Advances and Trends in Geodesy, Cartography and Geoinformatics – Molčíková et al. (Eds)
© 2018 Taylor & Francis Group, London, ISBN 978-1-138-58489-1

Classification of digitized old maps

M. Talich & O. Böhm
Research Institute of Geodesy, Topography and Cartography, Zdiby, Czech Republic

L. Soukup
Institute of Information Theory and Automation of the CAS, Prague, Czech Republic

ABSTRACT: Because of their importance as historical sources, old maps are steadily becoming more interesting to researchers and public users. However, the users are no longer satisfied only by simple digitization and online publication. Users primarily require advanced web tools for more sophisticated work with old maps.

This paper is concerned with classification of digitized old maps in form of raster images. An automatic classification of digital maps is useful tools. This process allows to automatically detect areas with common characteristic, i.e. forests, water surfaces, buildings etc. Technically it is a problem of assigning the image's pixels to one of several classes defined in advance. If the map is georeferenced the classified image can be used to determine the surface areas of the classified regions, or otherwise evaluate their position.

Unfortunately quite substantial difficulties can be expected when attempting to apply these tools. The main cause of these difficulties is varied quality of digitized maps resulting from damage caused to the original maps by time or storage conditions and from varying scanning procedures. Even individual maps from the same map series can differ quite a lot.

The review of the main classification methods with special emphasis on the Bayesian methods of classification is given. An example of this classification and its use is also given. Web application of raster image classification is introduced as well. The web application can classify both individual images and raster data provided via Web Map Services (WMS) with respect to OGC standards (Open Geospatial Consortium). After gathering the data, classification is applied to distinguish separate regions in the image. User can choose between several classification methods and adjust pertinent parameters. Furthermore, several subsequent basic analytical tools are offered. The classification results and registration parameters can be saved for further use.

1 INTRODUCTION

Because of their importance as historical sources, old maps are steadily becoming more interesting to researchers and public users. However, users are no longer satisfied with only a simple digitization and online publications of them. Users primarily require advanced web tools for more sophisticated work with the digitized old maps.

One of the useful tools is automatic classification of old digitized maps. This process allows to automatically detect areas with common characteristic, i.e. forests, water surfaces, buildings etc. From technical point of view is it a problem of assigning the image's pixels to one of several classes defined in advance. If the map is georeferenced the classified image can be used to determine the surface areas of the classified regions, or otherwise evaluate their position.

Unfortunately, quite substantial difficulties can be expected when attempting to apply these tools. The main cause of these difficulties is varied quality of digitized old maps resulting from damage caused to the original maps by time or storage conditions and from varying scanning procedures. Even individual maps from a single map series can differ quite a lot.

The basic prerequisite for processing old maps in this way is to have them scanned and published online in standardized way, because only then the maps will be available easily enough to research, develop and try sufficiently robust and efficient solutions to presented problems. Classification of raster images of the old maps will be further discussed in chapter 2 with special emphasis on the Bayesian methods of classification.

2 CLASSIFICATION OF RASTER IMAGES OF THE DIGITIZED OLD MAPS

2.1 *Problem formulation*

Regions with some characteristic features have to be localized in a digital image of digitized old map. These features could be uniquely derived from the given attributes of pixels, e.g. color of a pixel. The whole image has to be decomposed into disjoint regions and each region has to be attributed by a unique class according to the prevalent characteristic features. The set of classes has to be given in advance. The decomposition task, i.e. classification, results in assignment a specified class to each pixel in the given digital image.

2.2 *The required result*

Result of classification has to be in form of a new image that consists of homogenous disjoint regions of different classes. The regions are distinguishable by class labels or colors that are explained in the associated legend.

2.3 *Review of the main classification methods*

Vast number of different classification methods have been designed during short history of development of computer image processing. Two main groups of classification methods can be recognized: deterministic and statistic. Other distinction between classification methods is based on practical circumstances of solution of the classification problem. When some characteristic features of the classes are available, the classification is called supervised. If no preliminary data about classes are known in advance, unsupervised classification (cluster analysis) has to be performed. Statistical supervised classification, see e.g. (Webb 2003), (Denison, Holmes, Mallick & Smith 2002), presents more powerful tool than the other kinds of classification.

Statistical characteristics of the all admissible classes have to be known at statistical supervised classification. The most common way of gaining characteristic features of classes is to determine training sets in the given image. Training set is a region in the image which well represents certain class. Searching for other regions with similar characteristic features is the task of supervised classification. Principle of statistical supervised classification is based on geometric notion of feature space. Feature space is an Euclidean space of points, whose coordinates are features that characterize each pixel in the image. Typical example of feature space is color space RGB. Each pixel of digital image displays as a point with coordinates that are the features, e.g. color components Red, Green, Blue. Points in feature space create clusters that represent particular classes. Some points of these clusters correspond to pixels that belong to a training set. These points can be labelled by identifier of a corresponding class. With the aid of the labelled points of training sets, other points in feature space have to be labelled to complete the classification. Hence the classification task can be formulated as a rule for labelling pixels displayed in feature space. This rule, called classifier, can be searched for by means of several manners. The most common classifiers are e.g. linear, nearest neighbor or Bayesian classifiers. Bayesian classification that is the most important member of the family of statistical supervised classification will be studied in the sequel.

2.4 *Bayesian classification*

2.4.1 *Input data and assumptions*

A digital image is given where training sets are determined. Certain number of classes is chosen to distinguish regions of different characteristics. Let C be the set of the all classes. Each training set is assigned to a certain class. Each class has to be represented by one training set at least. Furthermore a prior probability $P(C)$ has to be known for each class C in $\mathcal{C}$. The prior probabilities describe general preliminary information about presence of classes in the given image.

2.4.2 *Solution of the problem*

The classification problem is solved by Bayesian classifier in this contribution. The Bayesian classifier stems from Bayes formula (see e.g. (Webb 2003)). This formula enables to compute the probability that a pixel with feature vector F belongs to class C. It is conditional probability $P(C\,|\,F)$. We can estimate opposite conditional probability $P(F\,|\,C)$ for any feature vector F and class C with the aid of the training sets. Expression $P(F\,|\,C)$ stands for probability that a pixel of class C has feature vector F. Under these assumptions for known prior probabilities $P(C)$ the Bayes formula has form:

$$P(C\,|\,F) = \frac{P(F\,|\,C)P(C)}{\sum_{\mathrm{T}\in\mathcal{C}} P(F\,|\,T)P(T)} \tag{1}$$

The last step of the classification procedure comprises assignment of class C to pixel with feature vector F to maximize posterior probability $P(C\,|\,F)$.

Crucial problem resides in computation of probabilities $P(F\,|\,C)$, since it is sensitive to input data in training sets. Three variants of Bayesian classification will be presented to cover most cases of determining training sets.

2.4.3 *Basic variant*

The simplest way of computation probabilities $P(F\,|\,C)$ is based on relative frequency of pixels in the training set. Let us denote n_C the overall number of pixels in training set of class C and $n_{C,F}$ the number of pixels with feature vector F in the same training set. Then the probability $P(F\,|\,C)$ can be approximately estimated by

$$P(F\,|\,C) = \frac{n_{C,F}}{n_C} \tag{2}$$

2.4.4 *Extended variant*

The extended variant is based on assumption, that clusters of the same class are normally distributed. Under this assumption each training set could be extended by adding other pixels with similar features as the original pixels selected by the actual training set in the chosen cluster.

Pixels, whose feature vectors are sufficiently close to the center of the cluster, could be treated as members of the actual class C. Such pixels can extend the actual training set to create a new, extended training set. The extended training set is more representative, but there is some risk that some of its pixels do not belong to the actual class C. If the risk is small (e.g. less than 0.05), it is possible to compute relative frequency (2) with greater values of n_C, $n_{C,F}$. Better estimation of probabilities $P(F\,|\,C)$ could be reached by this way. The problem is in definitions of riskiness and sufficient closeness to the center of cluster.

The distance of additional pixels from the center of cluster is measured by Mahalanobis distance. The limit distance below which the pixels are considered close has to be determined in accordance to the risk of appending wrong pixels.

2.4.5 *Nearest neighbor variant*

This variant is based on assumption of normality of clusters as in the previous variant. Indeed, membership of a pixel into a class is computed as a distance of the pixel from the center of the cluster. The pixel is assigned to the class, whose training set is the nearest to the pixel in question. The metrics for measuring the distance is derived from Mahalanobis distance. The distance between a pixel and a training set of class C is a posterior probability $P(C \mid E_h)$ which is given by Bayes formula in the consequent form.

$$P(P \mid E_h) = \frac{P(E_h \mid C)P(C)}{\sum_{T \in C} P(E_h \mid T)P(T)} \tag{3}$$

Symbol E_h depends on the risk of appending wrong pixels.

3 PRACTICAL ONLINE SOLUTION

Web application for practical solution of Bayesian classification was created. The application, named WACLASS, is available at www.vugtk.cz/ingeocalc/igc/classification/ and works as a client—server application.

The client part of the application supports all the user operations, namely design of classes, definition of training sets and so on. The actual classification runs on the server side of the application. This part of the application was programmed in Python language with the aid of web framework Django and image processing library PIL (Python Image Library). Client side of the application is based on standard web technologies such as HTML, Javascript, and SVG (Scalable Vector Graphics). It means that the application can be used on practically any computer that is connected to Internet with any web browser. More information about this web application with its features, data sources, possible variants of classification, analytical tools and application controls are in (Talich, Böhm & Soukup 2012).

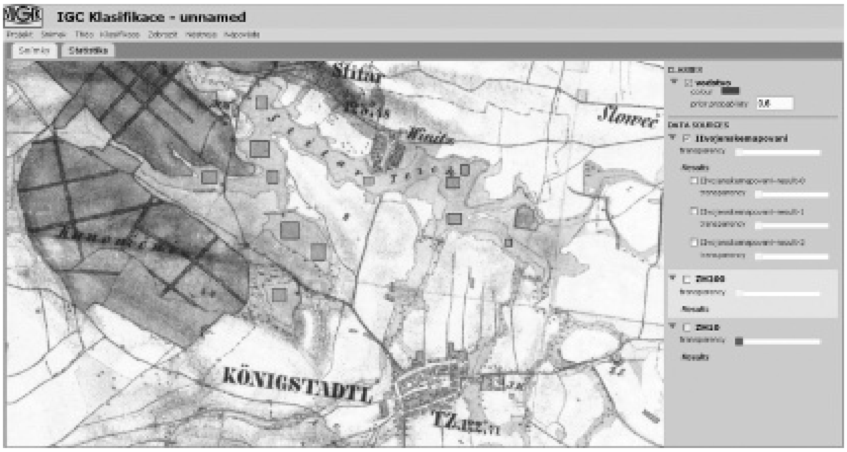
4 PRACTICAL EXAMPLE

This chapter presents an example of use of automatic classification for estimating the surface area of former lake Štítarský in the first half of 19th century. This lake lay near present day Vinice village near Městec Králové town. The lake can be found on II. Military Survey maps, but it no longer exists today in its former size.

By comparing the old maps with contemporary maps it is possible to see that only small remnants of the original water body are left. The original size of the lake can be determined by using web application for raster image classification mentioned above and accessible on the www.vugtk.cz/ingeocalc. This application is part of a knowledge system for decision support based on geodata created in VÚGTK between years 2006 and 2011. The application can display (among others) the maps of II. Military Survey provided as WMS (OGC 2006) and use it as data source.

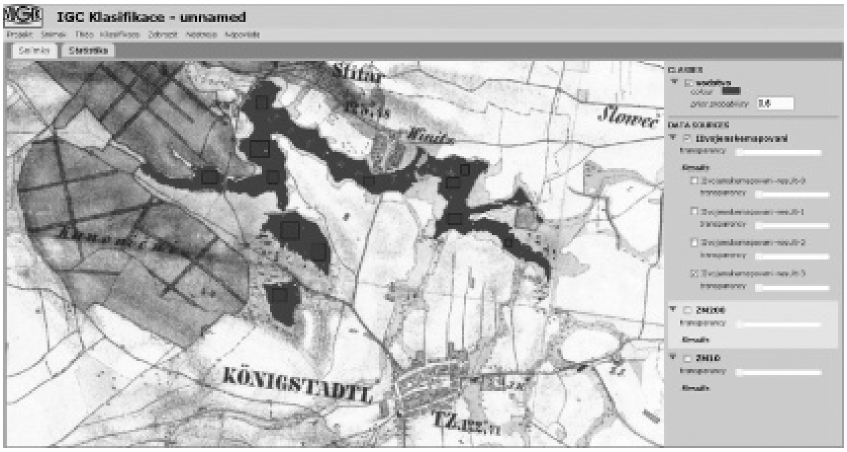
When the map is displayed it is necessary to select training areas—representative samples of the areas of interest. In this case, the area of interest is water surface and it is best represented by several rectangular areas inside the lake Štítarský (see Figure 1). Based on these training areas the application classifies the image, i.e. marks pixels with satisfying degree of similarity the pixels from training areas as water surface. Based on the characteristics of the processed image it might be necessary to try different classifying methods and/or adjust parameters of these methods.

When the result of classification is satisfying (see Figure 2), the surface areas of classified regions can be computed by using the “Statistics” utility from the “Tools” menu. This tool generates a table listing total surface areas of all classified regions in the image. In this case the area of water surfaces—the surface area of lake Štítarský in the first half of 19th century—is approximately 112 ha (hectare). For better demonstration a layer with contemporary map source can be displayed, for example the Basic Map of the Czech Republic 1:10 000 (as shown on Figure 3). The classified region then clearly designates the area where the lake used to be and changes to the area can be easily discerned.


IGC Klasifikace - unnamed
Projekt Snímky Třída Klasifikace Zobrazit Nástroje Nápověda
Snímky Statistika
CLASSES
vodstvo
colour
DATA SOURCES
IIvojenskemapovani
transparency
Results
IIvojenskemapovani-result-0
IIvojenskemapovani-result-1
IIvojenskemapovani-result-2
ZM200
ZM10
Štítar
Slower
KÖNIGSTADTL

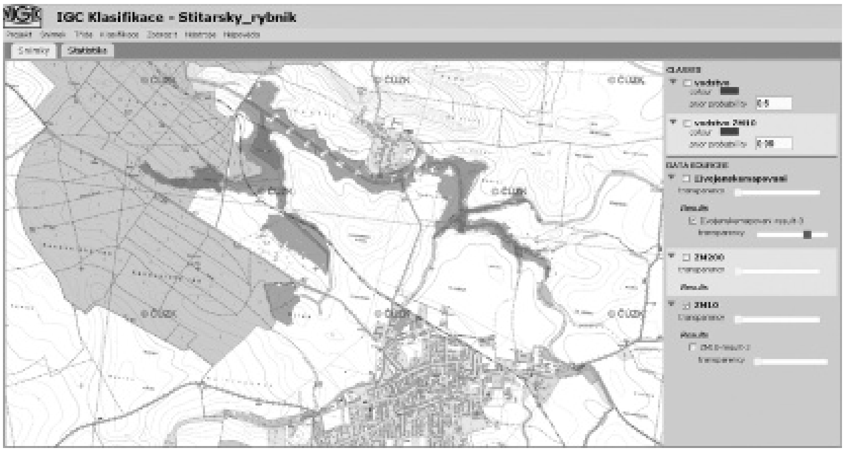
prior probability 0,6

Figure 1. Lake Štítarský, training areas selection.


IGC Klasifikace - unnamed
Projekt Snímky Třída Klasifikace Zobrazit Nástroje Nápověda
Snímky Statistika
CLASSES
vodstvo
colour
DATA SOURCES
IIvojenskemapovani
transparency
Results
IIvojenskemapovani-result-0
IIvojenskemapovani-result-1
IIvojenskemapovani-result-2
IIvojenskemapovani-result-3
ZM200
ZM10
Štítar
Slower
KÖNIGSTADTL

prior probability 0,6

Figure 2. Lake Štítarský, result of classification.



prior probability 0.5

prior probability 0.30

Figure 3. Lake Štítarský, classification result displayed over contemporary map source.

5 CONCLUSION

The goal of this contribution was to point out the fact, that today users of the old maps are not content with bare online accessible maps any more. Nowadays an added value is required—online tools that allow to use the digitalized maps more efficiently, more easily and to get more information from them than from the original paper maps. One of such useful tools can be the automatic classification of old digitized maps.

This article discusses the methods for automatic classification of raster images of the old maps and presents a practical online tools and a practical example of use of classification of old maps. One of the practical results of such a classification can be, for example, the discovery of the original area of the water surface (lake) from the old map in a certain period.

REFERENCES

Denison D.G.T., Holmes C.C., Mallick B.K. & Smith A.F.M. 2002. *Bayesian Methods for Nonlinear Classification and Regression*. Willey series in probability and statistics. John Willey & Sons. ISBN: 978-0-471-49036-4.

OGC 2006. The OpenGIS® Web Map Service Interface Standard (WMS) http://www.opengeospatial.org/standards/wms.

Talich M., Böhm O. & Soukup L. 2012. Classification of digitized old maps and possibilities of its utilization. e-Perimetron, Volume 7, No. 3: 136–146, ISSN 1790-3769, http://www.e-perimetron.org/Vol_7_3/Talich_et_al.pdf.

Webb, A.R. 2003. *Statistical Pattern Recognition*. Second Edition, John Wiley & Sons, Ltd, Chichester, UK. ISBN: 9780470845134.