TECHNICAL SCIENCES
Abbrev.: Techn. Sc., No 15(2), Y 2012

CATEGORIZATION OF SIMILAR OBJECTS USING BAG OF VISUAL WORDS AND *k* – NEAREST NEIGHBOUR CLASSIFIER

Piotr Artiemjew, Przemysław Górecki, Krzysztof Sopyła

Chair of Mathematical Methods in Computer Science
University of Warmia and Mazury Olsztyn

K e y w o r d s: Image categorization, *k* – Nearest Neighbor Classifier, Bag of Visual Words.

A b s t r a c t

Image categorization is one of the fundamental tasks in computer vision, it has wide application in methods of artificial intelligence, robotic vision and many others. There are a lot of difficulties in computer vision to overcome, one of them appears during image recognition and classification. The difficulty arises from an image variance, which may be caused by scaling, rotation, changes in a perspective, illumination levels, or partial occlusions. Due to these reasons, the main task is to represent represent images in such way that would allow recognizing them even if they have been modified.

Bag of Visual Words (BoVW) approach, which allows for describing local characteristic features of images, has recently gained much attention in the computer vision community. In this article we have presented the results of image classification with the use of BoVW and *k* – Nearest Neighbor classifier with different kinds of metrics and similarity measures. Additionally, the results of *k* – NN classification are compared with the ones obtained from a Support Vector Machine classifier.

Introduction

One of the most popular supervised classification methods is based on the searching *k* – Nearest Neighbors (*k* – *NN*) objects using a fixed similarity measure or metric. In the classification by means of *k* – *NN* method, the main problem is to identify the best method for computing a similarity between objects, and to find an optimal value of neighbors *k*. It is necessary to identify a measure which works best, but it is obvious that for different data the best method could differ.

In this article we have investigated the problem of selecting a proper measure and *k* parameter, in the domain of images represented by means of Bag of Visual Words (BoVW).

Our methodology involves using SIFT (Scale-Invariant Feature Transform) (LOWE 2004) feature extractor to obtain the collection of keypoints from the considered images. Successively, k-means clustering method is used for quantizing the keypoints into visual words (dictionary construction), which allows us to represent the images by means of the frequencies of visual words which are present in the image.

In the classification process we use different kinds of metrics and similarity measures. We use the Chisquare metric, Euclidean, Canberra, Manhattan, Normalized, Chebyshev distance and modifications of similarity measures like Cosine measure or the Pearson product-moment correlation coefficient (see DEZA E., DEZA M. 2009). Our results are evaluated by Leave One Out (LOO) method and compared with the results of our recent experiments obtained by applying different kernel functions in Support Vector Machine classifier (GÓRECKI et al. 2012c).

As concerning image representation, BoVW is employed, which is well known in generic visual categorization (CSURKA et al. 2004, DESELAERS et al. 2008, HERBRICH et al. 2000) and subcategorization (GÓRECKI et al. 2012b). In particular, an image is represented using frequencies of distinctive image patches (visual words), obtained in a two-step process. In the first step, a keypoint detector is employed to identify local interest points in a dataset of different images. Successively, the keypoints are quantized into visual words, so that one visual word corresponds to a number of visually similar keypoints. In most cases, K-means clustering is used to carry out the quantization, so that each centroid represents a visual word, and a set of visual words is called a "visual dictionary". By counting visual words in the particular image, a feature vector encoding frequencies of visual words is obtained. Given the feature vector, an image can be further classified into a predefined set of categories using supervised machine learning algorithm, such as k-NN or SVM.

Methodology

There are two issues of BoVW image categorization. The first one is the choice of keypoint detector/descriptor. There were many descriptors proposed in the literature, such as SIFT (LEWIS 1998), SURF (BAY et al. 2006), and more recently BRISK (LEUTENEGGER et al. 2011) and FREAK (ALAHI et al. 2012). Their common feature is robustness to changes in image rotation, scale, perspective and illumination. A comprehensive survey of keypoint detectors can be found in (MAK et al. 2008, THORSTEN et al. 1998). Another important

aspect is the choice of the classifier. In this paper SIFT detector and k – NN classifier were chosen.

Our process of image categorization consists of typical steps (CSURKA et al. 2004):

1. Identification of image keypoints – SIFT keypoints were detected for all images in the dataset.

2. Building the visual dictionary – all keypoints identified in the previous step were clustered into K visual words using – K means algorithm. During the experimental session we use different dictionaries.

3. Building the image representation – for each image, the keypoints are mapped to the most similar visual words and then the image feature vector $v = (v_1,...,v_K)$ is obtained, where v_i encodes the frequency of the i-th visual word.

4. Classification – we use the k – NN classifier, and LOO method to evaluate the effectiveness of the classification, the details are in the classification section.

Data

In our experiments we have clustered the keypoints into different numbers of visual words. In any case, the empty clusters are discarded, therefore the number of obtained visual words (attributes) could be smaller than number of clusters considered originally. Our datasets contain the following number of conditional attributes (visual words): 50, 100, 250, 499, 983, 2442, 4736, where the original number of considered visual words are 50, 100, 250, 500, 1000, 2500, 5000.

Fig. 1. An exemplary shoes from five distinctive classes of examined dataset

Fig. 2. An exemplary set of key points

Classification

In this article we use a classic way of classification based on k – NN methodology. We search for neighbors of the test objects in the whole dataset, and the major class is assigned to the test object, where ties are resolved hierarchically. Having obtained all training objects $\{y_i\}$, we classified the test object x in the following way:

(i) We have computed the distance between objects based on the chosen similarity measure or metric, that is $g(x, y_i)$, where g is metric (d) or similarity measure (p).

(ii) For the fixed test object, we have chosen k nearest training objects.

(iii) The most numerous class assigns the decision to the test object. In the case of draws, we have chosen last conflicted class.

Similarity measures and metrics

It is really important to identify the metric or similarity measure which works best for the considered data. For our k – NN classifier we get distance between objects according to the following functions:

The first one $d : X\, x\, X \rightarrow [0, \infty)$ fulfills conditions,

$$(i)\ d\ (x, y) = 0 \Leftrightarrow x = y$$

$$(ii)\ d\ (x, y) = d\ (y, x)$$

$$(iii)\ d\ (x, y) \leq d\ (x, z) + d\ (z, y)$$

which define a metric, and the second one $p : X \, x \, X \rightarrow [0, 1]$

$$(i) \; p(x, y) = 1 \Leftrightarrow x = y$$

$$(ii) \; p(x, y) = p(y, x)$$

$$(iii) \; p(x, y) \in [0, 1]$$

is a similarity measure. The Cosine measure applied in this article gives the values of similarity from the range [-1,1], which is an exception of the above definition.

One of the most popular is Euclidean metric (DEZA E., DEZA M. 2009), defined in the following way,

$$d \; (x, y) = \sqrt{(\Sigma(x_i - y_i)^2)}$$

The Cosine measure (DEZA E., DEZA M. 2009) works as follows,

$$p \; (x, y) = \Sigma_{i=1}^{n} \left(\frac{x \circ y}{||x|| * ||y||} \right)$$

Where the scalar product is defined as,

$$x \circ y = \Sigma \, x_i * y_i$$

Length of vectors is the following

$$||x|| = \sqrt{\Sigma x_i^2}; \;\; ||y|| = \sqrt{\Sigma y_i^2}$$

One of the simplest is Manhattan metric (DEZA E., DEZA M. 2009) defined below,

$$d \; (x, y) = \Sigma \, |x_i - y_i|$$

The normalized distance between objects based on division by the sum of attribute values is called Canberra metric (DEZA E., DEZA M. 2009),

$$d \; (x, y) = \Sigma \, \frac{|x_i - y_i|}{x_i + y_i}$$

And modification of Canberra metric is the Normalized metric normalized by the maximal and minimal values of attributes in their domains,

$$d\ (x,\ y)\ =\ \Sigma\ \frac{|x_i\ -\ y_i\ |}{\max_i\ +\ \min_i}$$

An interesting metric commonly used as a kernel of Support Vector Machine is Chisquare metric [6] defined in the following way,

$$d\ (x,\ y)\ =\ \Sigma\ \left(\frac{(x_i\ -\ y_i)^2}{x_i\ +\ y_i}\right)$$

Another metric, which determines the distance between objects as the maximal distance between attributes of objects, is the Chebyshev distance (DEZA E., DEZA M. 2009),

$$d\ (x,\ y)\ =\ \max\ (|x_i\ -\ y_i|),\ i\ =\ 1,2,...,n$$

The Pearson product-moment correlation coefficient, which measures the linear correlation of objects, is defined as follows,

$$p\ (x,\ y)\ =\ \ |r_{x,\ y}|$$

$$r_{x,\ y}\ =\ \frac{\Sigma\ (x_i\ -\ \bar{x})(y_i\ -\ \bar{y})}{\sqrt{\Sigma(x_i\ -\ \bar{x})^2}\ \sqrt{\Sigma(y_i\ -\ \bar{y})^2}}$$

$$\bar{x}\ =\ \frac{1}{n}\ \Sigma\ x_i,\ \bar{y}\ =\ \frac{1}{n}\ \Sigma\ y_i$$

Data preprocessing

The scalar length of feature vectors extracted from the dataset could disturb classification in case of data with a large number of visual words. For this reason we have normalized feature vectors by scaling their length to unit one, which is done by dividing all object's attributes by the scalar length of objects. It means that for all $x \in U$, and for all $a \in A$, we perform the following operation:

$$a\ (x)\ =\ \frac{a\ (x)}{||x\ ||},\ \text{where}\ \ ||x\ ||\ =\ \sqrt{\Sigma\ x_1^2}$$

Error evaluation

In this article we have evaluated the classification quality by using the standard Leave One Out method, where for n-objects of decision system we create n tests, where each time a different object is treated as test set and the rest of objects are the training set. The effectiveness of Leave One Out method was investigated among others in (MOLINARO et al. 2005). The LOO method is almost unbiased classification error estimator, hence the evaluation of classification quality is close to the real result.

Results of Experiments

The main goal of our experimental session was to identify the best similarity measure or metric and classification parameters for the prepared dataset with use of global k – NN method. Our dataset consists of 200 objects (images), which represent five categories of shoes, the cardinalities of decision classes are the following, 59, 20, 34, 29, and 58 images respectively. The exemplary images of each class and exemplary key points of selected image are in the Figure 1. and Figure 2.

After data normalization, the classification results for Cosine measure and Euclidean metric are the same, because distance between objects in Cosine metric is reduced to the computation of a scalar product of objects, that is equal to Euclidean distance between objects. The Cosine measure gives the same result before and after normalization, because the applied normalization is an internal part of this measure.

In the Table 1 we can see that the best results of classification for normalized and non-normalized data and considered dictionary sizes. For a smaller number of visual words in the range of 50–100 the classification

Table 1

Leave One Out; The result of classification for the best parameter k and a measure of distance between objects, before and after normalization; chi = Chisquare metric, euk = Euclidean metric, cos = Cosine measure, pea = Pearson product-moment correlation coefficient

	50		100		250		500	1000	2500		5000
Before norm	0.920		0.925		0.895		0.910	0.930	0.955		0.965
Measure	chi		euk		pea	euk	pea	pea	pea		pea
k	1		1		1	1	1	1	3		3
After norm	0.890		0.915		0.895		0.910	0.930	0.955		0.965
Measure	chi	man	pea	cos euk	pea	man	pea	pea	pea	cos euk	pea
k	2	4	3	2	1	1	1	1	3	1	3

is better without normalization. The reason is that the number of visual words, not their types, plays a dominant role in distinguishing decision classes. The optimal value of the k parameter is in the set {1, 2, 3}, the most effective measure turns out to be the Pearson product-moment correlation coefficient. The normalization does not have any influence on the Pearson product-moment correlation coefficient, since linear correlation between the objects is maintained, so the best result for a higher number of visual words is the same regardless of a normalization method. What is interesting, the best metric for 50 visual words, before and after normalization, is the Chisquare metric and additionally the Manhattan metric after normalization. For 100 words, the Euclidean metric is the best, (both before and after normalization), and Pearson's measure performs equally well if normalization is applied. For a higher number of words Pearson's measure is the best, but in a few cases the Euclidean and Manhattan metric have equal accuracy to Pearson's measure.

In the Table 2 we have shown the best results and parameters for all metrics chosen for the data with a different number of visual words. It turns out that we achieve the best results before normalization for the Chisquare,

Table 2
Leave One Out; The best result for all metrics and the parameter k before and after normalization

Metric	Before norm	No.of.visual.words		After norm	No.of.visual.words	k
Pearson	**0.965**	5000	3	**0.965**	5000	3
Chisquare	**0.920**	50	1	0.895	100	2
Manhattan	0.900	50	2	**0.910**	100	1
Cosine	**0.960**	5000	3	**0.960**	5000	3
Euclidean	0.925	100	1	**0.960**	5000	3
Normalized	**0.890**	50	1,2,4	0.875	50	1,3
					100	1
Canberra	**0.895**	50	4	0.865	50	3
Chebyshev	**0.875**	100	2	0.830	100	2

Normalized, Canberra, and Chebyszev metric, and after normalization tthe Manhattan and Euclidean metrics work better. As we mentioned before, Pearson's and Cosine measures work equally well before and after normalization.

Considering the best result of classification, we made an assumption that best parameter k has values 1, 2 or 3, and for this reason, in the Figures 3, 4, 5 and 6, we present the average of the classification results for these three parameters. Particularly, in the Figure 3 and 4 we have separate results for all dictionaries, and metrics before and after normalization. These generalized

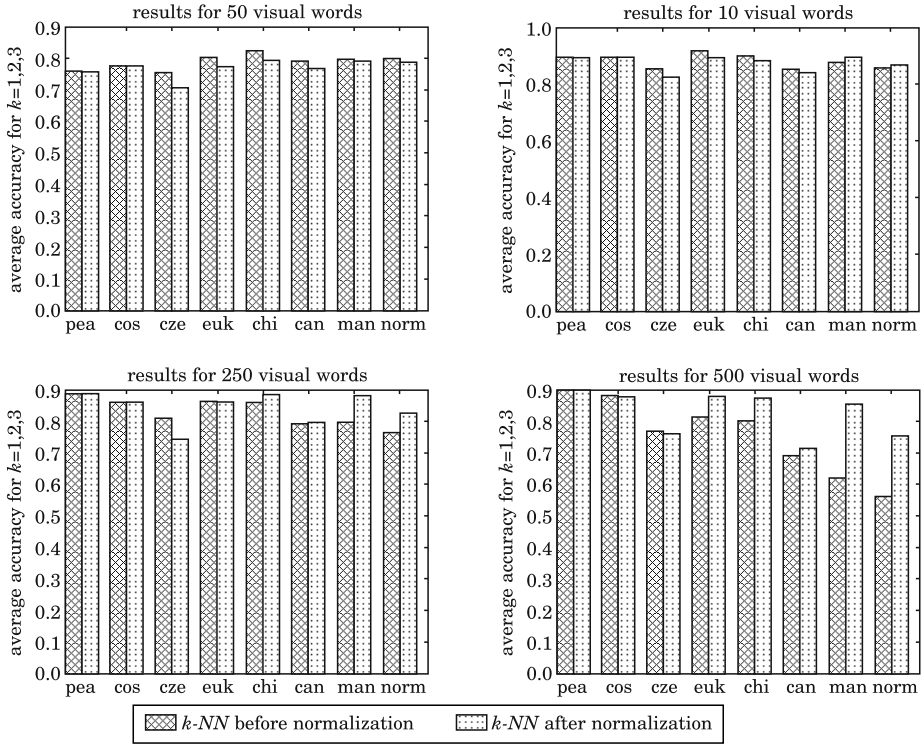

Fig. 3 Leave One Out; k – *NN*; Average result of accuracy for k=1,2,3 and 50–500 visual words

results lead us to the conclusion, that for 50 and 100 words, all metrics work better before normalization, except of Manhattan and Normalized metric in case of 100 words. For 250–5000 words, we obtain the best results for all metrics and measures after normalization, except for Chebyszev metric for 250 words.

In the Figure 5 and 6, we have results for non-normalized data and normalized data respectively. In the plots, we have the results for all metrics vs all dictionaries (the data from the Figure 3 and 4 shown in the different way). In the Figure 7, we have exemplary detail wesult for Pearson's measure with data after normalization. The conclusion is that for 50, 100 and 250 words all the metrics work really steadily before and after normalization. Pearson's and Cosine measure work optimally for all the dictionaries. In case of the Czebyszev metric after normalization, the result of classification is more consistent for all the dictionaries. The Euclidean metric works better after normalization. The result before and after normalization for the rest of the metrics is comparable.

In addition to our results we have compared results of a SVM classifier (CHAPELLE et al. 1997, FAN et al. 2005) with different kernel functions

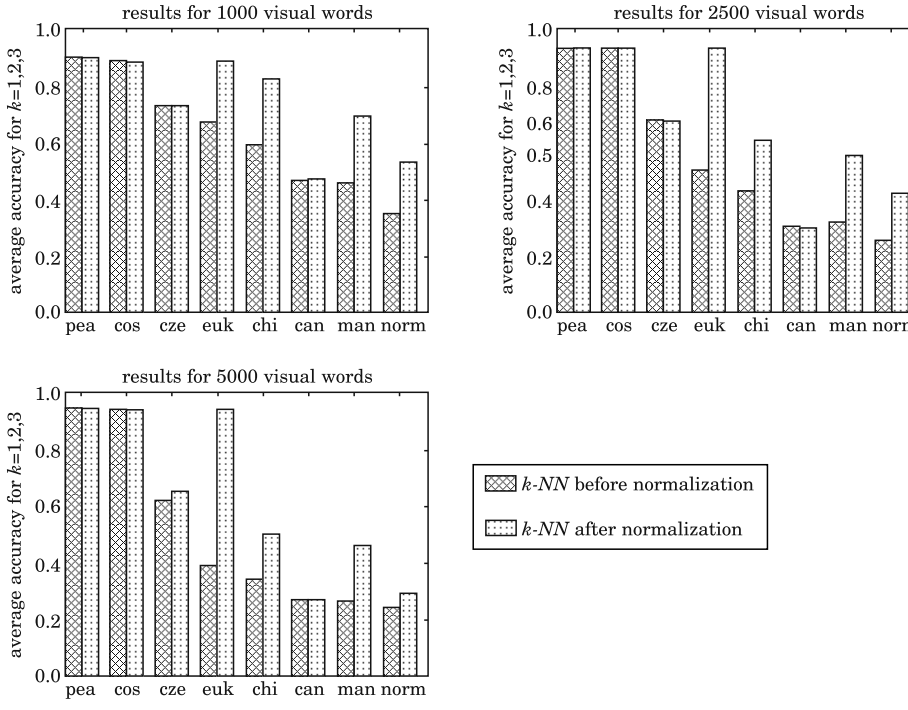

Fig. 4. Leave One Out; *k – NN*; Average result of accuracy for $k=1,2,3$ and 1000–5000 visual words

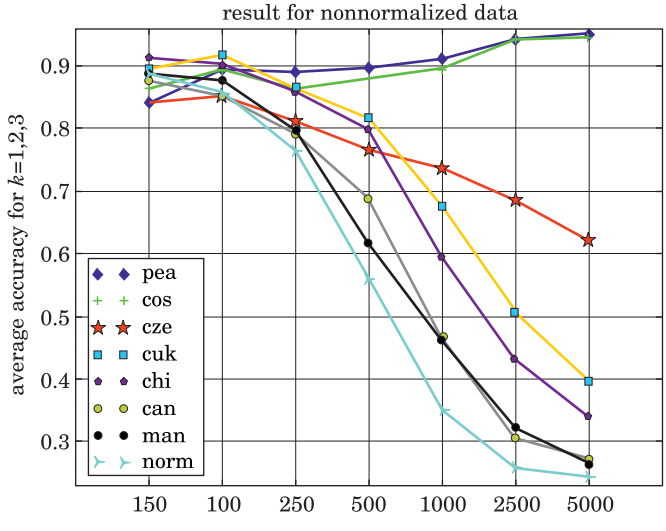

Fig. 5. Leave One Out; *k – NN*; For data before normalization; Average result of accuracy for $k=1,2,3$ and 50–5000 visual words

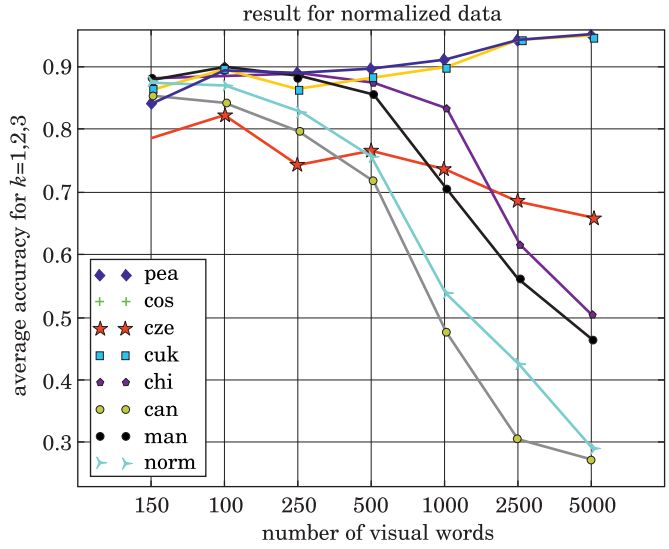

Fig. 6. Leave One Out; k – *NN*; For data after normalization; Average result of accuracy for k=1,2,3 and 50–5000 visual words

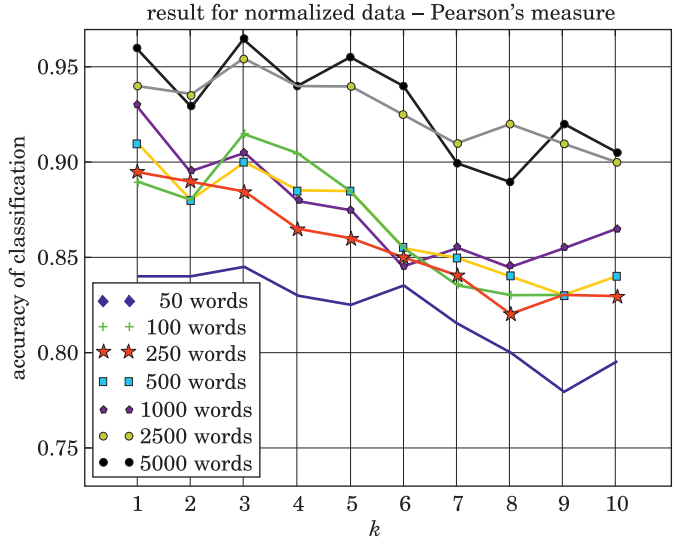

Fig. 7. Leave One Out; k – *NN*; For Pearson's measure with data after normalization; The result of accuracy for k=1,2,...,10 and 50–5000 visual words

(see Tab. 3) (GÓRECKI et al. 2012c), and global k – NN classifier with use of different similarity measures and metrics. We can summarize that for a number of visual words in the range of 50, 100, 2500, 5000, the results are comparable with the k – NN method and even better for a smaller number of words in the range of 50 and 100, but in range of 250, 500, 1000 visual words, the SVM classifier wins by about four percent of classification accuracy.

Table 3
Cross Validation 5; Overall accuracy of Support Vector Machine kernels in relation to visual dictionary size

Kernel	Number of visual words						
	50	100	250	500	1000	2500	5000
Linear	0.870	0.870	0.925	0.940	0.945	0.940	0.940
Chi^2	0.880	0.885	**0.930**	0.940	**0.960**	0.955	0.955
Histogram	0.855	**0.895**	**0.930**	0.935	0.940	**0.960**	**0.965**
RBF	**0.905**	0.890	**0.930**	**0.945**	0.945	0.940	0.940
Cauchy	0.900	0.890	0.900	0.930	0.915	**0.960**	0.905

Conclusions

The results of these experiments show an interesting dependence between the quality of classification by means of a global k– NN method, the number of visual words, and the normalization method. It has turned out that for smaller number of visual words in the range of 50, 100, we get better result for non-normalized data. In case of 50 visual words the best is the Chisquare metric, for 100 visual words the best is the Euclidean metric, and starting from 250 words the best similarity measure for the considered data turn out to be the Pearson product-moment correlation coefficient and Cosine measure. For a higher number of visual words in the range of 250-5000, we achieve better result after normalization. The optimal parameter k for global k – NN classifier and considered dataset is in the set $\{1, 2, 3\}$.

In the future we are planning to check the effectiveness of other methods of classification based on other publicly available datasets. Another goal is to apply other keypoint detectors in our research, so that their effectiveness can be compared with our current results.

Acknowledgements

The research has been supported by grant N N516 480940 from the National Science Center of Republic of Poland and grant 1309-802 from Ministry of Science and Higher Education of the Republic of Poland.

Translated by MATT PURLAND

Accepted for print 12.11.2012

References

ALAHI A., ORTIZ R. VANDERGHEYNST P. 2012. *FREAK: Fast Retina Keypoint*. IEEE Conference on Computer Vision and Pattern Recognition, Rhode Island, Providence, USA,

BAY H., TUYTELAARS T., VAN GOOL L. 2006. *Surf: Speeded up robust features*. ECCV, p. 404–417.

CHAPELLE O., HAFFNER P., VAPNIK V.N. 1999. *Support vector machines for histogram-based image classification*. Neural Networks, IEEE Transactions on, 10(5): 1055–1064.

CSURKA G., DANCE CH.R., FAN L., WILLAMOWSKI J., BRAY C. 2004. *Visual categorization with bags of keypoints*. Workshop on Statistical Learning in Computer Vision, ECCV, p. 1–22.

DESELAERS T., PIMENIDIS L., NEY H. 2008. *Bag-of-visual-words models for adult image classification and filtering*. ICPR, p. 1–4,

DEZA E., DEZA M. 2009. *Encyclopedia of Distances*, Springer.

FAN R., CHEN P., LIN Ch. 2005. *Working set selection using the second order information for training svm*. Journal of machine learning research, 6: 1889–1918.

GÓRECKI P., ARTIEMJEW P., DROZDA P., SOPYŁA K. 2012a. *Categorization of Similar Objects using Bag of Visual Words and Support Vector Machines*. Fourth International Conference on Agents and Artificial Intelligence. IEEE.

GÓRECKI P., SOPYŁA K., DROZDA P. 2012b. *Ranking by k-means voting algorithm for similar image retrieval*. In: ICAISC 1(7267): 509–517, of Lecture Notes in Computer Science, eds. L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L.A. Zadeh, J.M. Zurada, Springer.

GÓRECKI P., SOPYŁA K., DROZDA P. 2012c. *Different SVM Kernels for Bag of Visual Words*. In: International Congress of Young Scientist, SMI'2012, Springer.

HERBRICH R., GRAEPEL T., OBERMAYER K. 2000. *Large margin rank boundaries for ordinal regression*. MIT Press, Cambridge, MA.

LEUTENEGGER S., CHLI M., SIEGWART R. 2011. *BRISK: Binary Robust Invariant Scalable Keypoints*. Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2548–2555.

LEWIS D.D. 1998. *Naive (bayes) at forty: The independence assumption in information retrieval*. Springer Verlag, p. 4–15.

LOWE D.G. 2004. *Distinctive image features from scale-invariant keypoints*. Int. J. Comput. Vision, 60: 91–110.

MAK M.W., GUO J., KUNG S.-Y. 2008. *Pairprosvm: Protein subcellular localization based on local pairwise profile alignment and svm*. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 5(3): 416–422.

MIKOLAJCZYK K., LEIBE B., SCHIELE B. 2005. *Local Features for Object Class Recognition*. Tenth IEEE International Conference on Computer Vision (ICCV'05), 1: 1792–1799.

MOLINARO A.M., SIMON R., PFEIFFER R.M. 2005. *Prediction error estimation. a comparison of resampling methods*, Bioinformatics, 21(15): 3301–3307.

NISTR D., STEWNIUS H. 2006. *Scalable recognition with a vocabulary tree*. In IN CVPR, p. 2161–2168.

THORSTEN J. 1998. *Text categorization with support vector machines: learning with many relevant features*. Proceedings of ECML-98, 10th European Conference on Machine Learning, 1398: 137–142. Eds. C. N'edellec, C. Rouveirol. Springer Verlag, Heidelberg, DE.

TUYTELAARS T., MIKOLAJCZYK K. 2008. *Local invariant feature detectors: a survey*. Found. Trends. Comput. Graph. Vis., 3: 177–280.