A Comparison of Features in Parts-Based Object Recognition Hierarchies

Stephan Hasler, Heiko Wersing, and Edgar Körner

Honda Research Institute Europe GmbH
D-63073 Offenbach/Germany
stephan.hasler@honda-ri.de

Abstract. Parts-based recognition has been suggested for generalizing from few training views in categorization scenarios. In this paper we present the results of a comparative investigation of different feature types with regard to their suitability for category discrimination. So patches of gray-scale images were compared with SIFT descriptors and patches from the high-level output of a feedforward hierarchy related to the ventral visual pathway. We discuss the conceptual differences, resulting performance and consequences for hierarchical models of visual recognition.

1 Introduction

The human brain employs different kinds of interrelated representations and processes to recognize objects, depending on the familiarity of the object and the required level of recognition, which is defined by the current task. There is evidence that for identifying highly familiar objects, like faces, holistic templates are used that emphasize the spatial layout of the object's parts but neglect details of the parts themselves. This holistic prototypical representation requires a lot of experience and coding capacity and therefore can not be used for all the objects in every day's life. A more compact representation can be obtained when handling objects as combinations of shared parts. There is various biological motivation for such a representation. The experiments of Tanaka [1] revealed that there are high-level areas in primates ventral visual pathway that predict the presence of a large set of features with intermediate complexity, generalizing over small variations and being invariant to retinotopical position and scale. The combinatorial use of those features was shown by Tsunoda [2]. He observed that complex objects simultaneously activate different spots in those areas and that this activation is caused by the constituent parts. A parts-based representation is especially efficient for storing and categorizing novel objects, because the largest variance in unseen views of an object can be expected in the position and arrangement of parts, while each part of an object will be visible under a large variety of 3D object transformations.

In computer vision literature there is a similar distinction into holistic and parts-based approaches, depending on how feature responses are aggregated over

J. Marques de Sá et al. (Eds.): ICANN 2007, Part II, LNCS 4669, pp. 210–219, 2007.

the image. Parts can be local features of any kind. The response of a part detector at different positions in an image means that the part might be present several times but not that the probability is higher that the part is present at all. So each peak in the multimodal response map is handled as a possible instance of the part. In contrary to this, holistic approaches contain a layer that simply accumulates the real-valued response of single features of the previous layer over the whole image. This is only comparable to the biological definition if the configurational information is kept.

Approaches with strong biological motivation are presented in [3,4]. Here hierarchies of feature layers are used, like in the ventral visual pathway, where they combine specificity and invariance of features. So there are cells that are either sensitive to a specific pattern of activation in lower layers, in this way increasing the feature's complexity, or that pool the responses of similar features, so generalizing over small variations. The output layer of the feedforward hierarchy proposed in [3] contains several topographically organized feature maps which are used directly by the final classifier. Following the above definition this is a holistic approach. The similar hierarchy of [4] employs in the highest feature layer a spatial max-pooling over each feature map in the previous layer, which makes it a parts-based approach. Multimodal response characteristics and the position of the parts are neglected.

Most other approaches work more directly on the images. Very typical holistic approaches apply histograms, so e.g. in [5] the responses to local features are simply summed and in [6] it is counted how often a response lies in a certain range. In other holistic methods the receptive fields of the features cover the whole image. So e.g. in [7] features obtained by principal component analysis (PCA) on gray-scale images were used to classify faces. These features, so called eigenfaces, show a very global activation and do not reflect parts of a face. In contrary to PCA other methods produce so called parts-based features like the nonnegative matrix factorization (NMF) proposed in [8] or a similar scheme proposed in [9] yielding more class-specific features. Although during training the receptive field of each feature covers the whole image, it learns to reconstruct a certain localized region that contains the same part in many training views (e.g. parts of normalized frontal views of faces). But usually those features are used in a holistic manner, meaning that they are extracted at a single position in the test image and in this way are only sensitive to the rigid constellation of parts that was present during training. This limits the possibilities to generalize over geometric transformations, which is especially a drawback when using few training examples in an unnormalized setting. Also the holistic approaches perform bad in the presence of clutter and occlusion and often require extensive preprocessing as localization and segmentation.

Other parts-based recognition approaches also use the maximum activation of each feature, like the highest layer in [4]. In [10] the features are fragments of gray-scale images. The response of a feature is binary and obtained by thresholding the maximum activation in the image. The approach selects features based on the maximization of mutual information for a single class. This yields fragments of intermediate complexity. An image is classified by comparing binary activation vectors to stored representatives in a nearest neighbor fashion. Other approaches make use of the position and treat each peak in the response map as possible part instance. In the scale invariant feature transform (SIFT) approach in [11] gradient-histograms are extracted for small patches around interesting points (see Fig. 1c). Each such patch descriptor is compared against a large repertoire of stored descriptors, where the best match votes for the presence of an object at a certain position, scale and rotation. The votes are combined using a Generalized Hough Transform and the maximally activated hypothesis is chosen. A similar scheme is proposed in [12]. Here image patches are used as features and the algorithm is capable to produce a segmentation mask for the object hypothesis that can be used for a further refinement process. In the bags of keypoints approaches, e.g. [13], it is counted how often parts are detected in an image. In contrast to holistic histogram-based approaches the presence of a part is the result of a strong local competition of parts. Therefore it is more a counting of symbol-type information than a summation over real-valued signal-type responses. Parts-based recognition can be used to localize and recognize objects at the same time and works well in the presence of clutter and occlusion.

In Sect. 2 we first comment on the task we want to solve and the nature of the features required for this. Then we describe the investigated feature types and our feature selection strategy. We give results for a categorization problem in Sect. 3 and present our conclusions in Sect. 4.

2 Analytic Features

To generalize from few training examples, parts-based recognition follows the notion that similar combinations of parts are specific for a certain category over a wide range of variations. In this work we investigate how suitable different feature types are for this purpose and which effort is needed in terms of the number of used features. As has been argued in [10], it is beneficial that a single part can be detected in many views of one category, while being absent in other categories. So we need a reasonable feature selection strategy that evaluates which and how many views of a certain category a feature can separate from other categories and, based on those results, choose the subset of features that in combination can describe the whole scenario best. For simple categories a single feature can separate many views and therefore only few features are necessary to represent the whole category. For categories with more variation more features have to be selected to cover the whole appearance. This dynamic distribution of resources is necessary to make best use of the limited number of features.

How well certain local descriptors can be re-detected under different image transformations, as scale, rotation and viewpoint changes, was investigated in [14]. Although this is a desired quality, it does not necessarily state something on the usefulness in object recognition tasks. To underline that the desired features should be meaningful, i.e. offer a compromise between specificity and generality at low costs, and to avoid confusion with approaches that learn parts-based features, we will use the term analytic features.

We decided to compare patches of gray-scale images, for their simplicity, SIFT descriptors, for their known invariance, and patches of the output of the feedforward hierarchy in [3], because of the biological background.

A SIFT descriptor as proposed in [11] describes a gray-scale patch of 16x16 pixels using a grid of 4x4 gradient-histograms (see Fig. 1c). Each histogram in the grid is made up of eight orientation bins. The magnitude of the gradient at a certain pixel is distributed in a bilinear fashion over the neighboring histograms (in general four), where the orientation of the gradient determines the bin. The gradient magnitudes are scaled with a Gaussian that is centered on the patch, in this way reducing the influence of border pixels. Prior to the calculation of the histogram grid a single histogram with a higher number of orientation bins is computed for the whole patch. The maximum activated bin in this histogram is used to normalize the rotation of the patch in advance. Finally the energy of the whole descriptor is normalized to obtain invariance to illumination. In contrast to [11], we do not extract SIFT descriptors at a small number of interesting keypoints, but for all locations where at least a minimum of structure is present. In this way only uniform, dark background is neglected and on the category scenario in Fig. 3 on average one third of all descriptors is kept. We reduce the number of descriptors for each image by applying a k-means algorithm with 200 components. A similar cluster step was also done in [15] to improve the generalization performance of the otherwise very specific SIFT descriptors.

For the gray-scale patches we decided to use the same patch size as for the SIFT approach and the influence of the pixels is also weighted with a Gaussian that is centered on the patch.

The feedforward hierarchy proposed in [3] is shown in Fig. 1a. The S1-layer computes the magnitudes of the response to four differently oriented gabor filters. This activation is pooled to a lower resolution in the C1-layer performing a local OR-operation. The 50 features used in S2 are trained as to efficiently reconstruct a large set of random 4x4x4 C1-patches from natural images and are therefore sensitive to local patterns in C1. Layer C2 performs a further pooling operation and is the output of the hierarchy. Columns of 2x2 pixels are cut from the C2-layer as shown in Fig. 1b and used as feature candidates. Because of the two pooling layers, which offer a small degree of invariance to translation, a column of 2x2 pixels in C2 corresponds roughly to a patch of 16x16 pixels in the gray-scale image.

We will refer to the parts-based approaches as GRAY-P, SIFT-P, and C2-P. For SIFT-P each image i is described by the $J = 4 \times 4 \times 8 = 128$ dimensional representatives of the 200 k-means clusters $\mathbf{p}_{in}$, $n = 1 \dots 200$. For GRAY-P the $\mathbf{p}_{in}$ are the patches of image i at all distinct positions n ($J = 16 \times 16 = 256$). Similar to this for C2 each $\mathbf{p}_{in}$ is a column through the feature maps of image i at a distinct position n as shown in Fig. 1b ($J = 2 \times 2 \times 50 = 200$). The $\mathbf{p}_{in}$ show a large variety. Therefore we will use all $\mathbf{p}_{in}$ directly as feature candidates $\mathbf{w}_m$, where m is an index over all combinations of i and n, and select a subset

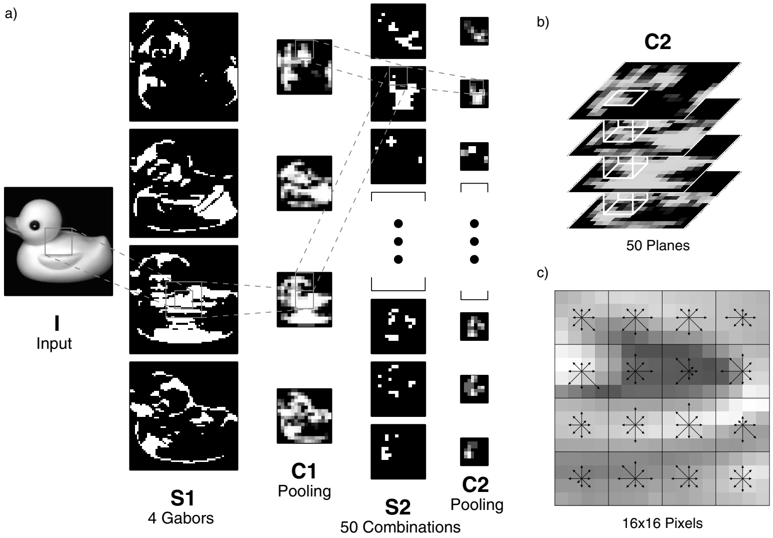

Fig. 1. a) Feedforward hierarchy in [3]. b) Columns of C2-layer are used as local features. c) SIFT descriptor [11] is grid of gradient histograms each with 8 orientations.

of those candidates with a strategy that is described later. The response r_{mi} of feature $\mathbf{w}_m$ on the image i is given by:

$$r_{mi} = \max_{n} \left(G(\mathbf{w}_m, \mathbf{p}_{in}) \right) . \tag{1}$$

For GRAY-P $G(\mathbf{w}_m, \mathbf{p}_{in}) = \frac{\sum_{j=1}^{J} h^j (w_m^j - \overline{w}_m)(p_{in}^j - \overline{p}_{in})}{\sqrt{\sum_j h^j (w_m^j - \overline{w}_m)^2 \sum_j h^j (p_{in}^j - \overline{p}_{in})^2}}$ is used which is the normalized cross-correlation, where $\overline{w}_m$ and $\overline{p}_{in}$ are the means of vector $\mathbf{w}_m$ and $\mathbf{p}_{in}$ respectively, and h^j is a weighting which decreases the influence of border pixels with a Gaussian. For C2-P the negative Euclidean distance $G(\mathbf{w}_m, \mathbf{p}_{in}) = -\sqrt{\sum_{j=1}^{J} (w_m^j - p_{in}^j)^2}$ shows better performance because of the sparseness in this layer. The similarity between SIFT descriptors is given by their dot product $G(\mathbf{w}_m, \mathbf{p}_{in}) = \sum_{j=1}^{J} w_m^j p_{in}^j$. The maximum activation per image is chosen as response and spatial information is neglected.

Reflecting the remarks on feature selection given above, we decided to use the following strategy: First we determine which views of a certain category each individual candidate feature $\mathbf{w}_m$ can separate. Therefore we compute the response r_{mi} for every training image with (1). Then the minimal threshold t_m is chosen that guarantees that all images with r_{mi} above or equal to t_m belong the same category (see Fig. 2):

$$t_m = \min \left\{ t | \forall_{\substack{i | r_{mi} \geq t \\ j | r_{mj} \geq t}} l_i = l_j \right\} . \tag{2}$$

Here l_i denotes the category label of image i. The images separated by the threshold is assigned a constant score $s_{mi} = k$ with respect to the feature $\mathbf{w}_m$.

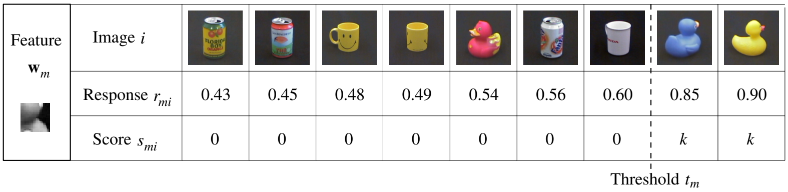

Fig. 2. Feature selection scheme. For visualization the images are sorted on their response r_{mi}. The threshold t_m separates views of a single category (here ducks) from all other images. To these views a score $s_{mi} = k$ is assigned.

When the scores s_{mi} are determined for the set of candidate features M an iterative process selects a given number of features by determining in each step the best candidate feature m with:

$$m = \arg\max_{m \in M} \left(\sum_i f \left(s_{mi} + \sum_{q \in Q} s_{qi} \right) \right) \tag{3}$$

and putting it from M into the set of already selected features Q. First $(Q = \emptyset)$ the feature is selected that is detected in the most views of a certain category. Then successively the feature which causes the highest additional score is selected. The function $f(z)$ controls how effective a new feature can score for a single image. When using a Heaviside function only a single feature can score for an image. Here we use a Fermi function $f(z) = \frac{1}{1+e^{-z}}$ and set $k = 3$. In this way the feature gets only a high score for images that were not separated yet, and a much lower score for images in which features have already been detected.

3 Results

We tested the performance of the different feature types on the categorization scenario shown in Fig. 3. The gray-scale images have a resolution of 128x128 pixels and show centered objects on dark background. The objects belong to ten categories, where each category contains nine objects. Five objects per category are used for training and the remaining four for testing. Each object is represented by 30 views taken during a rotation around the vertical axis.

For each approach we ranked the candidate features from the complete set of training images with the introduced selection framework. The first 75 selected features for each approach are shown in Fig. 4a. The gray-scale patches contain a lot of similar parts under different orientations. For C2 less complex patches are selected that sometimes have only activation at the border or even seem to stem from the background. The SIFT patches show the largest variety.

For the different tests we then varied the number of used features and the number of training views that were used by a single layer perceptron (SLP), as the final classifier. So first for each training and test image a vector was

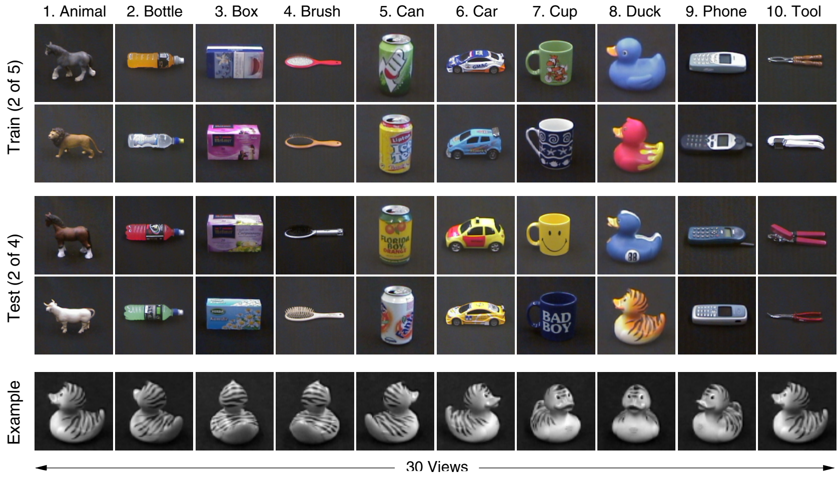

Fig. 3. Category scenario. Each category contains nine objects. Five are used for training and four for testing. Only two objects of both groups are shown here.

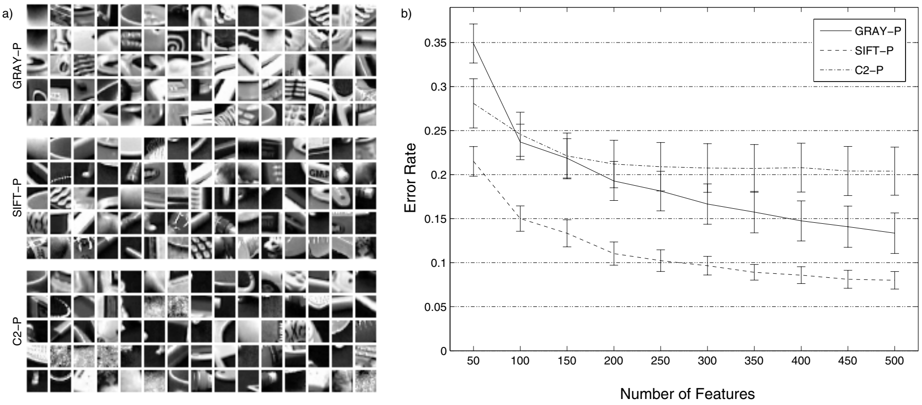

Fig. 4. a) First 75 top ranked features for parts-based approaches. For C2 the corresponding patch of the original gray-scale image and for SIFT the patch the descriptor of which is most similar to the selected k-means component is shown. b) Error rates depending on number of features for parts-based approaches.

calculated containing the responses of the selected features using (1). We let the SLP converge on the training vectors, and after this calculated the recognition performance on the complete set of test vectors. To increase both difficulty and objectivity we did not distribute the training examples equally over the single categories but repeated each test 50 times with random sets of training images and so obtained a mean performance together with a standard deviation.

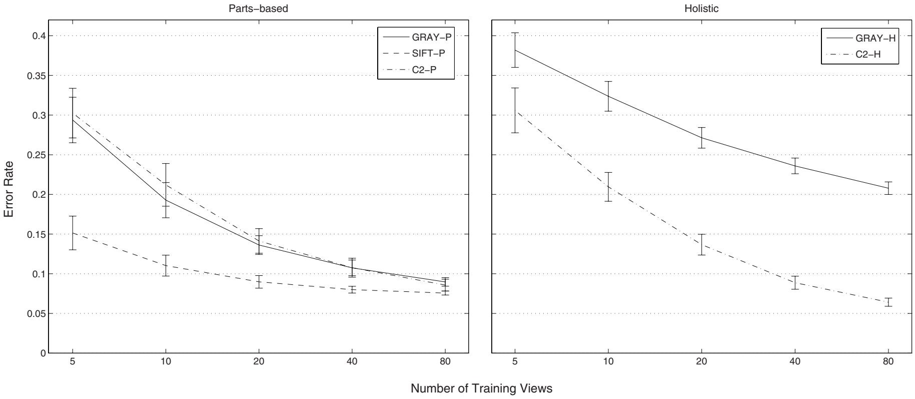

Fig. 5. Error rates depending on number of training views for different approaches

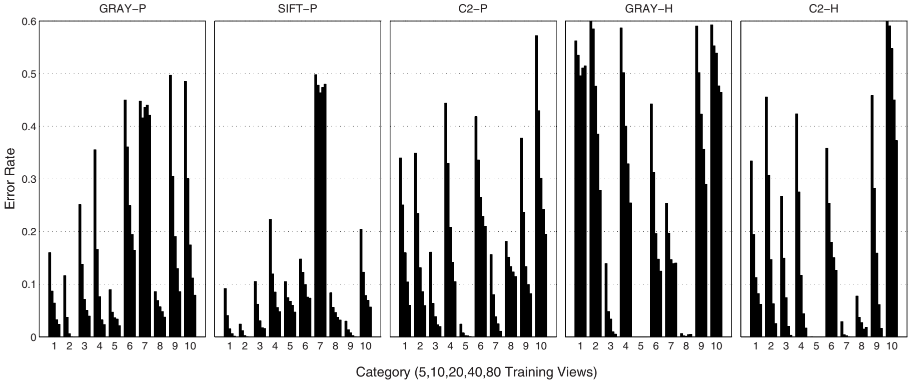

Fig. 6. Error rates of individual categories depending on number of training views

Fig. 4b shows how the classification performance depends on the number of selected features. For this test 10 random views were used per category for SLP training. SIFT-P outperforms C2-P and GRAY-P. For small numbers of features GRAY-P has the worst result, but shows the best improvement with increasing feature number, while the performance of C2-P saturates early. Maybe the variability that is gained via quantity helps to overcome the missing invariance of the very specific GRAY-P patches. C2-P patches make use of the invariance gained by the hierarchical processing from the beginning. But they maybe too general and only few qualitatively distinct features might exist.

Fig. 5 shows how the recognition performance depends on the number of views per category that were randomly chosen for the training of the SLPs. In this test we used 200 features for the parts-based approaches. On the right hand side of the figure we give also results of SLPs that were trained on the original gray-scale images (GRAY-H) and on the complete C2-activations (C2-H). For few training views SIFT-P is superior to the other approaches. C2-H is similar to C2-P and GRAY-P, and takes the lead when using a large number of views. GRAY-H shows the worst performance. More than the other approaches C2-H profits from an increase in the number of training views. This confirms the notion, that columns of C2-H, as used in C2-P, are invariant but too general. Although this is a drawback for C2-P, it helps C2-H together with position information to extrapolate well in the neighborhood of single views.

To provide reason for the shown differences Fig. 6 visualizes the same test with the mean error rates given for individual categories. SIFT-P especially works well for animals(1), bottles(2) and phones(9) but is outperformed by all other methods on cans(5) and cups(7), and by C2-H also on ducks(7). The performance of SIFT-P and GRAY-P on cups(7) is very poor and does not improve with more training views. The patches for SIFT-P and GRAY-P contain only few cup features but those are top-ranked and highly discriminative for the training images, but maybe too specific to generalize over the test images.

To conclude, the holistic approaches (C2-H, GRAY-H) are good for categories that do not vary much in shape during a rotation around the vertical axis, like cans(5) or cups (7). Also the results on ducks(8) are good because only the position of the head changes, while the body shape stays nearly unchanged. When the change of the global shape is more extreme during rotation SIFT-P performs better in comparison to the other approaches. This is especially true for categories where the rotation in depth looks like rotation in plane (bottle(2), brush(4), phone(9), tool(10)).

4 Conclusion

We evaluated the performance of different types of local feature when used in parts-based recognition. We showed that SIFT descriptors are good analytic features for most objects especially when the number of training views and the number of features is limited. The biological motivated feedforward hierarchy in [3] is powerful in holistic recognition with a sufficient number of training examples, but the patches from the output layer are too general and therefore show weak performance in parts-based recognition. This is interesting because also the calculation of a SIFT descriptor can be described as hierarchical processing: First features are used that extract the magnitudes for 8 different local gradient directions. Then a local winner takes all is applied over those features at each position. Each of the 16 histograms in the 4x4 grid integrates over each direction in a local neighborhood by summing the magnitudes (no non-linearity used as for pooling in [3,4]). Finally the SIFT descriptor stands for a more global activation pattern in the grid. Besides the normalization of rotation for SIFT, it would be interesting to investigate other reasons for the differences in performance in future work. This could be beneficial for both feature types.

The most related work in the direction of analytic features was done in [16], where Ullman introduced invariance over viewpoint in his fragments approach, or in the work of Dorko et al. in [15], where highly informative clusters of SIFT descriptors are used. Since both approaches have not been applied to scenarios with multiple categories, we hope that our comparative study provides further helpful inside into parts-based 3D object recognition.

References

1. Tanaka, K.: Inferotemporal Cortex And Object Vision. Annual Review of Neuroscience 19, 109–139 (1996)
2. Tsunoda, K., Yamane, Y., Nishizaki, M., Tanifuji, M.: Complex objects are represented in inferotemporal cortex by the combination of feature columns. Nature Neuroscience 4(8), 832–838 (2001)
3. Wersing, H., Körner, E.: Learning Optimized Features for Hierarchical Models of Invariant Object Recognition. Neural Computation 15(7), 1559–1588 (2003)
4. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust Object Recognition with Cortex-Like Mechanisms. IEEE Trans. Pattern Analysis and Machine Intelligence 29(3), 411–426 (2007)
5. Mel, B.W.: SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. Neural Computation 9(4), 777–804 (1997)
6. Schiele, B., Crowley, J.L.: Object Recognition Using Multidimensional Receptive Field Histograms. In: European Conference on Computer Vision, pp. 1039–1046. Cambridge, UK (1996)
7. Turk, M., Pentland, A.: Eigenfaces for Recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
8. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
9. Hasler, S., Wersing, H., Körner, E.: Class-specific Sparse Coding for Learning of Object Representations. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) ICANN 2005. LNCS, vol. 3696, pp. 475–480. Springer, Heidelberg (2005)
10. Ullman, S., Vidal-Naquet, M., Sali, E.: Visual features of intermediate complexity and their use in classification. Nature Neuroscience Vision Research 5(7), 682–687 (2002)
11. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
12. Leibe, B., Schiele, B.: Scale-Invariant Object Categorization Using a Scale-Adaptive Mean-Shift Search. In: Rasmussen, C.E., Bülthoff, H.H., Schölkopf, B., Giese, M.A. (eds.) Pattern Recognition. LNCS, vol. 3175, pp. 145–153. Springer, Heidelberg (2004)
13. Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, Springer, Heidelberg (2004)
14. Mikolajczyk, K., Schmid, C.: A Performance Evaluation of Local Descriptors. IEEE Trans. Pattern Analysis and Machine Intelligence 27(10), 1615–1630 (2005)
15. Dorko, G., Schmid, C.: Object Class Recognition Using Discriminative Local Features. In: INRIA (2005)
16. Ullman, S., Bart, E.: Recognition invariance obtained by extended and invariant features. Neural Networks 17(1), 833–848 (2004)