Proc. IEEE Conf. CVPR*, pages II:578–585.

Obdržálek, v. and Matas, J. (2002). Object recognition using local affine frames on distinguished regions. In *British Machine Vision Conference*, pages 113–122.

Puzicha, J., Hofmann, T., and Buhmann, J. (1999). Histogram clustering for unsupervised segmentation and image retrieval. *Pattern Recognition Letters*, 20:899–909(11).

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

Schmid, C. (2004). Weakly supervised learning of visual models and its application to content-based retrieval. *IJCV*, 56(1-2):7–16.

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