Image Categorization Using Local Probabilistic Descriptors

Extended version of the paper published in
Proceedings of the 18th International Conf. on Pattern Recognition, ICPR 2006

Katarina Mele[1,2], Jasna Maver[3], Dorian Šuc[1]
[1]Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, 1000 Ljubljana
[2]School of Computer Science and Engineering, University of New South Wales, 2052 Sydney
[3]Faculty of Arts, University of Ljubljana, Aškerčeva 2, 1000 Ljubljana
katarina.mele@fri.uni-lj.si, dorian.suc@fri.uni-lj.si, jasna.maver@ff.uni-lj.si

Abstract

Image categorization involves the well known difficulties with different visual appearances of a single object, but introduces also the problem of within-category variation. This within-category variation makes highly distinctive local descriptors less appropriate for categorization. In this paper we propose a family of local image descriptors, called ***probabilistic patch descriptors*** *(PPDs). PPDs encode the appearance of image fragments as well as their* ***variability*** *within a category. PPDs extend the usual local descriptors by modelling also the variance of the descriptors' elements, e.g. pixels or bins in a histogram. We apply PPDs to image categorization by using machine learning where the features are the matching scores between images and PPDs. We experiment with two variants of PPDs that are based on complementary local descriptors. An interesting observation is that combining the two PPD variants improves categorization accuracy. Experiments indicate benefits of modelling the within-category variation and show good robustness with respect to noise.*

1. Introduction

Image categorization, addressed in this paper, is a supervised learning problem with the goal of classifying new, unseen images into appropriate image categories by learning from a limited set of labelled images. Recognizing categories of objects, in addition to the well-known problems in computer vision, introduces the problem of within-category variation. Namely, considerably different objects are often in the same category. This makes the standard methods, which ignore the within-category variation, difficult to use.

Since images of the same category can differ substantially, but usually have some similar details, local image descriptors seem particularly interesting for image categorization. Many excellent local descriptors, including SIFT [9] and PCA-SIFT [7] have been developed recently. Their comparison is given in [11]. Although some of the best rated local descriptors are used for categorization task, it seems that they are not general enough and often too distinctive to be used as off-the-shelf tools for categorization tasks. Recently, some authors proposed region detectors designed for categorization task [6] and adaptations of local descriptors for more general recognition [12].

In this paper we introduce a family of local image descriptors called *probabilistic patch descriptors* (PPDs). PPDs encode the appearance of image fragments as well as their *variability* within a category. We apply PPDs to image categorization by using machine learning where the features are the matching scores between images and PPDs.

Ideas related to modelling the variability of patch descriptors appeared also in previous work [1, 4]. In contrast to our work, stable parts are assumed to be inside the regions representing an object. Our approach considers also parts of object's background, when this is informative for categorization, e.g. grass for categorizing cows.

Section 2 gives the idea of probabilistic patch descriptors. Categorization method is proposed in Section 3. Section 4 gives experimental evaluation using images of five categories and study the robustness of the method with respect to noise and occlusions. Finally we give conclusion and some directions for further work.

2. Probabilistic Patch Descriptor (PPD)

Probabilistic patch descriptor represents a set of similar patches by modelling probabilistic distributions of the elements of patch descriptors. PPD augments each element

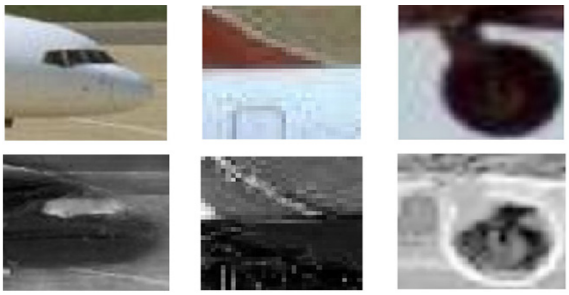

Figure 1. Three RGB-PPDs modelling a nose, a trunk , and a wheel of an airplane.

of the original "non-probabilistic"patch-descriptor with its variance. In the most simple case, where a patch is described by values of pixels, a PPD models probabilistic distributions of pixels' values. If a patch is described by a histogram, the PPD models also the variance in every bin. In a PPD each element of the patch descriptor is assumed to have a normal probability distribution $\mathcal{N}(\mu_i, \sigma_i), i = 1 \dots N_e$. Here N_e is the number of elements of the patch descriptor. Therefore a PPD named R can be given as a pair of vectors (μ^R, σ^R) denoting means and standard deviations of R.

In this paper we experiment with two probabilistic patch descriptors: RGB- and RIFT-PPD. RGB-PPD is based on a patch represented by a vector of pixels. In all our experiments we used patches of size 15×15. Since RGB-PPD models three color components, the number of elements of the patch descriptor is $N_e = 15 \times 15 \times 3$. Examples of three RGB-PPDs, modelling parts of an airplane are given in Figure 1. These RGB-PPDs were learned from images of the airplane category as described later in this section. Each PPD is described by two images illustrating its means (images in upper row) and variances (images in bottom row). Variances for R, G, and B color components are not shown separately, but instead the sum of all three components is shown. Darker colors in images of the bottom row denote smaller variances of spatially corresponding pixels in images above them. For example the airplane's nose has relatively small variance (dark colors in the bottom left image), but there is more variability in pilot's window that can be of different shape or color, or placed at different positions.

RIFT-PPD is based on RIFT, a variant of a SIFT descriptor [9]. SIFT models a patch at the scale-space peaks in a series of difference-of-Gaussian images. Such keypoints are highly distinctive and very sensitive to variability in appearance. This is prohibitive for localization of parts common to objects of one category, where we usually have high variability. Accordingly, our initial results with categorization using SIFT-PPDs did not give favorable results. To avoid the problem with high sensitivity of SIFT keypoints, we use boundaries [10] as the criteria to detect the interest points. Therefore, RIFT descriptors are actually SIFT descriptors that are calculated from patches centered at the boundary points in an image. Since RIFT (Rotational Invariant Feature Transform) is not applied on a patch at scale-space extreme it loses invariance to scale, but it still preserves invariance to rotation.

2.1. PPD Matching Score

The idea of PPD matching score is to model the probability that corresponding elements of two PPDs have a similar value. As elements are treated as random variables, this probability depends on their distribution and is generally smaller with higher variances of elements. Since we assume that PPD elements are independent variables, an appropriate measure of matching between PPDs T and R is the product of probabilities that values of corresponding elements differ by less than some small constant δ (see Eq. 1). For small δ, the integral in Eq. 1 has an approximate analytical solution that is linearly dependent on δ. We can discard δ, because we only need to compare matchings of different PPDs and we do not need the exact probability of matching. Therefore we can use the solution of the integral to define the *matching score* $m(T, R)$ between two PPDs:

$$\prod_i \int_{-\infty}^{\infty} \int_{x-\delta}^{x+\delta} \mathcal{N}(x; \mu_i^T, \sigma_i^T) \mathcal{N}(y; \mu_i^R, \sigma_i^R)\, dy\, dx. \qquad (1)$$

$$m(T, R) = \prod_i \frac{e^{-\frac{(\mu_i^T - \mu_i^R)^2}{2((\sigma_i^T)^2 + (\sigma_i^R)^2)}}}{\sqrt{2\pi}\, \sigma_i^T\, \sigma_i^R \sqrt{\frac{1}{(\sigma_i^T)^2} + \frac{1}{(\sigma_i^R)^2}}}\ . \qquad (2)$$

The above matching score defines a probability-based similarity measure between two PPDs. Matching score is also used to estimate the similarity of a PPD and an image. In this case, the matching score refers to the highest matching score over all possible positions of a patch descriptor on the image. This usually requires to calculate "non-probabilistic" patch descriptors for all possible patch positions on the image and transform these "non-probabilistic" patch descriptors into PPDs by assuming a small standard deviation σ_{noise} in values of the patch descriptors elements. We refer to the image patch descriptor with the highest matching score as *the most similar image patch descriptor*.

For example, the matching score of a RGB-PPD and an image refers to the highest matching score over all possible image patches having the same size as the RGB-PPD. To compute the similarity between a RGB-PPD and an image patch we assume that image pixels' values are distributed normally having small standard deviations σ_{noise} and the pixels' values as means. In our experiments, where pixels' take on values from $[0, 255]$, we used $\sigma_{noise} = 1$. This means that we assume that images are not very noisy.

2.2. Learning PPD

PPDs are learned for each category separately by randomly selecting a set of initial patches, calculating the corresponding patch descriptors and estimating the variances of the patch descriptors' elements. To obtain an initial set of patches for one category we select at random a number of patches from images of the selected category. Each initial patch is transformed into a so-called initial PPD I by setting μ^I to the patch descriptor values and setting σ^I to σ_{noise}. For each initial PPD we form a set of image patch descriptors that are most similar to the initial PPD. This set, denoted by $\mathcal{M}$, contains the most similar image patch descriptor from each image of the category.

An initial PPD and the corresponding set of the most similar patch descriptors are then used to form a PPD that is described with vectors μ and σ. The PPD retains means of the initial PPD, that is $\mu = \mu^I$. Variances are calculated from the set of most similar image patch descriptors. Here the contribution of each image patch descriptor $M \in \mathcal{M}$ is weighted by the average probability that two corresponding elements from I and M differ by less than δ.

In this way, we describe each category with a set of PPDs. As they are generated from randomly selected initial patches, some of these PPDs might be similar or do not discriminate well between different categories. Selection of representative PPDs is described in Section 3.

2.3. Accuracy Benefits of PPD

To show that modelling of within-category variation is beneficial for image categorization we compared the categorization accuracy using RGB-PPDs and their variant that does not model the variability, called *constant-variance RGB-PPDs* where the variance of all pixels' components is set to a constant $\sigma_{noise} = 1$.

We used images of five categories from Caltech-101 database [3]. These categories are airplanes, cars, motors, faces, and leopards. For each category we used 100 images, 50 for training and 50 for testing. RGB-PPDs were learned as described in the previous section. In this way we generated 50 RGB-PPDs of size 15×15 pixels for each category.

For categorization we used threshold classifiers that decide if an image belongs to a category based on a single PPD. Each threshold classifier consists of a PPD and a threshold value. If the matching score between an image and the PPD is above the threshold, the image is classified in the category of the PPD. In the training phase the threshold is set to the value of the matching score that recognizes the category of the PPD with the minimal error. Each of the 250 RGB-PPDs (50 RGB-PPDs from each of the five categories) was used with the other four categories, giving altogether 1000 threshold classifiers that were trained on the training images. In the same way, we also learned 1000 threshold classifiers that use *constant-variance RGB-PPDs*.

These threshold classifiers were evaluated on the testing images. The average errors are 29.1% for RGB-PPDs and 32.2% for *constant-variance RGB-PPDs*. According to the t-test the mean errors using RGB-PPDs are significantly better at 1% significance level and also at considerably smaller significance levels. This suggests that modelling the within-class variability of patch descriptors is beneficial for image categorization.

3. Categorization using PPDs

Our method for image categorization learns PPDs from training images and use PPDs to construct features that are then used for classification using standard machine learning methods. PPDs used in our experiments were RGB-PPDs, RIFT-PPDs and their combination denoted by RGB&RIFT-PPD. We decided for the combination of RIFT and RGB based PPDs, since these two descriptors are somehow complementary. RGB based PPDs emphasize homogenous regions. Here colors play an important role. The boundaries between regions inside a RGB-PPD are usually the parts with the largest variation and consequently the shapes of regions are just slightly indicated. On the other hand, RIFT-PPDs are calculated from grayscale images and discard the color information, but use a kind of gradient that captures the shapes of edges.

To prevent overfitting, training images are divided into two sets, called *PPD-learning images* and *PPD-matching images*. PPD-learning images, consisting of 50% randomly selected images are used to learn PPDs. Other images, called PPD-matching images are used to calculate matching scores between images and PPDs. These matching scores are the features for machine learning. In this way, features are calculated on an independent set of images, not on the images they were generated from.

The method has the following four steps. First, given PPD-learning images of a category, learn a set of PPDs as described in Section 2.2. Second, the set of representative PPDs is selected by removing similar PPDs and PPDs that do not match their category significantly better than other categories. Second, we remove similar PPDs. For each pair of PPDs of the same category we calculate the average probability that two corresponding elements of the patch descriptor match and remove one PPD, if this probability is above 0.01. Third, we retain only PPDs that have significantly higher matching scores on images from its category than on images of at least one other category. To compare matching scores corresponding to two categories we used the two-sample t-test at 1% significance level. In this way, we get the set of representative PPDs that are in the fourth step used to compute features for standard machine learning

methods. Namely, for each PPD-learning image we form a feature vector by computing the matching scores of the image with all representative PPDs. These feature vectors and the corresponding category labels are then used to learn a classifier with machine learning methods. Images with unknown category labels, i.e. test images, are categorized by the learned classifier based on corresponding features. In our experiments we used a Support Vector Machine (SVM) with linear kernel.

This method requires that objects from one category are of similar size on all images. It does not require alignment, segmentation or other preprocessing steps. Except for extreme changes, image sizes do not affect the success of the method as far as objects of same category are of similar size. Because PPDs are usually generated from slightly different object scale, some scale invariance is already incorporated into the model. Robustness to changes in scale is increased by treating images at three different scales: 90% 100% and 110% of the original image size. This requires to compute matching scores using images at three different sizes as described in Section 2.1. It should be noted that some scale robustness is already incorporated into PPDs by modelling the variability of a patch descriptor. If a detail of an object category often appears in different sizes, this is reflected also in the variance of the corresponding PPD.

4. Experiments using PPDs

Here we describe experimental results with categorization using images of five categories from the Caltech-101 Object Categories database [3]. For comparison of results we used categories that were frequently used by other authors. These are images of airplanes, cars, motors, faces, and leopards. We did not do any alignment, segmentation or other preprocessing. PPDs used in our experiments were RGB-PPDs, RIFT-PPDs and their combination RGB&RIFT-PPDs. With RGB&RIFT-PPDs each image is described both with features based on RGB-PPDs and features based on RIFT-PPDs. The features obtained from PPDs are used for image classification using SVM.

In experiments we used 125 randomly selected images of each category. 100 images were always used for training, other 25 as a test set. All the results were obtained by five times five-fold cross-validation. Since the method randomly selects initial patches from PPD-learning images, we initially studied also the variations in results that come from using different initial patches. We noticed that the variations in the results obtained by different (random) selection of initial patches are very small and that the selection of initial patches does not effect the categorization accuracy much.

When learning PPDs we used 200 initial patches from PPD-learning images of each category. On average 20% of initial PPDs of a category were discarded due to the similarity. With RGB-PPDs, after removing non-discriminative PPDs there was about 100 representative RGB-PPDs left in each category. With RIFT-PPDs this number was about 130.

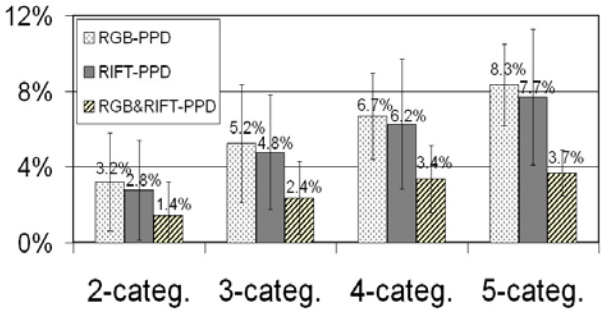

Figure 2. Error rates for all combinations of 2, 3, 4, and 5 categories using RGB-PPD, RIFT-PPD and their combination RGB&RIFT-PPD.

The matching scores of RGB-PPDs and images are calculated efficiently using convolution. Still, this is the most time-consuming operation since a large number of the matching score computations is required. On a Pentium III computer it requires about a minute to calculate features for one image of size $200 \times 200 \times 3$. With RIFT-PPDs computing the matching scores is much faster, since RIFT descriptors are calculated only at sampled boundary points.

4.1. Experimental Results

We experimented with five image categories. In all experiments we separately give error rates for categorization into two, three, four and five categories. All these results are averages over all possible combinations of k categories, i.e. average of $\binom{5}{k}$ combinations.

Figure 2 gives the error rates for RGB-PPDs, RIFT-PPD, and their combination called RGB&RIFT-PPDs. These results are given also in Table 1. The first observation was that all three methods perform quite well. For example, with the categorization into two categories, the average error rates for all three methods are less than 3.2%. As expected, the error rates slightly increase with increasing number of categories. The results with RIFT-PPDs are slightly better than with RGB-PPDs, although the improvements are not significant. The most notable are the improvements of RGB&RIFT-PPDs with respect to both other methods. This happens consistently with different categories and different number of categories. The average error rates fall from 5.86% and 5.36% for RGB- and RIFT-PPD, respectively, to only 2.71% for the combination of both PPD types. The improvements of RGB&RIFT-PPDs with respect to both RGB-PPDs and RIFT-PPD are significant at 1% significance level with all different numbers of categories we used, i.e. two, three, four and five categories.

Table 1. Average error rates (in percentages) with different variants of PPDs, with added noise, cropped images and reduced training set. Standard deviations are in the brackets.

PPD type	2categ.	3 categ.	4 categ.	5 categ.
RGB	3.2 (2.6)	5.2 (3.1)	6.7 (2.3)	8.3 (2.2)
RIFT	2.7 (2.6)	4.7(3.0)	6.2 (3.4)	7.7 (3.6)
RGB&RIFT	1.4 (1.8)	2.3 (1.9)	3.3 (1.8)	3.7 (1.2)
RGB, 5%	3.1 (2.5)	5.1 (3.0)	8.9 (2.2)	8.9 (2.2)
RGB, 20%	11.3(14.9)	14.7 (10.7)	18.7 (7.3)	21.7 (3.5)
RGB, crop	6.2 (5.4)	10.1 (5.2)	13.5 (4.7)	16.0 (4.8)
RGB, red.	3.3 (4.2)	5.3 (3.3)	9.1 (3.5)	10.2 (2.5)

Table 2. The published EERs for category recognition problems with related methods and the true-positive rates for the six-category problem with our method

	airp.	cars	mot.	fac.	leop.	bg.
[4], EER	90	90	93	96	90	-
[15], EER	68	-	84	-	-	-
[8], EER	-	94	94	-	-	-
[14], EER	84	90	93	83	-	-
[5], EER	98	99.9	-	99.9	-	-
RGB&RIFT-PPDs, TP rates	94	94	94	94	89	89

With RGB-PPDs we evaluated also the robustness with respect to noise, occlusions and small training set. To assess the robustness with respect to noise, we corrupted the test images with 5% and 20% of uniformly distributed noise. Here, the noise percentage denotes the percentage of noise range in the range of the pixel's value.

The occlusions were simulated by cropping 40% of images from left or from right side. The results given in Table 1 (row RGB, crop) show that the method is very robust with respect to noise in images and occlusions. In the last experiment we used only 25 images with no noise for training. The results in Table 1 (last row) show that the error rates increase only slightly when compared to using 100 training images. These results show that the method is well-suited also for learning from a small number of images.

4.2. Comparing Results with Related Work

Here we compare the results with other authors that used the same image database. These authors give the results with the "recognition of a single category". Namely, they consider a binary learning problem where an image is classified either in a single object category or as so-called *background*. This is different than our method, which is designed to classify into a set of predefined categories. These differences in the learning problem require some consideration when comparing the results.

To compare the results we added a background category consisting of 125 images that were selected at random from so-called "Background-Google" category. In contrast to other categories we did not use PPD-learning images for this category, since we did not expect to learn any representative PPDs for the background category. Therefore, SVM uses only the matching scores with PPDs of the other five categories to classify an image in one of the six categories.

Table 2 gives the five-fold cross-validation true-positive rates for each of the six categories. The table also gives published equal-error rates (EER) of related methods. The results in Table 2 should be compared carefully. First, our results were obtained with categorization into six categories, i.e. five original categories plus background category. The results of other authors are obtained with binary classifiers. In general, binary classification is more robust. Second, the table gives true-positive rates for our method, but EER with other methods. EER is found by varying a threshold defining when an image is classified as an object category or as background and can be compared to true-positive rate when the categories have the same number of images. Although the results cannot be directly compared, they suggest that our method is comparable to the state-of-the-art methods also when used in the presence of a background category, for which the method was not designed in the first place.

We compared our method also to the method of Csurka *et al.* [2] that is designed for categorization into multiple categories, and is in this sense more similar to our method. When we repeated our experiments with their experimental settings, the average error rate was 3.6% with our method and 3.9% with their method.

5. Conclusions

We presented probabilistic patch descriptors that are designed for categorization of still images. These descriptors are based on the usual patch descriptors, but explicitly model the within-category variation of a patch descriptor. Experimental results in Section 2.3 show that modelling the within-category variation is beneficial for image categorization. PPDs are learned in an unsupervised manner for each category separately. They are applied to image categorization by machine learning using features calculated as the matching scores between images and PPDs.

We developed two variants of probabilistic patch descriptor, called RGB-PPD and RIFT-PPD. The two PPD variants give in a way complementary descriptions of an image fragment. The combination of these two descriptors

significantly improves categorization accuracy. This is in itself an interesting, although not a surprising result. Except for the two methods [13, 5], that use grayvalues and different moment descriptors with a Boosting algorithm, we are not aware of any other work on combining different local descriptor for categorization of still images. The experimental results demonstrate benefits of modelling the within-category variation, show that the method gives results that are comparable to the state-of-the-art categorization methods, can cope with noise and occlusions, and learns well also from a small set of training images. As information of spatial arrangement of PPDs is not a part of the model, the proposed approach can also handle objects with non-rigid shape, e.g. leopards.

We plan to experiment also with PPDs based on other image descriptors, e.g. moments or steerable filters, that show best matching performance among low dimensional descriptors [11], and to explore the perspectives of combining different local descriptors for image categorization.

Acknowledgements

The work described in this paper was supported by Slovenian Ministry of Education, Science and Sport, the European Commission's Sixth Framework Programme under contract no. 029427 as part of the Specific Targeted Research Project XPERO, and partially by Slovenian IT company SRC.SI. Many thanks to Prof. Arcot Sowmya and Prof. Yael Moses for fruitful discussions.

References

[1] E. Borenstein and S. Ullman. Learning to segment. In *ECCV04*, volume 3, pages 315–328, 2004.

[2] G. Csurka, G. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Proceedings ECCV*, Prague, Czech Republic, May 2004.

[3] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples an incremental bayesian approach tested on 101 object categories. In *CVPRMW*, volume 12, page 178, 2004.

[4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, volume 2, pages 264–271, June 2003.

[5] M. Fussenegger, A. Opelt, A. Pinz, and P. Auer. Object recognition using segmentation for feature detection. In *ICPR (3)*, pages 41–44, 2004.

[6] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *ECCV*, volume 1, pages 228–241, Prague, Czech Republic, 2004.

[7] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *CVPR (2)*, pages 506–513, 2004.

[8] B. Leibe and B. Schiele. Scale-invariant object categorization using a scale-adaptive mean-shift search. In *DAGM-Symposium*, pages 145–153, 2004.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[10] D. R. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *PAMI*, 26(5):530–549, 2004.

[11] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005.

[12] K. Mikolajczyk, A. Zisserman, and C. Schmid. Shape recognition with edge-based features. In *British Machine Vision Conf.*, volume 2, pages 779–788, 2003.

[13] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *ECCV (2)*, pages 71–84, 2004.

[14] J. Thureson and S. Carlsson. Appearance based qualitative image description for object class recognition. In *ECCV 2004*, pages 518–529, 2004.

[15] M. Weber. *Unsupervised Learning of Models for Visual Object Class Recognition*. PhD thesis, CIT, 2000.