Improving Feature Extraction of Composite Cartographic Information in Low-Quality Maps

Stefan Leyk and Ruedi Boesch

ABSTRACT: We describe an extraction method for area features, which are defined by composite cartographic elements and derived from historical manually produced maps of low graphical quality. Composite elements appear in many existing topographic maps of the 19th and 20th century, which provide unique information about the landscape in the past. To advance recent research efforts we further developed a method for extracting forest area in the Siegfried Map, which is represented by a set of circle-like forest symbols and boundary regions for closed forest.
First, a prototype search identifies forest symbols that can be characterized by "idealized" combined properties using geometric attributes of connected components, morphological properties of the local image plane and the degree of spatial association between similar objects. Next, the complete set of forest symbols is iteratively determined. Forest symbol candidates in the vicinity of prototypes are labeled as prototypes if they reach conditions of spatial association. Finally, spatial expansion determines the forest net area, which is described by the set of recognized forest symbols, and continues to fill gaps between forest net area and boundaries as well as larger objects within forest area.
The automated extraction from three map pages resulted in accuracies of 95% (Kappa) and indicates a high robustness for automated processing of the whole map series. The developed approach represents a methodological framework for the extraction of areas, which are described by composite map elements, within similar cartographic documents.

KEYWORDS: Pattern recognition, historical topographic maps, composite map symbols, feature extraction, retrospective land cover change analysis.

Introduction

Pattern recognition in cartographic documents aims at the delineation and extraction of spatial information and its incorporation into GIS as raster or vector data (Chen et al., 1999). To establish methods for recognition in maps is particularly challenging because of the complexity of map contents (Cordella and Vento, 2000; Watanabe, 2000) as well as the presence of single and composite map elements (Llados et al., 2002). These composite elements consist of sets of low-level symbols whose spatial distribution and pattern can define higher-level spatial objects such as paths (Gamba and Mecocci, 1999), line-work (Yamada et al., 1993; Zhong, 2002), or forest areas (Leyk et al., 2006). The recognition of such complex objects is particularly problematic in maps of low quality such as historical maps, which are often hand-drawn documents with vague information about the underlying concepts for map production. Such historical maps are valuable sources for land cover change analysis and represent the most reliable data source of historical land cover before aerial photography could be used for interpretation. A large number of historical map series exists (Figure 1), which indicates the high demand for such extraction methods.

Forest area in the Siegfried Map, the Swiss national topographic map of the 19th century, is one typical example of composite spatial information. The representation of forest consists of distributed individual circular symbols bounded by line objects or unbounded,

for closed or open forests, respectively. Line objects, which represent the boundary of forest areas, are frequently fragmented, merged with other objects or missing due to the manual drawing process. This results in an inherent inconsistence and requires the description of such objects by combining geometric, morphological and structural map attributes. The Siegfried Map also represents an example of low graphical quality due to ageing, blurring, false coloring and mixed coloring. Recent research efforts resulted in a first prototype for multi-step forest extraction (Leyk et al., 2006). This prototype aimed at the recognition of all relevant map objects such as text, bedrock, line-work or buildings. The drawback of this approach was its complexity and thus its lacking robustness if map properties such as font size of map text changed drastically. Nevertheless this recognition prototype was successful in that it provided the fundamental principles to successfully extract forest as an area object and indicated limitations where improvement was needed.

This paper describes follow-up research to these earlier efforts aiming at a simpler but more robust forest extraction process, which does not require the recognition of objects of any other map category. The extraction process is based on the combination of different information domains such as geometric and morphological attributes of objects (connected components), analysis of spatial association of similar objects and morphological characteristics within the local image plane. The described sequence of methodological steps can be transferred to similar recognition problems where composite map elements have to be extracted. In this paper the approach is demonstrated to work for maps of low graphical quality. Thus it would allow the extraction of spatial composite information from other historical maps of similar time periods (Figure 1).

(a) (b) (c) (d)

Figure 1: Subsections of historical topographic map series with composite elements describing forest area: (a) Military Geographical Institute, Poland 1930, 1:25 000; (b) Royal Prussian Surveying Unit, Map of Western Russia, 1915, 1:100 000; (c) Imperial and Royal Military Geographical Institute, Austria, Map of the Austrian-Hungarian Monarchy and foreign map pages, Russia, 1878, 1:75 000, (d) Federal Topographic Bureau, Swiss Topographic Map (Siegfried Map) 1912, 1:25 000.

Data and Material

The Siegfried Map is the national topographic map series of Switzerland of the 19th century and was published in the scales of 1:25,000 in the Swiss Midlands and 1:50,000 in the mountainous areas. Forest belongs to the black color layer. The graphical representation of the black layer is characterized by merged objects of different categories, inconsistent shapes and varying

dimensions of map symbols, which are consequences of manual production techniques (engraving). In addition to these problems the scanned maps suffer from ageing effects, blurring, false and mixed coloring and thus result in low image quality (Figure 1d). The map data in this paper are preprocessed to carry out color image segmentation using a method that is described in Leyk and Boesch (in press). The result of this segmentation process is an image that contains the reconstructed black, red and blue map color layers (Figure 3a).

Methods

The method is based upon morphological attributes, which are derived from connected components and from the local image plane, as well as upon the spatial association of similar objects.

The different stages of this extraction process are:

- Low-level recognition of prototypes (individual forest symbols)
- Defining the composition of symbols and their spatial extent
- Identification of boundary and embedded objects

Low-level recognition of prototypes

In a first step forest symbol prototypes are identified based on geometric and morphological properties of connected components, a prototype test in the local image plane ("Octopus" test), a test for spatial containment, as well as a test for spatial association among similar objects in the local environment of the symbol of consideration.

Region-based candidate search

Connected components of the black layer in the color-segmented image (Figure 3a) are processed to derive geometric and morphological attributes of the resulting regions. We used attributes such as area, number of holes, perimeter, dimensions in horizontal and vertical direction, and circularity. This attribute space is used to define and label an initial set of "first-level" candidates (Figure 3c) that meet the general geometric constraints of forest symbols. We used broad ranges of admitted attribute values to ensure that all individual symbols are included regardless of their fragmentation and distortion, which are consequences of manual production, scanning and color segmentation.

A subset of these first-level candidates is defined by a filtering step in attribute space to identify objects, which reach conditions of "idealized" forest symbols. These candidates are labeled as "direct prototype candidates" (Figure 3c). They meet much stricter constraints of geometric and morphologic properties, which indicate a high similarity to a non-distorted and non-fragmented forest symbol.

Morphological test in the image plane

Due to the often distorted and fragmented forest symbols their original shape is not a reliable feature for recognition. For this reason it was necessary to develop a specific morphological test for circle-like local environments, which is independent from connected components (Figure 3b). This "Octopus" test has a distant similarity to concepts found in the generalized Hough-transformation (Ballard, 1981). Instead of using a functional model, a geometric model is combined with a spectral model to describe a single forest symbol. A forest symbol is characterized by a court of bright pixels in the center from which a raising and falling edge transition can be observed in eight search directions according to eight limbs of an octopus

(Figure 2). In contrast to common edge models in computer vision, which depend on second order differentiation, we use a weaker edge definition. Starting from the center, a valid edge is identified if a transition from bright to dark is followed by a transition from dark to bright (direction E in Figure 2). Intermediate dark pixels between raising and falling edge transitions are valid (direction W). A center is considered bright if there is a majority of bright pixels (here 8 of 9 pixels within the central 3x3 box). Direction SE represents a valid edge transition; directions S and SW are invalid because there is no transition to dark. Dark and bright are based on value ranges in lightness and hue. Due to the coarse resolution and the small size of forest symbols (typical box size is between 11x11 and 15x15 pixels) a finer differentiation of angles for search directions is not meaningful and would result in highly increased computational burden.

In imitation of the Hough space, the accumulator space of the Octopus is defined by the edge test in eight directions. If a sufficient number of edge evaluations are valid (6 in Figure 2), the shape in the local image plane is recognized as a circular symbol that meets the morphological conditions of forest symbols and the central pixel is labeled, accordingly.

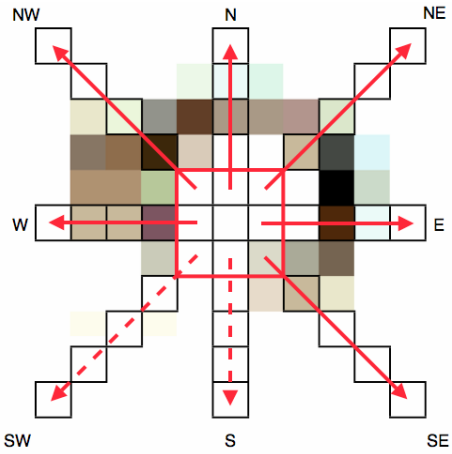


Figure 2: Illustration of the Octopus test to examine for edges in eight different directions.

Containment test and spatial association of prototypes

Next the pixels that are labeled as Octopus are tested for containment within the boundaries of one of the direct prototype candidates. If the containment condition is true, the candidate is further examined; if not, the candidate is degraded to a first-level candidate.

The local neighborhood of the prototype candidate is examined for the presence of a minimum number of other candidates (first-level or prototypes). If this condition is reached the candidate is labeled as a prototype (Figure 3c). We examined for the presence of at least three other candidates within a 50x50 pixel window.

This three-step test strategy, which combines attributes at the object level, shape descriptors of the local image plane as well as the degree of spatial association, aims at highest certainty that only well-shaped forest symbols of certain dimensions are identified as prototypes.

Defining the composition and spatial extent of forest symbols

The derived primary set of prototypes represents the starting point to examine each first-level candidate whether it belongs to the spatial set of forest symbols or not. This represents an iterative process that includes a test for spatial association of the candidates in the vicinity of prototypes with similar objects. Once the complete set of spatially “associated” forest symbols is

identified the spatial extent, which these symbols describe, is determined to derive the forest net area.

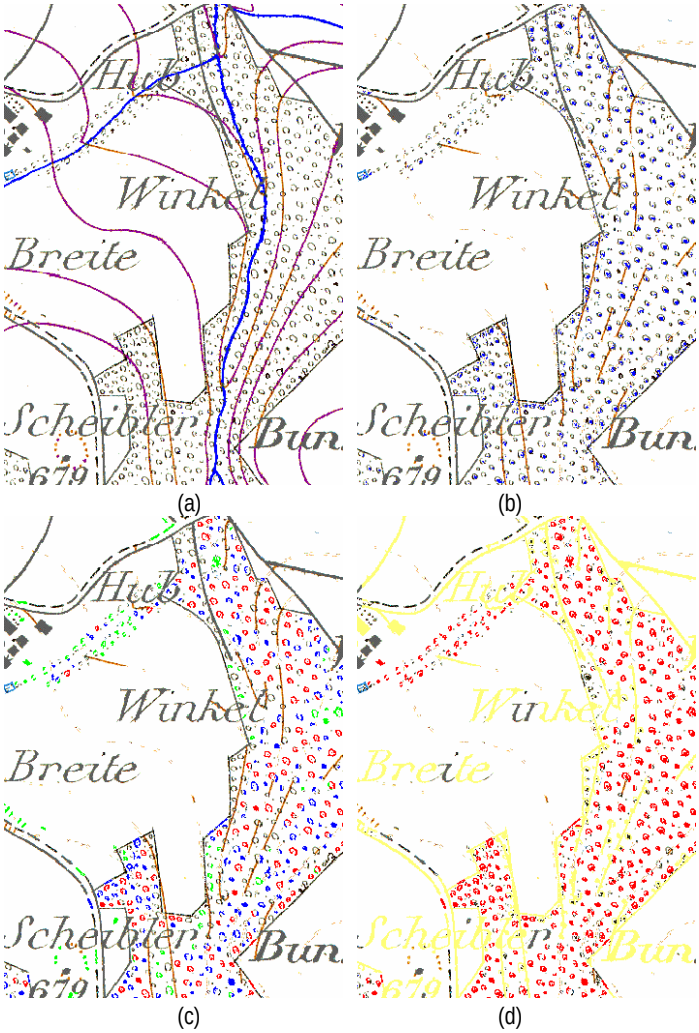


Figure 3: Prototype search and symbol composites: (a) Color-segmented image; (b) Labels of approved Octopus tests (blue); (c) Types of candidates after the prototype search (prototypes: red, second-level candidates: blue, remaining first-level candidates: green); (d) Composition of forest symbols after iterative analysis for spatial association (red) and large objects (yellow).

Iterative analysis of symbol composites

First-level candidates in the vicinity of the primary prototypes are labeled as “second-level” candidates (Figure 3c). These second-level candidates are then examined whether they are in spatial association with a minimum number of similar objects or not (first-level, second-level or prototype). If this condition is reached the second-level candidate is labeled as a prototype (Figure 3c); if not the candidate is eliminated from the set of potential candidates. First-level candidates that are found in the search box are labeled as second-level candidates to initiate the next iteration. This procedure is repeated until no new second-level candidate is labeled and provides the complete set of forest symbols (Figure 3d).

Spatial expansion

The spatial extent of forest net area is determined by expanding contiguous forest color between individual pixels of the identified forest symbols (set of prototypes), which are less distant from

each other than a predefined distance l_{netfor} (l_{netfor}=18 pixels) (Figure 4a). Expansion takes place along search paths of length l_{netfor} in each of the eight neighbor directions d_{netfor}. All examined locations along the search path between the starting pixel and the most distant forest pixel found are labeled as forest.

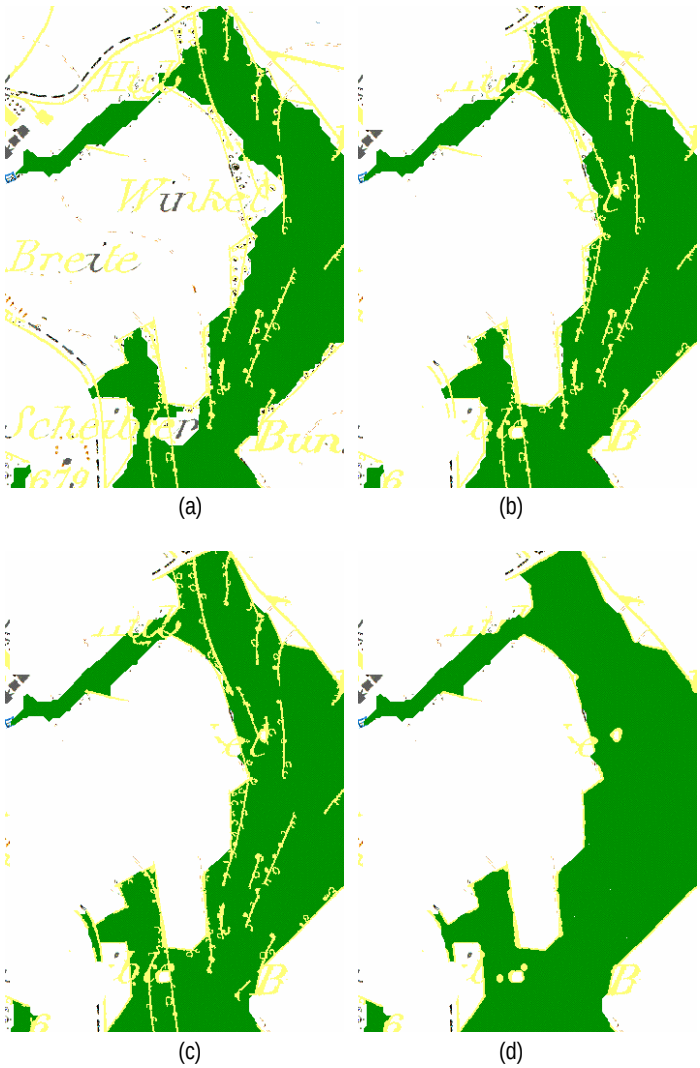


Figure 4: Stages of spatial expansion: (a) based on the set of forest symbols; (b) including nearby Octopus labels (study area clipped using buffered forest area); (c) Expansion after filling gaps between forest and large objects in the vicinity; (d) Filling of embedded large objects of low local dimension.

There is only one constraint included with regard to large objects. Large objects are connected components with an area greater than 100 pixels and include, e.g. road segments or boundary regions. If a pixel of a large object is encountered, the search path is examined in full length. If a forest pixel is found after encountering the large object, expansion takes place but is restricted to the pixels between the starting and the encountered large-object pixel. If no forest can be found nothing is done. This constraint ensures that large objects within forest such as roads are completely embedded but not crossed and that forest is not expanded uncontrolled. Spatial expansion continues, iteratively, until no new location is labeled as forest. The result is the spatially expanded forest net area, which is described by the set of associated forest symbols (Figure 4a).

In a next step the remaining Octopus labels, which are located in the vicinity of the net forest margin, are included in the expansion process (Figure 4b). Thus forest symbols, which do not reach the geometrical requirements needed because they are merged with forest boundary fragments, bedrock, text or remaining color layer segments, can still be used for expansion. If the Euclidean distance between an Octopus label and the closest forest pixel is less than a given threshold (10 pixels) the Octopus label is converted to forest color and included in the iterative expansion process. This step further expands forest net area towards large objects without closing the gaps (Figure 4b).

Boundary identification and filling embedded objects

The derived forest net area is buffered using the observed mean distance between the boundary and the closest non-merged forest symbols, multiplied by two. The whole image is clipped using this buffered area to constrain the boundary search to this region (Figure 4b). This intermediate step increases efficiency and prevents incremental extraction errors.

If a forest pixel has a background pixel in its direct neighborhood in direction d_{bound}, search paths of length l_{bound} (l_{bound}=8) are defined in the same direction. If a large object is encountered the pixels between the starting pixel and the first large-object pixel along this path are labeled as forest. There is no requirement to identify a forest pixel after encountering a large-object pixel. This process is repeated until there is no new forest pixel labeled (Figure 4c).

The final step aims at filling line-like large objects that are embedded in forest area using search paths similar to the boundary identification. If a forest pixel has a large-object pixel in the direct neighborhood, search paths are defined in this direction. If another forest pixel is found within a very short distance (3 pixels) the local dimension of the large object encountered is considered very low. Consequently, the pixels along this search path are labeled as forest (Figure 4d).

Results and Discussion

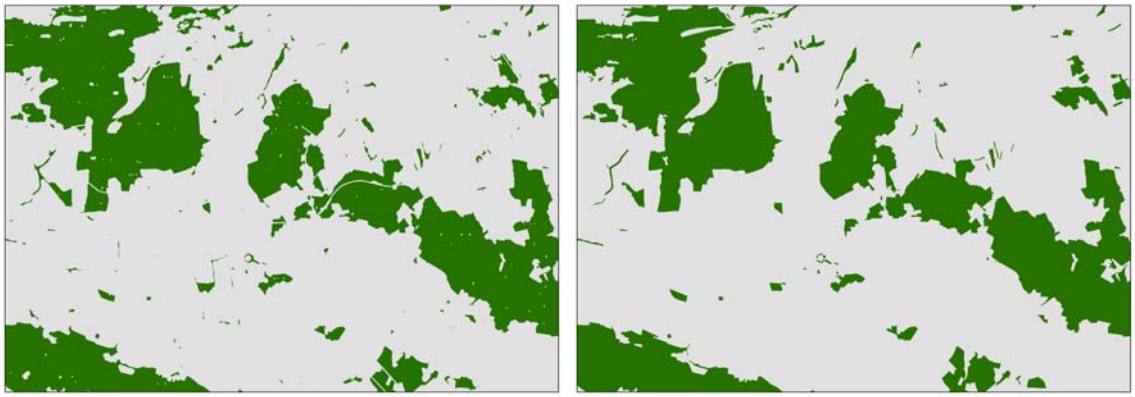
We compared the forest extraction with a visual inspection (manual digitization) done by an experienced cartographic interpreter to evaluate the performance of the method. The underlying presumption is that the human being is the most reliable interpreter of cartographic information due to his/her ability to visually explore spatial and topological relationships, even if the graphical representation lacks in completeness, quality or precision.

Accuracy assessment: We tested three complete map pages with 7000 × 4800 pixels each to obtain some first estimations of the robustness of the extraction approach, which was developed on a different set of map pages. A simple confusion matrix was established and different class-specific and global accuracy measures could be derived. Sensitivity and specificity (Fielding and Bell, 1997) are the conditional probabilities that forest or non-forest, respectively, is correctly classified. Positive predictive power (PPP) and negative predictive power (PNP) (Fielding and Bell, 1997) indicate the probabilities that a pixel is forest or non-forest if the extraction classifies it as forest or non-forest, respectively. To estimate the overall classification accuracy some global measures were calculated, i.e., percent correctly classified (PCC) or accuracy (ACC) (Michie et al., 1994), Kappa coefficient of agreement (K) (Cohen, 1960) and normalized mutual information criterion (NMI) (Forbes, 1995). K and NMI represent more conservative measures than simple accuracy estimates, which make full use of the information contained in the confusion matrix.

Table 1: Extraction results as validated for three whole map pages of 4800x7000 pixels each.

Confusion matrix derived measures	Map page 042	Map page 043	Map page 211	Total
Sensitivity	0.97	0.98	0.97	0.98
Specificity	0.98	0.98	0.97	0.98
Positive predictive power PPP	0.95	0.96	0.93	0.95
Negative predictive power PNP	0.99	0.99	0.99	0.99
PCC/ACC	0.98	0.98	0.97	0.98
Kappa	0.95	0.96	0.93	0.95
NMI	0.85	0.86	0.79	0.83

Performance: The presented results (Table 1) demonstrate a very high global accuracy of 95% in average (Kappa). The parameters defined proved to be valid for each map page tested. Thus the extraction process runs very robust with a certain independence from the color segmentation process except in some situations, which are described below. Sensitivity (0.98) and specificity (0.98) indicate high conditional probabilities that forest and non-forest, respectively, are correctly classified. PPP shows the smallest class-specific probability measure in Table 1 (PPP=0.95). This can be partly explained by a trend of the extraction process to "over-detect" forest. Existing small groups or chains of forest symbols are very likely to be detected and will be extracted as forest area (Figure 5, 6). In some instances the interpreter did not delineate these symbol groups as forest because of their small size or elongated shape (Figure 6). This effect results in a slight decrease in PPP. A subsequent filtering step to eliminate forest patches with areas below a threshold value would further improve the accuracy.



(a) (b)

Figure 5: Forest/non-forest classifications of map page 042 (1913) (a) automatically extracted and (b) visually inspected.

Observed minor problems: Some problems occurred where map text crossed forest boundary regions. The expansion of forest area would often not allow to bridge objects of this size resulting in some underestimations (decreasing PNP). In other cases text objects could result in some overestimations (decreasing PPP) if they are located outside the forest boundary but falsely identified as boundary objects (Figure 6).

If elongated color layer fragments of e.g., elevation contours remained in the image they are completely embedded in forest area by the expansion process. Only if such an object is located in the direct vicinity of the forest boundary, it is treated as a boundary object in some instances and can result in decreasing PNP and sensitivity.

Another problem observed is the false recognition of forest area where backyard signature that surrounds buildings shows geometric properties similar to forest symbols. This problem could be reduced by combining the extraction with a topological analysis of building symbols. Similarly, riverbanks, which have a punctuated signature but occasionally show geometric properties similar to forest, are misclassified as forest. These two problems result in decreasing PPP and specificity.

There are some examples where broken and fragmented forest boundary objects allow for the expansion of forest area over the boundary especially if other forest symbols outside but in close vicinity to the boundary are found (Figure 6). These “overshoots” represent a minor problem due to a strict expansion control but contribute to decreasing PPP and specificity.

Figure 6: Detail of the extraction: (a) Color-segmented image; (b) Automated extraction (green, superimposed over the map); (c) Visual inspection (green, superimposed over the map).

The most important drawback is the missing ability to differentiate between bedrock and text, which require semantically different actions during extraction (expansion vs. clipping). Since these objects strongly vary in shape and size and are frequently merged with other objects a reliable differentiation based on geometry or morphology is nearly impossible. Even text recognition of isolated labels is very demanding, because text is often curved or rotated and has varying character spacing within the same label word. A feasible solution to this problem would be to label these conflicting locations, treat them as described in this paper, and include an interactive component at the end of the automatic extraction process. The analyst would return to these locations and correct the result if the large object encountered represents bedrock. This interaction is at the cost of full automation but will ensure a minimum recognition error – especially in mountainous regions where bedrock can have considerable spatial extents. As can be seen in Figure 5 there are small holes in the extracted forest area. These are examples where text elements create a bright court similar to bedrock and thus require a revisiting to verify the category of the encountered object.

Conclusions

The presented methodological framework describes the procedure for robust automated extraction of areas, which are defined by composite cartographic elements in mainly hand-drawn historical map documents. The strength of this approach is that extraction proceeds very class-oriented without trying to recognize other map categories – one mentioned problem in recent research efforts to develop a recognition prototype for the investigated map (Leyk et al., 2006). The conservative strategy of testing for combined salient attributes before adding an object to the set of forest symbols and conducting constrained expansion allowed to minimize extraction errors. Thus the extraction performed equally well in maps where color image segmentation showed some problems or fragmentation of the forest boundary regions was observed.

In general the presented results, their reproducibility and the automation level reached give reason to continue such extraction efforts for the Siegfried Map but also for comparable maps from similar time periods. Whereas such historcial documents impose significant problems of inherent uncertainty (Leyk and Zimmermann, 2007) retrospective landscape research for land cover change, urban development or conservation will greatly benefit from gaining access to such unique historical information for large areas. Thus the authors hope to see an increasing number of research efforts to make these historcial documents available for GIS-based analysis.

With regard to the presented method, an improved analysis for spatial association between similar objects will be tested. This test will take into account the directions in which similar objects are found. Thus additional constraints can be formulated for forest symbols that are located inside a forest patch as opposed to those located at the margin of forest patches. These methodological extensions will be linked to the problem of composite elements that cause the described overestimations of forest area in urban environments. Further steps will be taken to focus on the problem of differentiating between bedrock and text elements inside forest areas as mentioned above. While this paper focused on forest extraction we will also examine the extraction of other composite elements such as wetlands or vineyards.

REFERENCES

Ballard, D.H. (1981) Generalizing the Hough Transform to Detect Arbitrary Shapes. ***Pattern Recognition***, 13, pp. 111-122.

Chen, L., Liao, H., Wang, J., and Fan, K. (1999) Automatic Data Capture for Geographic Information Systems. ***IEEE Transactions on Systems, Man, and Cybernetics C***, 5, 2, pp. 205–215.

Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales. ***Educational and Psychological Measurement***, 20, pp. 37-46.

Cordella, L. and Vento, M. (2000) Symbol Recognition in Documents: A Collection of Techniques?. ***International Journal of Document Analysis and Recognition***, 3, pp. 73-88.

Fielding, A.H. and Bell, J.F. (1997) A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models. ***Environmental Conservation***, 24, 1, pp. 38–49.

Forbes, A.D. (1995) Classification Algorithm Evaluation: Five Performance Measures Based on Confusion Matrices. ***Journal of Clinical Monitoring and Computing***, 11, pp. 189-206.

Gamba, P. and Mecocci, A. (1999) Perceptual Grouping for Symbol Chain Tracking in Digitized Topographic Maps. ***Pattern Recognition Letters***, 20, pp. 355-365.

Leyk, S. and Boesch, R. (accepted) Colors of the past: Color image segmentation in historical topographic maps based on homogeneity. ***GeoInformatica***.

Leyk, S., Boesch, R., and Weibel, R. (2006) Saliency and Semantic Processing – Extracting Forest Cover from Historical Topographic Maps. ***Pattern Recognition***, 39,5, pp. 953-968.

Leyk, S., Zimmermann, N. (2007) Improving land change detection based on uncertain survey maps using fuzzy sets. ***Landscape Ecology,* 22**, pp. 257-272.

Llados, J., Valveny, E., Sanchez, G. and Marti, E. (2002) Symbol Recognition: Current Advances and Perspectives, in: D. Blostein, Y.-B. Kwon (Eds.), Fourth IAPR Workshop on Graphics Recognition. ***Lecture Notes in Computer Science***, 2390, Springer, Berlin, pp. 104–128.

Michie, D., Spiegelhalter, D. and Taylor, C. (1994) ***Machine Learning, Neural and Statistical Classification***, Ellis Horwood.

Watanabe, T. (2000) Recognition in maps and geographic documents: features and approach, in: A.K. Chhabra, D. Dori (Eds.), Third IAPR Workshop on Graphics Recognition. ***Lecture Notes in Computer Science***, 1941, Springer, Berlin, pp. 39–49.

Yamada, H., Yamamoto, K., Hosokawa, K. (1993) Directional Mathematical Morphology and Reformalized Hough Transformation for the Analysis of Topographic Maps, ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 15, 4, pp. 380–387.

Zhong, D. (2002) Extraction of Embedded and/or Line-Touching Characterlike Objects, ***Pattern Recognition***, 35, pp. 2453–2466.

Stefan Leyk, Assistant Professor, Department of Geography, University of Colorado, Boulder. Boulder, CO 80309. E-mail:<stefan.leyk@colorado.edu>.

Ruedi Boesch, Research Associate, Federal Research Institute for Forest, Snow and Landscape (WSL). Birmensdorf, CH-8903, Switzerland. E-mail:<ruedi.boesch@wsl.ch>.