Ensembles of Gradient Based Descriptors with Derivative Filters for Visual Object Categorization

Enas A.A. Eqlouss, Azizi Abdullah, and Siti Norul Huda Sheikh Abdullah

Pattern Recognition Research Group
Center for Artificial Intelligence
Faculty of Information Science and Technology
University Kebangsaan Malaysia 43600 Bangi, Selangor
bello_pearla@yahoo.com,{azizi,mimi}@ftsm.ukm.my

Abstract. This paper describes several ensemble methods that combine multiple edge and orientation based histograms with support vector machine classifiers. The aim is to enhance learning speed and accuracy performance by using the chosen classical primitive filters on different edge and orientation descriptors. For efficiently describe images using these descriptors, the combination of a few basis features or edge filters are used. The stronger filter operator responds to edge-like structures, the more sensitive it to orientation. Thus, using more than one edge filter allows to capture more edge information to completely describe the structure of image content. One problem in combining these different descriptors is that the input vector becomes very large dimensionality, which can increase problems of overfitting and hinder generalization performance. The intuitively designed ensemble methods namely product, mean and majority are then used to combine support vector machines classifiers derived from the multiple orientations of edge operators. The results indicate that the ensemble methods outperform the single and naive classifiers.

Keywords: object categorization, ensemble rules, support vector machines, compass filters.

1 Introduction

An image is an ill-defined entity consisting of many image features such as lines, edges or textures. These features provide a more abstract and informative description of images than pixels for recognition. The images can be of a complex nature, however, it is not impossible to describe the generic meaning by using these low-level features. During the last decade a large number of novel algorithms have been researched for recognition and became one of the most interesting topics in computer vision [1][2][3]. The algorithms use descriptors describing an image and a machine learning algorithm for classification. However, it is common experience for computer vision researches to get low accuracy performance due to complex transformations with high inter-class variations from the query of digital images.

In the literature, it is often difficult to determine which image features are most discriminative to describe the information in an image. Good image features are crucial because they can give a compact representation and help in capturing meaningful patterns

J.-H. Kim et al. (Eds.): *Robot Intelligence Technol. & Appl. 2012*, AISC 208, pp. 33–43.
DOI: 10.1007/978-3-642-37374-9_4

in the image. However, one of the most widely used features in describing images is the construction of edge and orientation features. For this reason, gradient based descriptors such as SIFT[4], histograms of oriented gradients [5] and MPEG-7 edge histogram [6][7] have become popular and nowadays widely used in image recognition systems. In this paper, we show the advantages of using ensembles of a set of edge filters instead of pixel difference in the performance of the well known edge-based descriptors. One problem in gradient computation using pixel differences in these descriptors is sensitive to noise and other artifacts, which can increase problems of feature indexing. A common solution to the problem is to compute smoother approximations of the image derivative using filters such as Gabor filters[8][9][10] and Gaussian-weighted Principle Independent Component Analysis (GPICA) of an image [11]. In contrast to these previous works, here we use a set of classical filters namely Robinson's filters to approximate the whole structure of the image. One problem with a single filter for feature description is that the stronger a filter operator responds to edge-like structures, the more sensitive it to orientation and the operator only respond to edges in a narrow range of orientations. Thus, using more than one edge operator provides wider range of directions and magnitudes for feature description. After that, each gradient based histogram of the operator outputs and giving them as input to a learning classifier such as a support vector machine (SVM) [12] has been shown to lead to promising results.

In [13][10], the authors computed descriptors from feature maps at different scales and orientations using filters. However, when these methods are used to combine many gradient based descriptors in a single large input vector, this may lead to overfitting the data and worse generalization performance. Therefore, in [13], proposes a series of planes where successive convolutions and subsampling operations at different scales to construct feature maps for feature indexing. The system is based on neural network architecture that consists of six different layers, where the last two layers carry out the classification task using the features extracted in the previous layers namely n_1 and n_2 layers. Layer n_2 receives inputs from layer n_1 that contains a number of partially connected sigmoid neurons of the network. In contrast, [10] constructs a set of different feature maps from Gabor Wavelets features at different scales and orientations. After that, the Principle Component Analysis is applied to reduce dimensionality of feature maps and nearest neighbor classifier for classification. In this paper, we describe several ensemble methods that combine multiple descriptors from different feature maps. The aim is to enhance learning speed and accuracy performance by using the chosen classical primitive filters of different edge and orientation descriptors. It basically constructs a set of individual support vector machines classifiers from these descriptors. After that, the ensembles are used for learning and combine probability outputs from all classifiers.

Contributions of this Paper. (1) The ensembles are used for learning and combine multiple outputs of filter based classifiers to enhance learning speed and accuracy performance. (2) We compare the accuracy of the proposed method with single and naive approaches on 20 classes from Caltech-101 dataset.

The rest of the paper is organized as follows: Section 2 reviews the related researches. After that we describe our system for image recognition in section 3. The system basically uses gradient based descriptors compute feature vectors from feature maps that are used to construct support vector machine classifiers. Section 4 describes the ensemble methods and how we used the support vector machine as classifiers. Experimental results on 20 classes from Caltech-101 dataset are shown in Section 5. Section 6 concludes this paper.

2 Related Work

Edge and orientation are important elements for object recognition purposes than pixels information. These information are typically represented by combination of a few basis features. Recently, most studies are focusing on using multiple edge and orientation information to completely describe images for satisfactory recognition result. Using these information may help in recognizing different structures of images in wider range. The information can be extracted using a single filter such as Sobel or a set of different filters such as compass filters or Gabor filters. However, using multiple filters are more informative for feature indexing due to different orientations or magnitudes of image structure can be extracted.

2.1 Robinson Compass Filters

One advantage using edge operator such as Robinson filters is that a wider range of orientations and magnitudes information can be extracted. It contains eight different filters with orientation spaced at 45°. The main reason why we used this filter is that it is simple and not requiring expensive operation to convolve images. The Robinson filter contains eight major orientations namely vertical right, vertical left, bottom horizontal, top horizontal, bottom right diagonal, top left diagonal, bottom left diagonal and top right diagonal with its coefficients in the range of -2 and 2.

2.2 MPEG-7's Edge Histogram

Human eyes are very sensitive to the intensity changes. Thus, texture information is important to check homogeneity and non-homogeneity between images. We used the MPEG-7 edge histogram [14] to compute texture information. The edge histogram describes a non-homogeneous texture and captures a local spatial distribution of edges. Given an input image, the image is partitioned into 4 x 4 overlapping blocks or 16 sub-block. After that each sub-block is convolved with the following five orientation filters. As a result each block holds a total of five different orientations or 5-bin for description. The maximum of the most dominant edges is determined by comparing it with other edges' strength. Then the maximum of these results is compared with a threshold. Finally, the descriptor with 80-bin histogram for intensity component is constructed for the input image by excluding the no-edge information. We named them as EH_G to represent the edge histogram with intensity information.

2.3 Histograms of Threshold-Oriented Gradients (HTOG)

Shape is important to discriminate between objects. Local shape histograms are represented by edge orientations within an image subregion quantized into N bins. We model the shape by using intensity signals, and then we compute orientations by detecting the signal changes that are visible and significant in a certain angular range.

The most popular gradient base histogram that extract information about edges and shape is HOG[5] and SIFT[4]. The histogram basically describes an image by a set of local histograms. In contrast to this previous work, here we used histograms that count occurrences of thresholded gradients in a local part of the image. After that, the image is divided into 4 x 4 sub regions to obtain the spatial relationship between edge attributes. Subsequently, the gradients dx and dy are computed at each point in each region by using the filters in the x and y directions, respectively.

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & -1 \end{bmatrix}$$

To compute the magnitude and orientation of the gradient the following formulas are used:

$$m(x,y) = \sqrt{dy^2 + dx^2} \tag{1}$$

$$\Theta(x,y) = arctan(dy/dx) \tag{2}$$

where m is the magnitude, Θ is the orientation of the gradient, and dy and dx are gradients in vertical and horizontal directions, respectively.

In order to determine the occurrence histogram for different orientations, a threshold is used to choose the strongest edges. The edge is considered as a noise or weak response if $m(x,y)$ is below the threshold (in our experiments set to 10), and not counted in constructing the histogram. Otherwise, all Θ's which have a magnitude above the threshold are selected and then quantized into N bins. In our experiments, N = 8 gave the best results. Finally, the descriptor with 128 bins is constructed for the whole region (consisting of 4x4 blocks). Each bin in the histogram represents the number of occurrences of edges that have a certain orientation. We chose several angular ranges to recognize different structures of images and to enrich the semantic description of the edge information. We found two angular ranges i.e., 180° and 360° to be optimal in our dataset. An angular range of 180° maps angles between 180° and 360° to the range between 0 and 180 degrees. We named the two resulting descriptors HG_180_G and HG_360_G to represent the HTOG with intensity information.

2.4 Block-Based SIFT (Scale Invariant Feature Transform)

SIFT[4] describes an image by constructing histograms of gradient orientations around a set of interest points. Thus, we also applied the descriptor as one of our main descriptors. The original SIFT version uses an interest points detector to detect salient locations which have certain repeatable properties. In contrast with this approach, we believe that using fixed partitioning blocks gives a simpler method with the same or better performance on our dataset. Furthermore, using this approach the spatial relationships between the SIFT features can be represented more efficiently, i.e. we do not need clustering and less computational time for constructing the descriptor. Therefore, fixed regions without orientation alignment are constructed over the image and instead of 'salient points' we compute the center of each region.

To compute the descriptor, an input image (whole image) is smoothed with the same smoothing function and differentiated using the same dx and dy filters. After that, the center point of the region is determined by dividing its width and height by 2. The descriptor is then constructed by a circular region around the center point of the region. The circular region radius is determined by taking the $min(\frac{width}{2}, \frac{height}{2})$, where width and height are the sizes of the region. After that, the descriptor breaks apart a window around the center point into 4x4 sub-blocks and calculates a gradient orientation histogram, whereby each gradient is weighted by its magnitude to better reflect strong orientations. Each histogram has 8 bins and in total there are 128 bins per histogram for each region. Our use of SIFT differs from the HTOG in the following ways: it uses a circular region instead of a rectangular block and it does not use a threshold on the magnitude. In this way we compute complementary features with SIFT and HTOG. We also used SIFT descriptors with 180° and 360° angular ranges to enrich its visual information. We named them S_180_G and S_360_G to represent the SIFT descriptors with intensity information.

3 Ensemble of Gradient Histogram Based Filters

The main idea behind histogram gradient-based filters is to use the most probable edge or orientation frequency distributions to depict different feature map images. Using filters basically may help to recognize different maxima at edge structures of images efficiently and enrich the semantic description of visual information. The stronger a filer operator responds to edge-like structures, the more sensitive it to orientation. Thus, using more than one edge operator allows to capture more edge information to completely describe image content.

The system uses multiple gradient based descriptors to describe feature maps extracted from eight different Robinson filters. It consists of five layers, excepting the input plane that receives an image as a whole of any sizes, without requiring any local pre-processing of the input image such as brightness correction, contract adjustment, etc. The system starts by convolving pixels neighborhood of an image with eight 3 x 3 masks (a through h of Fig. 1), resulting eight different magnitude images or feature maps with maxima at edge locations. The different filters used in the stage is to provide wider range of directions and magnitudes for specific structural feature description. After that, gradient based descriptors are applied to compute feature vectors from the feature maps. These descriptors compute the first order image derivatives using these filters as convolution masks. Next, the computed feature vectors are used to construct support vector machine models for predictions. Once trained, for a given test object, SVM probability outputs of all models are then combine using ensemble learning methods for final classification. Fig. 1 shows the overall purposed system.

4 Classification Methods

4.1 SVM Classifier

We employ an SVM [15] to learn to classify the images. The one-vs-one approach is used to train and classify images in the Caltech-101 dataset. For the SVMs, we use

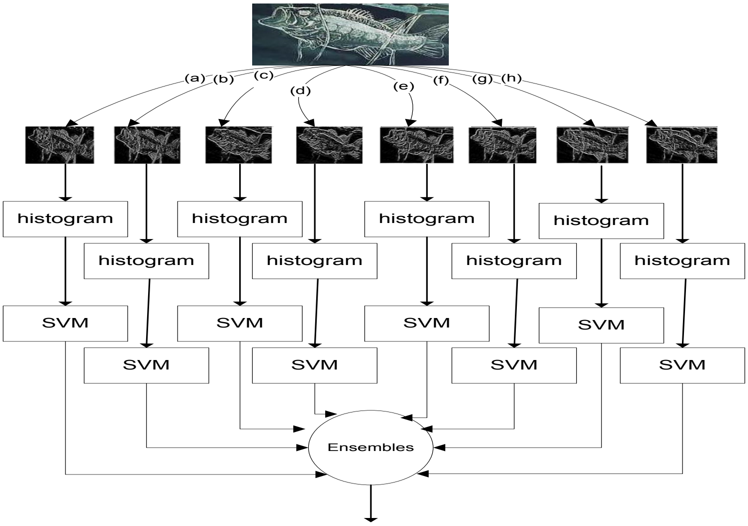

Fig. 1. Architecture of the ensembles of gradient-based descriptors using eight Robinson filters

Radial-Basis-Function (RBF) kernels in all experiments. Initially, all attributes in the training and testing sets were normalized to the interval [-1,+1] by using this equation:

$$x' = \frac{2(x - min)}{(max - min)} - 1. \quad (3)$$

The normalization is used to avoid numerical difficulties during the calculation and to make sure the largest values do not dominate the smaller ones. We also need to find the SVM parameters C and γ that perform best for the descriptors. To optimize the classification performance, the parameters were determined by using the libsvm grid-search algorithm [16]. We tried the following values $\{2^{-5}, 2^{-3}, ..., 2^{15}\}$ and $\{2^{-15}, 2^{-13}, ..., 2^{3}\}$ for C and γ, respectively. The values which gave the best accuracy performance with 5-fold cross-validation are picked and used to train on the training set.

4.2 Ensemble Methods for Combining Classifiers

Our previous works [17][18] showed that combining multiple features and classifiers with ensemble methods significantly increases classification performance. Ensemble methods have received considerable attention in the machine learning community to increase the effectiveness of classifiers. In order to construct a good ensemble classifier, the ensemble needs to construct accurate and diverse classifiers and to combine outputs from the classifiers effectively [19]. There exist several methods to obtain and combine the diverse classifiers. Here we employ three ensemble algorithms namely (1) product rule (2) mean rule [20] and (3) majority voting.

The product rule is one of the simplest and most efficient ways for combining outputs of classifiers [20]. When the classifiers have small errors and operate in independent feature spaces, it is efficient to combine their (probabilistic) outputs by multiplying them. Thus, we use this product rule to determine the final decision of the ensemble. First the posterior probability outputs $P_j^k(x^k)$ for class j of n different classifiers are combined by the product rule:

$$P_j^p(x^1,...,x^n) = \prod_{k=1}^{n} P_j^k(x^k) \tag{4}$$

where x^k is the pattern representation of the k^{th} descriptor. Then the class with the largest probability product is considered as the final class label belonging to the input pattern.

When estimators of the different classifiers contain large errors, it can be more efficient to combine their estimated probabilities by the mean rule [20] as follows:

$$P_j^m(x^1,...,x^n) = \frac{1}{n}\sum_{k=1}^{n} P_j^k(x^k) \tag{5}$$

Majority voting is the simplest and intuitive rule in combining multiple classifiers. It counts the collective judgement sets or votes for every classifier and applies a score. Let $d_{n,j} \in 0,1$ denote the decision outputs of the n^{th} classifier M_n, n=1...L and j=1...c, where L is the number of classifiers and c is the number of classes. If the n^{th} classifier selects class j, then $d_{n,j} = 1$ for correct, and zero for error. The vote will result in ensemble decision for t input class if:

$$\sum_{n=1}^{L} d_{n,t} = \max_{j=1}^{c} \sum_{n=1}^{L} d_{n,j} \tag{6}$$

Similar to the product rule and mean rule the class with the largest score is considered as the final class label.

In the experiments we will compare these ensemble methods to the naive approach that combines the feature vectors computed at all feature maps in one large feature vector.

5 Experiments and Results

For our comparison between the different descriptors and ensemble algorithms, a variety of image classes were chosen. The images should be common and familiar to machine vision researchers, and therefore we used a well known dataset, i.e. Caltech-101 [21]. The dataset contains various image sizes and were categorized into 101 different classes. However, in our experiment, only the first 20 classes were chosen for evaluation due to computational restrictions. In the dataset, each image consists of different sizes and contains different viewpoints, which makes the recognition process more challenging.

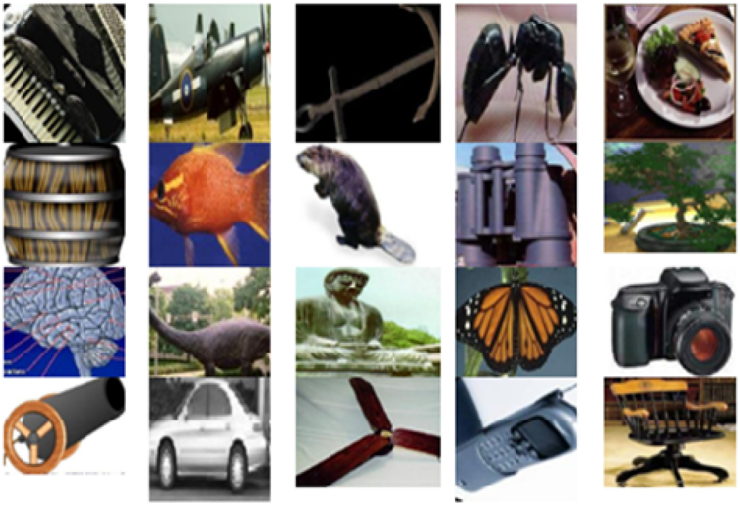

Fig. 2. Image examples with ground truth for different groups namely accordion, airplane, anchor, ant, background, barrel, bass, beaver, binocular, bonsai, brain, brontosaurus, Buddha, butterfly, camera, cannon, car side, ceiling fan, cell phone and chair respectively

5.1 Caltech-20 Dataset

The Caltech-101 is one of the most popular and widely used datasets to demonstrate the performance of object recognition systems [21]. It consists of 101 categories depicting real world object images such as camera, airplanes, bonsai, anchor, etc. In our experiments, we used the first 20 categories (in alphabetical category order) and a total of $20 \times 30 = 600$ images for evaluation. These images are all in JPEG format with medium resolution about 300 x 300 pixels and both in color and gray level representation. Fig. 2 shows the ground truth for the 20 different classes we used of the Caltech-101 dataset.

For evaluating the ensemble methods and the other single descriptors, we used 15 training and 15 testing images for each image class. To compute the performances of the different methods, we choose 10 times different training and test images randomly from a set of candidate images in the 20 classes of the Caltech-101 dataset. Finally, we report the performance using mean and standard deviation to verify significances of the obtained classification results.

5.2 The Detection of Intensity Changes

The process of extracting information from edges and orientations can be divided into three main tasks. The first task is to transform pixels in RGB color space into a more robust color space. In our case, YIQ color model is used to describe intensity information. In YIQ color space, only the **Y** component is used since this channel represents the intensity information. We used intensity values in the range of [0...255] per pixel, where the value 0 represents the minimum brightness and 255 the maximum brightness. The second task is to provide feature maps for the grayscale image. In this step, the image will be convolved using eight different Robinson filters. Once the feature maps are constructed, the last step is to apply gradient based descriptors that compute the first order image derivative to describe images and support vector machine algorithm for learning.

Table 1. The average classification accuracy (mean and SD) of the single and combination classifiers. M1=HG_180_G, M2=HG_360_G, M3=S_180_G, M4=S_360_G, and M5=EH_G. F1-F8=Filter, N=Naive, E1=Product Rule, E2=Mean Rule and E3=Majority Voting. The best result is marked in boldface.

	F1	F2	F3	F4	F5	F6	F7	F8	N	E1	E2	E3
M1	48.50 ±2.94	52.20 ±3.50	49.67 ±1.71	52.13 ±1.74	49.13 ±2.80	50.87 ±2.96	50.37 ±3.44	49.00 2.03	58.83 ±2.17	59.93 ±1.97	**60.07 ±1.95**	58.73 ±2.22
M2	48.47 ±2.46	50.47 ±1.57	50.0 ±1.97	50.67 ±2.16	48.93 ±2.07	50.20 ±2.49	50.33 ±1.86	50.53 ±2.30	- -	**60.7 ±1.83**	58.0 ±2.04	48.7 ±2.46
M3	52.83 ±2.87	56.37 ±3.70	54.50 ±2.47	55.47 ±3.41	53.87 ±3.70	57.14 ±2.29	53.67 ±2.68	56.03 ±2.10	61.37 ±3.24	**61.67 ±3.57**	61.43 ±3.55	59.87 ±3.98
M4	51.03 ±2.40	55.30 ±3.25	52.73 ±2.99	54.73 ±3.57	52.50 ±3.44	55.90 ±2.53	52.00 ±2.38	55.60 ±1.92	- -	**59.67 ±3.47**	59.50 ±3.43	58.7 ±2.40
M5	47.93 ±2.12	52.10 ±2.52	46.63 ±2.68	52.23 ±2.51	48.50 ±2.63	52.63 ±2.69	47.87 ±1.90	53.23 ±2.28	**59.40 ±2.38**	59.37 ±2.57	59.17 ±2.21	57.73 ±2.68

5.3 Results on Caltech-20

Table 1 shows the average classification accuracy and standard deviation of the different descriptors to classify images using the RBF kernel. The result shows that the average classification accuracy for each descriptor is best if all filter classifiers are combined i.e. naive and ensemble methods. It indicates that using a single filter is insufficient to completely describe objects due to narrow range of orientations. In contrast, using a larger set of filters produces wider range of orientations to enrich description of images. Besides, it provides better cooperation between classifiers to improve the final performance of the combination algorithms. In our experiments, we do not report results on HG_360_G and S_360_G of the naive method because previous results show that ensemble methods slightly outperform this approach. Thus, we believe both naive and ensembles have sufficiently rich information to describe objects using eight Robinson filters using a single descriptor.

We extended our experiments to combine all classifiers of the different descriptors on 20 classes. We compare the combination based on all filters combined with three ensemble methods namely product rule, mean rule and majority voting. The results are reported in Table 2. In this experiment, combining all classifiers of the different descriptors with the mean rule gives the best performance of 68.23%. This result shows that a combination of the descriptors performed very well with an ensemble of support vector machines. We do not report results on naive combination due to computation restriction of feature vector size. Besides, our previous works [17][18] show naive gives no improvement to increase classification result due to overfitting problem.

Table 2. The average classification accuracy of the different combination classifiers on 20 classes. M=Classifiers based on all filters combined. The best result is marked in boldface.

	Product Rule	Mean Rule	Majority Voting
M	67.93±2.19	**68.23±2.16**	67.67±2.74

6 Conclusions

In this paper, we have introduced an approach for recognizing objects in digitized images using classical compass filters. We reported a significant comparison of using filters in describing and classifying images namely (a) using a filter with a single feature descriptor, (b) using a set of different filters with a single feature descriptor, and (c) filters with different feature descriptors. The system uses different gradient based descriptors to compute the first order image derivative using eight Robinson's filters as convolution masks. A possible problem using a single filter to compute feature vectors, is that this operator only respond to edges in a narrow range of orientations. Thus, for a completely describe objects using edge information, a wide range of orientations is needed. Still, the best idea to describe images on the 20 classes is to use a combination of different filters and feature descriptors. This may be caused by its ability to keep structural relationships or cooperations between feature maps of the images. Using a single filter only to describe images gives worse result due to losing information about structures.

In future work we want to use color information to enrich semantic information for describing objects. And to use the spatial pyramid approach to construct multiple spatial resolution levels for each convolution mask.

Acknowledgment. The authors want to thank the Faculty of Information Science and Technology, University Kebangsaan Malaysia, for providing facilities and financial support under the exploration Research Grant Scheme No. UKM-TT-07-FRGS0253-2010, UKM-ERGS/1/2011/STG/UKM/02/41 and UKM-ERGS/2011/STG/IKM/2/48(TK). We also want to thank to Sanusi Azmi for helpful comments. The first author also wants to thank to the government of Libya for scholarship.

References

1. Kinnunen, T., Kamarainen, J.K., Lensu, L., Lankinen, J., Kälviäinen, H.: Making visual object categorization more challenging: Randomized caltech-101 data set (2010)
2. Larlus, D., Jurie, F.: Latent mixture vocabularies for object categorization and segmentation. Image Vision Comput. 27(5), 523–534 (2009)
3. Winn, J., Criminisi, A., Minka, T.: Object categorization by learned universal visual dictionary (2005)
4. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection (2005)

6. Manjunath, B.S., Salembier, P., Sikora, T. (eds.): Introduction to MPEG-7: Multimedia Content Description Language. Wiley (2002)
7. Won, C.S., Park, D.K., Park, S.J.: Efficient use of mpeg-7 edge histogram descriptor 24(1), 23–30 (2002)
8. Moreno, P., Bernardino, A., Santos-Victor, J.: Improving the sift descriptor with smooth derivative filters. Pattern Recogn. Lett. 30(1), 18–26 (2009)
9. Koenderink, J.J., van Doorn, A.J.: Generic neighborhood operators. IEEE Trans. Pattern Anal. Mach. Intell. 14, 597–605 (1992)
10. Zhu, J., Vai, M.I., Mak, P.U.: A new enhanced nearest feature space (enfs) classifier for gabor wavelets features-based face recognition (2004)
11. Arlt, B., Brause, R.W., Tratar, E.: Mascot: a mechanism for attention-based scale-invariant object recognition in images. Technical Report 00, 02, Informatik (2009)
12. Vapnik, V.N.: The nature of statistical learning theory (1995)
13. Garcia, C., Delakis, M.: Convolutional face finder: A neural architecture for fast and robust face detection. IEEE Trans. Pattern Anal. Mach. Intell. 26(11), 1408–1423 (2004)
14. Lux, M., Becker, J., Krottmaier, H.: Caliph and emir: Semantic annotation and retrieval in personal digital photo libraries (2003)
15. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. 20(3), 273–297 (1995)
16. Hsu, C.W., Chang, C.C., Lin, C.J.: A Practical Guide to Support Vector Classification (2000)
17. Abdullah, A., Veltkamp, R.C., Wiering, M.A.: An ensemble of deep support vector machines for image categorization (2009)
18. Abdullah, A., Veltkamp, R.C., Wiering, M.A.: Ensembles of novel visual keywords descriptors for image categorization (2010)
19. Dietterich, T.G.: Ensemble methods in machine learning (2000)
20. Tax, D.M., Duin, R.P., Breukelen, M.V.: Comparison between product and mean classifier combination rules (1997)
21. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Comput. Vis. Image Underst. 106(1), 59–70 (2007)