

# Image features classification using Bayesian model

Vojtěch Dornák<sup>1,2,a)</sup>, Martin Čermák<sup>1,2</sup>, Marek Pecha<sup>2</sup> and Lukáš Pospíšil<sup>1</sup>

<sup>1</sup>*Department of Mathematics, Faculty of Civil Engineering, VSB-TU Ostrava, Czech Republic*

<sup>2</sup>*Department of Applied Mathematics, FEECS, VSB-TU Ostrava, Czech Republic*

<sup>a)</sup>Corresponding author: vojtech.dornak@vsb.cz

**Abstract.** This paper presents the first results of the Bayesian model in the problem of image classification using feature extraction. The proposed approach is similar to well-known Markov processes and is based on conditional probabilities determined from the given data using the regression process with Kullback-Leibler divergence. In the contribution, we shortly review the complete process consisting of the Scale Invariant Feature Transform technique for the detection of key features, the so-called Bag of Visual Words, and the classification. We compare the new methodology with the standard approach based on Support Vector Machine on a non-trivial benchmark.

## Introduction

In this paper, we introduce the Bayesian model and the Support Vector Machine classification techniques. We experiment with these techniques on classifying photographs of cats and dogs from the Oxford-IIIT Pet Dataset [1], transformed into a space of significant features using the Scale Invariant Feature Transform technique [2].

It has been shown in [3], that classifying the transformed data using the Support Vector Machine is a viable option. In [4], these classification techniques have been explored on multiple different datasets. In this paper, we compare the results of the two classification techniques.

## Scale Invariant Feature Transform

Scale Invariant Feature Transform (SIFT) is a commonly used local feature extractor, firstly proposed by David Lowe in 2004 [2]. It describes an area around a few selected key points by the means of a vector called a descriptor. These descriptors are invariant to scale, rotation, illumination and affine transformations.

Let the scale-space representation of an image  $L(x, y, \sigma)$  be the convolution of the input image with the Gaussian kernel. The scale-invariant key points are selected as extrema of the scale normalized Laplacian of Gaussian (LoG):

$$\Delta_{norm}L(x, y, \sigma) = \sigma \left( \frac{\partial^2 L(x, y, \sigma)}{\partial x^2} + \frac{\partial^2 L(x, y, \sigma)}{\partial y^2} \right). \quad (1)$$

Since the determination and the evaluation of LoG is time-consuming, it is approximated by the Difference of Gaussian (DoG). The DoG is computed by subtracting two adjacent scale-space representations of the image, which is performed across different scales, where the scales are constructed by progressively convolving the original image with the Gaussian kernel.

Each candidate pixel is compared with the 8 surrounding pixels as well as with the 9 pixels in the scale above and the 9 pixels in the scale below. If the candidate pixel intensity is lower or higher than the 26 surrounding pixels, it is selected as a key point. To detect the sub-pixel location of extrema, the DoG is interpolated using a quadratic Taylor expansion at each candidate. The key point is discarded, if the offset from the key point is greater than 0.5 in any dimension, as this indicates that the extremum is close to a better key point. key points located along edges are also considered poorly located and therefore are discarded.

The key point descriptors for the object must be almost identical across different scales, rotations, illuminations and other transformations. To ensure descriptor invariance to rotation, the key point orientation needs to be determined. It is selected as a dominant gradient magnitude, from an orientation histogram.

The descriptor is constructed from a window of the size  $16 \times 16$  pixels around the key point. This window is rotated according to the previously selected orientation of the key point. The window is divided into 16 ( $4 \times 4$ ) sub-windows. In each sub-window, an 8 bin histogram weighted by the gradient magnitudes is constructed. These histograms form a descriptor vector.

The SIFT extractor provides us with a various number of descriptors for each image. Since we require each image to be represented by a single vector, we apply the Bag of Words technique (BoW). The BoW represents each sample in a dataset as a multi-set of its words. In our case, the words are a categorization of descriptors from all images. The categorization is done using  $k$ -means clustering.

## Support Vector Machines

The Support Vector Machine is a supervised learning model designed for binary data classification [5]. It is well known and commonly used for classification in various fields. In the case of image classification, promising results have been proposed in [3].

Let  $T := \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$ , be the training dataset, where  $n$  is the number of the samples,  $\mathbf{x}_i \in \mathbb{R}^m$ ,  $i \in \{1, 2, \dots, n\}$ , are samples and  $y_i \in \{-1, 1\}$  are corresponding sample labels. The classification model is represented in the form of the hyperplane  $H : \omega^\top \mathbf{x} - \tilde{b} = 0$ , where  $\omega \in \mathbb{R}^m$  is the normalized normal vector of the hyperplane  $H$ , and  $\tilde{b} = \frac{b}{\|\omega\|} \in \mathbb{R}$  is the bias from the origin.

By augmenting the vector  $\omega$  and each sample  $\mathbf{x}_i$  with an additional auxiliary dimension, so that  $\hat{\omega} \leftarrow [\omega \ B]^\top$ ,  $\hat{\mathbf{x}}_i \leftarrow [\mathbf{x}_i \ \beta]^\top$ , where  $B \in \mathbb{R}$ , and  $\beta \in \mathbb{R}^+$  is a user defined variable, the bias  $\tilde{b}$  is included in the problem. Such approach is well known as the no-bias classification [6].

The hyperplane  $H : \hat{\omega}^\top \hat{\mathbf{x}} = 0$  can be determined using a constrained optimization problem in the primal formulation:

$$\arg \min_{\hat{\omega}, \xi_i} \frac{1}{2} \|\hat{\omega}\|^2 + \frac{C}{p} \sum_{i=1}^n \xi_i^p \quad \text{s.t.} \quad \begin{cases} y_i(\hat{\omega}^\top \hat{\mathbf{x}}_i) \geq 1 - \xi_i, \\ \xi_i \geq 0 \text{ if } p = 1, i = 1, \dots, n, \end{cases} \quad (2)$$

where  $\xi_i = \max(0, 1 - y_i(\hat{\omega}^\top \hat{\mathbf{x}}_i))$  is the hinge loss function,  $C \in \mathbb{R}^+$  is a penalty, that penalizes misclassification error. The appropriate penalty  $C$  can be determined using the grid-search technique combined with cross-validation method.

Exploiting the Lagrange duality and evaluating the Karush-Kuhn-Tucker conditions for (2), we obtain the  $l1$ -loss  $l2$ -regularized SVM formulation

$$\arg \min_{\lambda} \frac{1}{2} \lambda^\top H \lambda - \lambda^\top \mathbf{e} \quad \text{s.t.} \quad \mathbf{o} \leq \lambda \leq C \mathbf{e}, \quad (3)$$

for  $p = 1$ . For  $p = 2$ , we get the  $l2$ -loss  $l2$ -regularized SVM formulation

$$\arg \min_{\lambda} \frac{1}{2} \lambda^\top (H + C^{-1} \mathbf{I}) \lambda - \lambda^\top \mathbf{e} \quad \text{s.t.} \quad \mathbf{o} \leq \lambda. \quad (4)$$

In the proposed optimization problems,  $H = Y^\top G Y$ ,  $G = X^\top X$  denotes the Gramian matrix,  $X = [\hat{\mathbf{x}}_1 \ \dots \ \hat{\mathbf{x}}_n]$ ,  $Y = \text{diag}(\mathbf{y})$ ,  $\mathbf{y} = [y_1, y_2, \dots, y_n]^\top$ ,  $\mathbf{e} = [1, 1, \dots, 1]^\top \in \mathbb{R}^n$ ,  $\mathbf{o} = [0, 0, \dots, 0]^\top \in \mathbb{R}^n$ .

In the  $l2$ -loss  $l2$ -regularized SVM formulation, the Hessian matrix  $H$  regularized by the matrix  $c^{-1} \mathbf{I}$  is symmetric positive definite. In this case, the objective function is strictly convex and the problem (4) has unique solution [7].

## Bayesian Model

The method is based on law of total probability for two categorical processes. Let  $x_t \in X := \{x_1, \dots, x_{K_x}\}$ ,  $t = 1, \dots, T$  be the input variables and let  $y_t \in Y := \{y_1, \dots, y_{K_y}\}$  be output labels. Let  $\Pi_{x_t}^{K_y}$  be a stochastic vector of relative frequencies of word categories  $X$  in image  $t$  and  $\Pi_{y_t} \in \mathbb{R}^{K_y}$  be a vector of probabilities of image classification. Then

using on the law of total probability, there exists a left stochastic matrix of conditional probabilities  $\Lambda \in \mathbb{R}^{K_y \times K_x}$  with components

$$\Lambda_{ij} = P(y_t = y_i | x_t = x_j), i = 1, \dots, K_y, j = 1, \dots, K_x. \quad (5)$$

The determination of the optimal  $\Lambda^*$  can be written as regression problem

$$\Lambda^* = \arg \min_{\Lambda \in \Omega_\Lambda} \sum_{t=1}^T \text{dist}(\Pi_{y_t}, \Lambda \Pi_{x_t}), \quad (6)$$

where  $\Omega_\Lambda$  is a feasible set of left stochastic matrices. In our approach, the  $\text{dist}(\Pi_{y_t}, \Lambda \Pi_{x_t})$  is chosen as Kullback-Leiber divergence [8]:

$$\text{dist}(\Pi_{y_t}, \Lambda \Pi_{x_t}) = - \sum_{i=1}^{K_y} \Pi_{y_t}^i \ln \frac{[\Lambda \Pi_{x_t}]_i}{[\Pi_{y_t}]_i} = - \sum_{i=1}^{K_y} [\Pi_{y_t}]_i (\ln [\Lambda \Pi_{x_t}]_i - \ln [\Pi_{y_t}]_i) \propto - \sum_{i=1}^{K_y} [\Pi_{y_t}]_i \ln [\Lambda \Pi_{x_t}]_i. \quad (7)$$

Since the objective function of (6) is convex, we can use Jensen's inequality to obtain

$$- \sum_{i=1}^{K_y} [\Pi_{y_t}]_i \ln [\Lambda \Pi_{x_t}]_i \leq - \sum_{i=1}^{K_y} [\Pi_{y_t}]_i \left( \sum_{j=1}^{K_x} [\Pi_{x_t}]_j \ln (\Lambda_{ij}) \right) = - \sum_{i=1}^{K_y} \sum_{j=1}^{K_x} [\Pi_{y_t}]_i [\Pi_{x_t}]_j \ln \Lambda_{ij}. \quad (8)$$

The problem (6) with objective function (8) has analytical solution, which can be derived using Lagrange multipliers

$$\Lambda_{ij}^* = \frac{\sum_{t=1}^T [\Pi_{y_t}]_i [\Pi_{x_t}]_j}{\sum_{i=1}^{K_y} \sum_{t=1}^T [\Pi_{y_t}]_i [\Pi_{x_t}]_j}. \quad (9)$$

Please notice the similarity of obtained formula with the estimation of Markov chain transitional matrix estimated by Matimum Likelihood Principle [9]. In this paper, we compare solution (9) with numerical solution of original problem (6) without using the Jensen's inequality. Since the feasible set  $\Omega_\Lambda$  is a closed convex set and we can easily derive the derivatives of objective function, the problem can be solved numerically using the Spectral Projected Gradient method [10]. We compare both approaches in our experiments.

## Benchmark Results

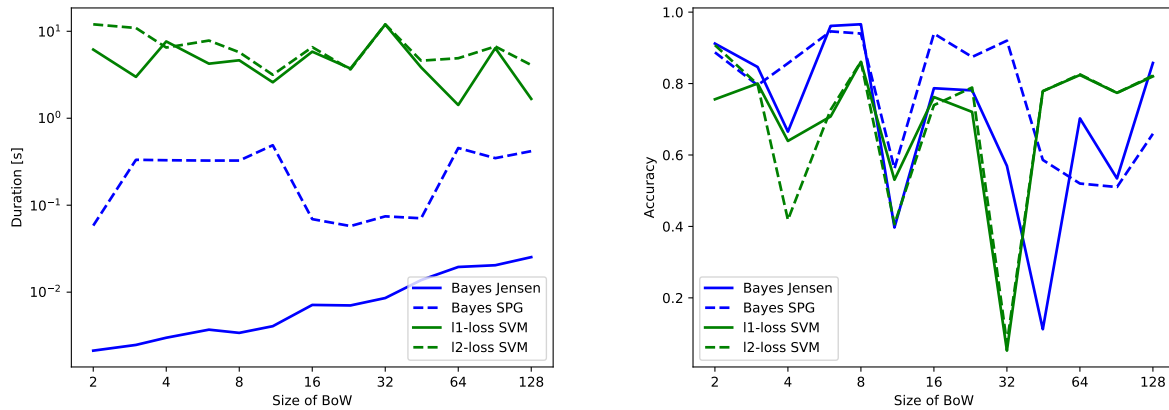
To test our extraction-classification pipeline, we use the Oxford-IIIT Pet Dataset [1]. The dataset contains over 7,000 photographs of different cat and dog breeds. The dataset provides a trimap for each photograph, allowing us to cut out the animal from the background. This allows us to concentrate only on classifying the data relevant to the animal. We assign 90% of the data to the training subset and 10% to the testing subset.

We run the presented experiments on the Salomon supercomputer [11] at IT4Innovations. As the underlying SVM solver, we use PermonSVM [12]. The Bayesian Model using the Jensen inequality is implemented in Python, while the Bayesian Model solved using the Spectral Projected Gradient method is trained using an Octave code.

From the graphs FIGURE 1 we can observe that for the lower BoW sizes both versions of the Bayes classifier perform better in regards to accuracy and the duration of training. However, with the larger BoW sizes, both the  $l1$ -loss and the  $l2$ -loss regularized SVM performs similarly to the Bayes classifier.

## Conclusion

In our contribution, we presented the Bayesian model in image classification and demonstrated the efficiency of this novel approach on the standard non-trivial benchmark. Our first results indicate better performance than commonly used SVM. Additionally, our method can be easily extended to the classification of images into more classes, whereas the SVM approach can be directly used only for binary classification. However, this feature as well as the comparison to other commonly used techniques (such as Neural Networks) would provide a better understanding of the efficiency of this promising method. This will be the aim of our further research.



**FIGURE 1.** The results of image classification benchmark: the computational time (left, lower is better) and the classification accuracy of proposed methods (right, higher is better).

## Acknowledgement

This contribution has been completed thanks to the financial support provided to VSB-Technical University of Ostrava by the Czech Ministry of Education, Youth and Sports from the budget for conceptual development of science, research and innovations for the 2021 year.

## REFERENCES

- [1] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, “Cats and dogs,” in *IEEE Conference on Computer Vision and Pattern Recognition* (2012).
- [2] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision* **60**, 91–110 (2004).
- [3] V. Dornák, M. Čermák, M. Pecha, and L. Pospíšil, “SIFT feature extraction applied in SVM classification,” (2020 (accepted)).
- [4] V. Dornák, “Algorithms used for classifying visual signals using methods for extracting significant features,” Diploma thesis, VŠB - Technical University of Ostrava 2021, supervised by Martin Čermák.
- [5] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* (1992), pp. 144–152.
- [6] M. Pecha and D. Horák, “Analyzing l1-loss and l2-loss support vector machines implemented in PERMON toolbox,” in *International Conference on Advanced Engineering Theory and Applications* (Springer, 2018), pp. 13–23.
- [7] Z. Dostál, *Optimal Quadratic Programming Algorithms, with Applications to Variational Inequalities*, Vol. 23 (SOIA, Springer, New York, US, 2009).
- [8] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The annals of mathematical statistics* **22**, 79–86 (1951).
- [9] J. R. Norris, *Markov chains*, Cambridge series in statistical and probabilistic mathematics (Cambridge University Press, 1998).
- [10] E. G. Birgin, J. M. Martinez, and M. Raydan, “Nonmonotone spectral projected gradient methods on convex sets,” *SIAM Journal on Optimization* **10**, 1196–1211 (2000).
- [11] IT4Innovations: National Supercomputing Center, VSB-Technical University of Ostrava, Salomon cluster documentation – hardware overview, 2017.
- [12] V. Hapla, D. Horák, and M. Pecha, PermonSVM, <http://permon.vsb.cz/permonsvm.htm> (2017).