A Comparison Of Suitable Object Recognition Methods For Mobile Voiced Visual Assistant

Vassili Kovalev, Igor Safonov
Biomedical Image Analysis Group
United Institute of Informatics Problems, Minsk, Belarus
vassili.kovalev@gmail.com, safonov@tut.by

Abstract

The objective of this work is to find an optimal object recognition method for Mobile Voiced Visual Assistant (MVVA). MVVA is under development in the Biomedical Image Analysis Group of the Institute. It's aimed to assist visually-impaired people in recognition and audio interpretation of surrounding scenes and objects in real time. In this paper we consider some applicable recognition methods based on color co–occurrence matrices.

Keywords: *object recognition, co–occurrence matrix, SVM, PCA, Mobile Voiced Visual Assistant.*

1. INTRODUCTION

The subject of our study is to find an optimal recognition method for future use in developing of Mobile Voiced Visual Assistant (MVVA). MVVA is aimed to assist visually-impaired people in recognition and audio interpretation of surrounding scenes and objects in real time (indoors or outdoors).

Let us remark that the use of special devices is not planned here. It is anticipated that the software will work as an application on any portable device with a camera. Such devices as ultra-mobile PC, netbooks and mobile phones are implied. These are resulted in some specific requirements to the software.

First of all, the application should work with images of low or medium quality.

The second requirement is the recognition algorithm's stability. The algorithm must provide good results under different environment. Along with this, in some algorithms the correct choice of parameters might be crucial for obtaining good results. Therefore an important step of the analysis is to estimate these parameters. In order to achieve this, we have developed and implemented a plan of experimental research the results of which are reported with this study.

The structure of the article is as follows. Section 2 reviews input data, hardware and software tools. In the Section 3 research stages themselves are described. The final section summarizes the results of this experimental investigation.

2. INITIAL DATA, THE HARDWARE AND SOFTWARE TOOLS

As a development environment we have used R, a system for statistical computing and graphics [1]. It consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in form of R script files. The add-on package e1071 was employed here too.

Input data were 24-bit images of 320×240 pixels in size (Fig.1). All the images came from Logitech QuickCam Pro 9000 web camera. We consider 73 images. A variety of the data can be grouped into certain classes. Eight classes are considered in framework of this study that conditionally categorized into: «bag», «chair», «cup», «doorway», «flower», «phone», «wardrobe» and «window».

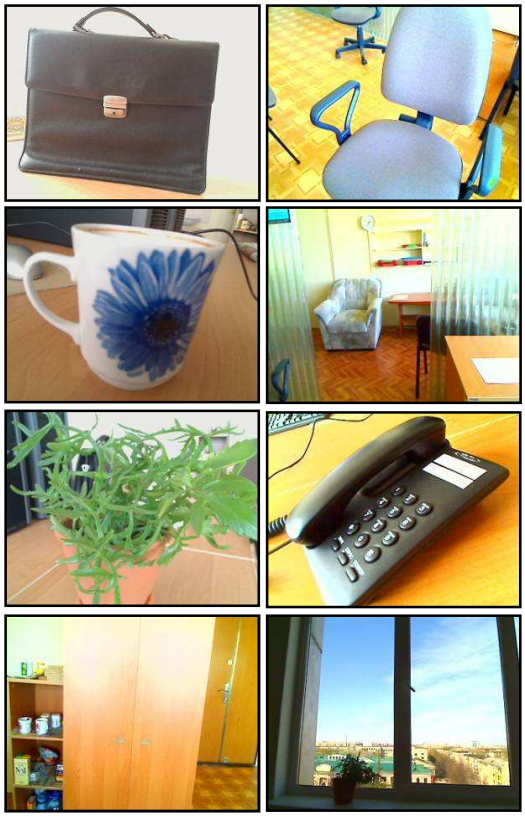

Figure 1: Example of input images.

Pictures of each class are split into a training and test sets. If *N* is a number of images from some class, then training set contains *N/2* elements, and test set contains the rest of *N-N/2* elements. We utilize the re-sampling techniques to generate test and training sets which are the random subsets of the original image collection of each type.

Image descriptors were created by way of a vectorization of color co-occurrence matrices [2,3]. It should be noted that for 24-bit color RGB images the corresponding color co-occurrence matrix can be very large. We reduce color space from 24-bit down to 8-bit using the common quantization scheme known as “3-3-2”.

Thus, all further manipulations were carried out in the system for statistical computing and graphics R with the whole set of 73 image files. The files contain pre-calculated co-occurrence matrices of 8-bit images. At every stage of experimentation the

resultant estimates reported in this paper are the mean values computed over 300 iterations for reliability. On each iteration step, the original data set is randomly re-sampled with replacement to generate independent replications.

3. RESEARCH STAGES

We will subsequently pass through the four stages of the assessment changing the conditions of experiments on:

- Selection of optimal parameters for a support vector machine with RBF-kernel;
- Changing the type of co-occurrence matrices and their control parameters;
- Selecting the way of pre-processing of co-occurrence matrices;
- Recognition with weighted distances.
- Most of these steps are very comprehensive computational tasks.

To reduce the feature space and to decrease computing cost, a simple descriptors' preprocessing was used. Positions, where the elements in each descriptor are equal to zero, were excluded from further analysis. Such a technique allows to significantly reduce the dimensionality of the data. The descriptors can reach 65,536 elements at most. After removal of zero elements, the descriptors contain no more than 5000 elements.

Detailed information on each experiment is given below.

3.1 Selection of optimal parameters for a support vector machine with RBF-kernel

Support vector machines (SVMs) [4] are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an *n*-dimensional space, an SVM will construct a separating hyperplane in that space, one which maximizes the margin between the two data sets.

For classification task, we use C-classification with the RBF kernel

$$K(x, y) = e^{-\gamma\|x-y\|^2}, \gamma > 0.$$

In [5] authors suggested that in general RBF is a reasonable choice. The RBF kernel nonlinearly maps samples into a higher dimensional space, so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. In addition to that, there are only two parameters while using RBF kernels: C and γ.

The calculations were performed with grid-search on C and γ using cross-validation. Since doing a complete grid-search may still be time-consuming, we used a coarse grid first. After identifying a better region on the grid, a finer grid search on that region was conducted.

For each data set we first use a coarse grid on the initial interval $C=2^{-5}, 2^{0},\ldots,2^{15}$ and $\gamma=2^{-50}, 2^{-48},\ldots,2^{10}$ (Fig.2) . After finding the best (C, γ) we conduct a finer grid search on the neighborhood of the point (Fig.3).

As illustrated in Fig.3, a better rate 84,89% at (2^{3},2^{-33}) is gained.

3.2 Changing the type of co-occurrence matrices and their parameters

The image recognition experiments on the first stage are carried out for the co–occurrence of colors of neighboring pixels only. It is reasonable to suppose that if we add to the existing descriptors information about the neighboring pixels that are located at a greater distance from each other, the recognition quality could be improved.

Another way of looking to further improvements of descriptors is to consider so–called concatenated descriptor, which can be obtained by concatenation of co-occurrence descriptors of original images and their pyramided versions (i.e., with every second pixel row and column eliminated).

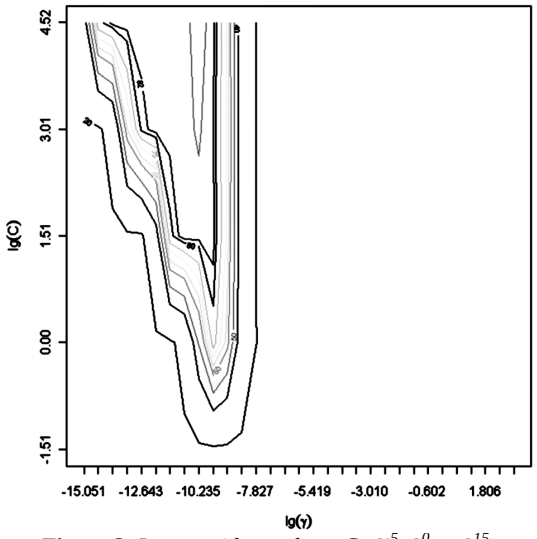

Figure 2: Loose grid-search on $C=2^{-5}, 2^{0},\ldots,2^{15}$ and $\gamma=2^{-50}, 2^{-48},\ldots,2^{10}$.

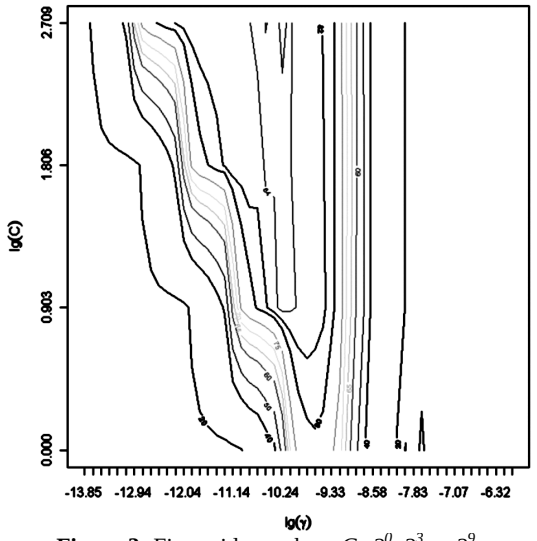

Figure 3: Fine grid-search on $C=2^{0}, 2^{3},\ldots,2^{9}$ and $\gamma=2^{-46}, 2^{-45.5},\ldots,2^{-20}$.

The decimation consists in removal of every second row and every second column from original image.

If $\|U_{ij}\|, i=\overline{1,n},\ j=\overline{1,m}$ is co–occurrence matrix of an original image, then a vector $u_{image}=(u_{11},u_{12},\ldots,u_{1m},u_{21},\ldots,u_{2m},\ldots,u_{n1},\ldots,u_{nm})$ is its descriptor. In a similar manner, if $\|V_{ij}\|, i=\overline{1,k},\ j=\overline{1,l}$ is co–occurrence matrix of decimated image, then as its descriptor we

have a vector $v_{image} = (v_{11}, v_{12}, \ldots, v_{1l}, v_{21}, \ldots, v_{2l}, \ldots, v_{k1}, \ldots, v_{kl})$, for $k = \left\lceil \frac{n}{2} \right\rceil$, $l = \left\lceil \frac{m}{2} \right\rceil$.

Further calculations are performed on normalized vectors of the form $w = u_{image} \cup v_{image}$. The normalization is achieved by division of each vectors' part into total pairs of first (305522 pairs) and second (75962 pairs) images respectively

The descriptors, thus defined, second data set presents. Such an experiment have been performed in perfect analogy to the previous stage. In that case the greatest possible recognition efficiency amount to 84.67%.

3.3 Selecting the way of pre-processing of co-occurrence matrices

Partially pre–processing of co–occurrence matrices was occurred in each of the previous experiments. The pre–processing consists in reduction of common zero elements. This allows to increase processing speed.

But in addition, any image from digital camera has a noise. It does not carry any useful information and it must be eliminated from further analysis.

For this purpose we applied thresholding. The threshold was defined as a percentage of the total number of occurrence pairs in the picture. The next threshold values were defined: 0,1% 0,2% 0,5%. For the source images with a total of 305,522 pairs after thresholding we can get 306, 611 and 1528 pairs respectively. For decimated images with a total number of 75,962 pairs we can get 76, 152 and 380 pairs.

Of course, the number of descriptors' elements significantly reduced after the operation.

In Table 1 elements count are listed after thresholding of descriptors.

Table 1: Elements count in descriptors.

	Initial number	After thresholding		
		0,1%	0,2%	0,5%
Basic descriptor	256^2	563	423	245
Descriptor of pyramided image	256^2	323	229	155
Concatenated descriptor	2×256^2	886	652	400

Further experiments were conducted to determine the quality of recognition for the filtered descriptors (Figure 4).

It can be seen that recognition quality is best achieved on descriptors of original image with threshold 0,5% (88,69%). That corresponds to the analysis of 245 features (look at Table 1).

With combining the two types of descriptors appreciable recognition quality was not gained. When it is considered that concatenated descriptor construction and its further use requires additional resources (processing time and additional memory to store), there is little point in using of concatenated descriptors.

Another possible approach to the pre–processing co–occurrence matrices is to use principal component analysis (PCA) [6]. PCA is a useful statistical technique that has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension.

Experiments have been performed to determine quality of recognition after rearrangement of co-occurrence matrices by PCA. For three sets of data, was taken number of principal components with cumulative proportion of 0,5, 0,6, 0,7, 0,8 and 0,9. Obtained principal components were used as features to perform recognition with the help of SVM.

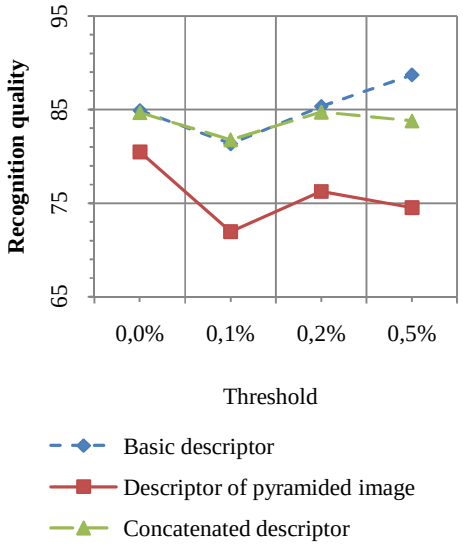

Fig.4: Recognition quality before and after thresholding.

Considering Table 2, it can be concluded that the best recognition quality (86,51%) is achieved for the descriptors of decimated image and the cumulative proportion equal to 0,5. It according to analysis of 12 principal components only.

Table 2: Relationships between cumulative proportion of PCA variance and recognition quality, %.

	0,5	0,6	0,7	0,8	0,9
Basic descriptor	79,52	83,31	81,73	74,05	63,62
Descriptor of pyramided image	86,51	77,17	73,58	66,44	54,78
Concatenated descriptor	83,24	83,97	81,37	74,01	59,55

3.4 Recognition with weighted distances

The next stage of the study is recognition with minimal weighted distances to a class object.

For each recognition object weighted distance to all images of each class was calculated. Minimum distance was a criterion of belonging of some class. To calculate the distance between the descriptors, four different distances were taken: Euclidean distance, Chebyshev distance, Manhattan distance and Canberra distance.

For two vectors, $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ the distances are defined as follows:

- Euclidean distance: $d_{euclidean}(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$,

- Chebyshev distance: $d_{chebyshev}(X,Y) = \max_i |x_i - y_i|$,
- Manhattan distance: $d_{manhattn}(X,Y) = \sum_{i=1}^{n} |x_i - y_i|$,
- Canberra distance: $d_{canberra}(X,Y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{x_i + y_i}$.

The results of the experiment are summarized in Table 3.

You can see that the most effective recognition was 67,64%. The value is achieved on descriptors of original image with Canberra distance. With the use of Canberra distance we achieve the best recognition results on other types of descriptors.

Table 3: Recognition quality for different types of descriptors and different distances.

	Euclidean			**Chebyshev**		
Basic descriptor	59,15			55,94		
Descriptor of pyramided image	57,04			45,85		
Concatenated descriptor	59,64			49,53		
Basic descriptor with thresholding (0,1%; 0,2%; 0,5%)	59,89	59,48	59,44	56,35	56,44	56,19
Descriptor of pyramided image with thresholding	56,55	56,16	57,03	45,63	45,11	45,93
Concatenated descriptor with thresholding	59,36	60,04	60,24	48,72	49,81	49,58
	Manhattan			**Canberra**		
Basic descriptor	63,14			67,64		
Descriptor of pyramided image	55,43			62,43		
Concatenated descriptor	58,94			65,20		
Basic descriptor with thresholding (0,1%; 0,2%; 0,5%)	63,01	62,73	62,23	66,71	66,57	66,25
Descriptor of pyramided image with thresholding	55,05	54,77	55,41	61,22	60,78	61,26
Concatenated descriptor with thresholding	58,23	58,72	58,31	63,86	64,07	63,79

4. CONCLUSION

The paper examined several approaches for recognition of 24–bit color images sized 320×240 pixels, acquired with the help of the web camera. An investigation on the influence of different factors on the object recognition quality has been accomplished. Specifically, the following factors were studied:

- Selection of optimal parameters for a support vector machine with RBF-kernel;
- Changing the type of co-occurrence matrices and their control parameters;
- thresholding of descriptors
- principal component analysis
- Selecting the way of pre-processing of co-occurrence matrices;
- Recognition with weighted distances.

Based on the experiments, it may be deduced that the best recognition quality (88,69%) is achieved using SVM classifier with basic color co-occurrence descriptors of original images and with the matrix element selection threshold equal to 0,5% of the total number of pixel pairs.

These results will be considered on the implementation of Mobile Voiced Visual Assistant aimed to a practical use.

Acknowledgments. This work has been partially supported by the ISTC grant B-1682.

5. REFERENCES

[1] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008, http://www.r-project.org/

[2] Kovalev V. and Volmer S. *Color Co-Occurrence Descriptors for Querying-by-Example*, Int. Conf. on Multimedia Modelling, Oct. 12-15, Lausanne, Switzerland, IEEE Comp. Society Press, pp. 32-38, 1998.

[3] Kovalev V.A. *Towards image retrieval for eight percent of color-blind men*. Int. Conf. on Pattern Recognition, Cambridge, UK, 23-26 Aug 2004, IEEE Comp Society Press, Vol. 2, pp. 943-946, 2004.

[4] Vapnik V. *The nature of statistical learning theory*. New York, NY: Springer–Verlag, 1995.

[5] C.–W. Hsu, C.–C. Chang, C.–J. Lin. *A practical guide to support vector classification* . Technical report, Department of Computer Science, National Taiwan University. July, 2003.

[6] Jolliffe I. T. *Principal Component Analysis, Series: Springer Series in Statistics*, 2nd ed., Springer, NY, 2002.

About the authors

Vassili Kovalev is a Head of Biomedical Image Analysis Group, United Institute of Informatics Problems. His contact email is vassili.kovalev@gmail.com.

Igor Safonov is a junior researcher of Biomedical Image Analysis Group. His contact email is safonov@tut.by.