Geoff Groom¹, Gregor Levin¹, Stig Svenningsen², Mads Linnet Perner²

Historical Maps – Machine learning helps us over the map vectorisation crux

Keywords: Denmark, topographic maps, land category, OBIA, CNN, map symbols.

Summary: Modern geography is massively digital with respect to both map data production and map data analysis.

When we consider historical maps, as a key resource for historical geography studies, the situation is different. There are many historical maps available as hardcopy, some of which are scanned to raster data. However, relatively few historical maps are truly digital, as machine-readable geo-data layers. The Danish “Høje Målebordsblade” (HMB) map set, comprising approximately 1100 sheets, national coverage (i.e. Denmark 1864–1920), and geometrically correct, topographic, 1 : 20 000, surveyed between 1842 and 1899, is a case in point. Having the HMB maps as vector geo-data has a high priority for Danish historical landscape, environmental and cultural studies. We present progress made, during 2019, in forming vector geo-data of key land categories (water bodies, wetland, forest, heath, sand dune) from scanned HMB printed map sheets. The focus here is on the role in that work of machine learning methods, specifically the deep learning tool convolutional neural networks (CNN) to map occurrences of specific map symbols associated with the target land categories. Demonstration is made of how machine learning is applied in conjunction with pixel and object based analyses, and not merely in isolation. Thereby, the strengths of machine learning are utilised, and the weaknesses of the applied machine learning are acknowledged and worked with. Symbols detected by machine learning serve as guidance for appropriate values to apply in pixel based image data thresholding. The resulting map products for two study areas (450 and 300 km²) have overall false-positive and false-negative levels of around 10% for all target categories. The ability to utilise the cartographic symbols of the HMB maps enabled production of higher quality vector geo-data of the target land categories than would otherwise have been possible. That these methods are in this work developed and applied via a commercial software (Trimble eCognition) recognizes the significance of a tried-and-tested and easy-to-use, graphical-user-interface and a fast, versatile processing architecture for development of new, complex digital solutions. The components of the resulting workflow are, in principle, alternatively usable via various free and open source software environments.

Introduction

Increasingly, the conversion of paper or image media historical maps to machine readable digital geo-data layers is moving from time and resource demanding visual interpretation of the maps and manual digitisation to less time-consuming automated production methods. In a Danish context, analyses of historical maps have most often been elaborated only for a limited number of map sheets and relatively small study areas (e.g. Fritzbøger and Caspersen 2002; Jensen and Jensen 1979; Kristensen et al. 2009; Svenningsen et al. 2015) due to a reliance on visual, manual methods. There are, however, exceptions, such as Dam (2005), who, for all of Denmark, manually produced geographical information from late 18th century maps.

¹ Aarhus University, Denmark [gbg@bios.au.dk]

² Royal Danish Library, Denmark

In Denmark there is a unique opportunity to develop and apply digital map processing methods for automated extraction of land categories from historical maps as all publicly produced topographical maps are made freely available as high-resolution scans by the national mapping agency (SDFE). An automated extraction of land categories from historical topographic maps will provide the opportunity to generate countrywide map layers, representing the location and extent of different land categories at different points in time. Such map layers can be applied in spatially detailed assessments of land use and land cover (LULC) changes. Being spatially explicit, the map layers can be used in assessments at any spatial level, from a specific location or plot to different spatial delineations, such as parishes, municipalities, watersheds or the whole country.

The topographical maps ‘Høje målebords’ (HMB) represent the first large-scale topographic mapping of Danish territory. The HMB map series consists of about 1100 sheets at a scale of 1 : 20 000 surveyed between 1842 and 1899 and was developed by the Danish military as a response to a growing need for detailed topographical information about the landscape for military purposes.

Digital map processing is a relatively young research area, which grew out of disciplines such as graphics recognition and document analysis (Chiang et al. 2016, Freeman and Pieroni 1980, Liu et al. 2019). Due to various graphical quality issues and the general complexity of the map contents, extraction of information from maps is generally more difficult compared to processing of other data sources, such as written documents or technical drawings (Lladós et al. 2002). In maps, layers of geographic features, such as roads, contour lines, text labels, and land categories, often overlap with each other, increasing the degree of content complexity and colour mixing. Furthermore, due to inconsistency of applied symbols and colours, processing of manually drawn maps is even more challenging (Leyk and Boesh 2009). General techniques for graphics recognition and document analysis can therefore not be directly applied to map processing. In most historical maps, colours, symbols, linear features and text items, represent explicit, overlapping layers of geographic information. Colours often represent thematic information of land categories, such as forest or water. Symbols, and to some degree text items describe different types of geographic characteristics. Linear features can represent geographic features, such as roads or watercourses, boundaries of land categories, administrative boundaries or topographical characteristics, such as contour lines. A major objective of digital map processing is to spatially isolate these overlapping layers of geographic information from each other and to filter out those map elements that obstruct the extraction of the geographical layer of interest.

Through 2019, resources were available to enable the authors to examine the possibilities of developing, undertaking and evaluating automated machine-readable geo-data layer production from HMB digital image sources, for five target land categories: water bodies, wind blown sand (dunes), wetland (mainly meadow), heath and forest (Levin et al. 2020), (*Figure 1*). The applied methodology was one of object-orientated image analysis (OBIA), largely reliant on expert knowledge derived rule based image segmentation and object labelling operations, rather than statistical classification methods. OBIA methods of that form are widely used in production of geo-data layers from image data as they enable high levels of control over mapping results (Blaschke et al. 2014, Chen and Weng 2018, Chen et al. 2018). For the mapping of water bodies, forest, heath and wetland the HMB colouration represents the major basis for formation of image objects. The core target class object mapping process thereby involves identification of

the relevant image layers, followed by selection of an appropriate value to apply in threshold based image object segmentation (Levin et al. 2020). However, the colour variability between the different map sheets (*Figure 2*) means that if constants (fixed values) are applied for the thresholding, results are generally poor across the set of map sheets. The HMB maps include the use of distinct symbols that associate with the target classes wetland and heath (also forest). Machine learning methods such as CNN are making increasingly marked in-roads into many areas of applied image work, often with impressive results (Ball et al. 2017). Occurrences of map symbols can be detected by application of the machine learning methods like CNN (Quan et al. 2018). In the HMB maps target class symbols lie within the zones of target class colouration. It was therefore postulated that CNN predictions of symbol locations can form a basis for determination of the appropriate threshold for good target class image segmentation. The aim of this paper is to present details of the CNN symbol prediction assisted thresholding approach, with a basic evaluation of its effectiveness.

Figure 1. Examples from the HMB map series, that includes areas having the HMB representation of the HMB legend categories (left-right), wetland (blue-green), heath (pink), forest (brown), water bodies (blue) and dunes (even black stipple).

Figure 2. Examples from the HMB map series, showing variations in the colouration applied for the HMB legend categories wetland (blue-green); each image is displayed without display stretching. Also illustrated in these examples is the variability in the form of the wetland symbol (examples indicated by added red box)

Materials

The work reported here uses two study areas, selected to (a) enable use of the results of the developed information production method as an indication of their more general applicability and (b) to enable analyses of the resulting land category mappings for a meaningful set of parish extents. The first of these objectives implied choice of study areas that between them included extensive areas of all the target land categories and source map data with differences typical of the range found across the full HMB map series. One study area is formed of the bounding-box of the parishes of Klejtrup, Nørre Onsild, Sønder Onsild,

Lindum, Øls and Hvornum, in eastern Jutland. That study area, of approximately 195 km², includes extensive areas of forest, wetland and heather. The second study area is formed of the bounding-box of the parishes of Hirtshals, Horne and Asdal, in northern Jutland. That study area, of approximately 90 km², includes extensive areas of heath, wetland and dune. Because the HMB sheets do not follow parish boundaries, nine HMB sheets were needed to completely cover the selected parishes in the study area in eastern Jutland, while six HMB sheet were needed to for the northern study area. Therefore data was extracted from all of these HMB sheets, thus creating an extended study area, covering approximately 422 km² and 13 parishes for the study area in eastern Jutland and approximately 200 km² and seven parishes for the study area in northern Jutland (*Figure 3*).

Figure 3. Locations of the two study areas.

There are two forms of HMB source material that the project that work reported here is part of, has used. One is from the digital archive of the Danish National Mapping Agency (SDFE) of scannings made of one set of HMB paper maps. These source data are publically available via SDFE web portal:

<https://download.kortforsyningen.dk/content/dtkh%C3%B8je-m%C3%A5leboardsblade>

as georeferenced (ETRS_1989_UTM_Zone_32N, resampled to a pixel size of 2 m) image data that have been mosaicked and then cut into a set of 20 x 20 km tiles as .tif (8-bit unsigned, LZW lossless compressed) files. The other source is the archive of HMB paper sheets held by the Danish Royal Library (KB) and are the original paper sheets, including margins. Use of items from this archive in this work data required scanning and geo-referencing, which was applied to just the map sheets needed in order to cover the project's study areas. Scanning of the KB sourced map sheets was done using a ScannTECH 600i-fb, with the standard settings of that scanner for colour artwork. A digitisation resolution of 600 dpi was used, which is appropriate for 1 : 20 000 scale source material (Tobler, 1988). The KB archive paper map sheets and the map sheets that were used for the SDFE scanning are not entirely the same raw material. Firstly, only approximately two-thirds of the sheets in the KB collection are full-colour with the remainder (mainly eastern Denmark) being just 3-colour (black, blue, brown), whereas the entire set that were scanned for the SDFE dataset are full colour. Where maps of both sources are full-colour, the actual colourations differ, with evidence that they relate to two sets of paper maps that were coloured independently, rather than just differential effects of time and storage conditions. The differences between the data from the two sources present interpretation challenges in a few places but, countering that, the situation is one of beneficial complementarity. Therefore, wherever possible and meaningful, the thresholding is made on an image layer from both image sets.

Methods

Numerous raster image processing methods can be applied to enhance the image data with respect to one or other aspect of the colouration. In general, initial gentle kernel-based filtering is useful to smooth out single pixel anomalies and is applied in this work. Applied enhancements are Hue-Saturation-Intensity, histogram normalisation and arithmetic combinations of the R-G-B (in raw-smoothed and in histogram normalised form) image data. Visual inspection indicates which derived image layer(s) are most likely to provide a satisfactory initial set of image objects for a target class. In general, the same derived image layers provides a good basis for class discrimination for each map sheet. For example, the relative greenness ($G/(R+B)$) based on the histogram normalised R-G-B layers are generally a potent means for mapping the wetland target class (Figure 4).

Figure 4. Three examples of the HMB data for areas with the legend class wetland, together (rows 2 and 4) with their representation as the derived image layer RGB-histogramNormalised-relativeGreenness; rows (1) and (2) are of data from the SDFE archive and rows (3) and (4) are of data from the KB map sheet collection

Simple threshold based image segmentation is used rather than alternative OBIA segmentation methods, such as multi-resolution segmentation (Baatz and Schäpe 2000), as the former involves fewer parameters and provides a high level of user control over the resulting objects.

An effective symbol CNN model, for an HMB symbol, is acquired by testing of sets of symbol target and non-target training data and CNN parameter settings for an independent set of HMB

images. For a specific map sheet, the use of a CNN model based symbol mapping is enabled as simple thresholding of the symbol CNN heatmap to give a set of image objects, followed-by filtering-out of objects that did not meet size and shape expectations for the symbol. Remaining symbol objects are then grown to include a zone of the surrounding image data. This set of objects is then filtered with respect to a relatively broad range of expected object means in one or more image layers (*Figure 5*). Most of the remaining symbol associated objects should then relate to areas of the target class. These objects are then used to derive the threshold that would be most appropriate to apply for initial target class image objects across the current image data. The appropriate threshold is determined through stepwise threshold lowering with testing for under-mapping and over-mapping with respect to the grown symbol objects. This iterative process also flags cases where an image layer cannot be used to effectively map a target class. The derived threshold value is used as a baseline value to apply in various subsequent threshold based segmentation and filtering operations in order to derive a full mapping of the target class.

Figure 5. One of the examples from Figure 4 with the corresponding wetland symbol raw CNN heatmap (centre) and symbol heatmap hotspot objects, grown and filtered to zone objects (right).

Results

The accuracy of the vector target class map layers produced by the entire workflow, including various stages of morphological and contextual refinement of the initial image objects (Levin et al 2020), have been assessed against a 100 m grid set of reference point data, approximately 29,000 points in total. These reference data have been derived by independent HMB image visual interpretation of the target class membership of each point. The final results of that accuracy assessment will be reported in Levin et al. (2020). An earlier accuracy assessment of initial target class mappings indicated false-negative and false-positive levels of around 10% for most classes on most map sheets. Those detected patterns of errors were applied for further development of the workflow.

Here, as an assessment of just the contribution of CNN symbol mapping to the adaptive threshold setting, a simple visual assessment is made between thresholding with a pair (one each for SDFE,

KB) of constants for three map sheets, and thresholding with pairs of variables, derived via the symbol CNN symbol predictions (*Figure 6*). For the left map sheet the constant based thresholding is nearly as effective as the CNN assisted variable thresholding, but for the other two map sheets it gives a clearly far inferior result.

Figure 6. Result mapping of the target class wetland for three HMB map sheets (top row) with a constant threshold values for all three map sheets (middle row) and variable thresholds set via use of CNN symbol predictions (bottom row).

Discussion

Work with historical maps, including their conversion to machine-readable geospatial information, should not be viewed as just a technical, computing task, but also one that requires understanding of the historical context of the source materials and thereby how their contents should be understood. The HMB maps are the primary source for detailed information about the landscape before the 20th century. However, despite the importance of the maps as a source for information about the landscape of the past, the surveys and production behind the maps have only to a limited degree been subject to research. Although the military origin of the maps and the possible implications for landscape analysis have been recognized (Brandt, 2004; Münier, 2009), our understanding of the land classification system is still limited to what is stated in the map legends (Korsgaard, 2006). This means that information about the categorisation and representation of land area types and landscape elements in the map is sparse and in principle limited to the signature tables published with the map series. Although some of the land area types shown in the legend seem to be directly comparable to modern LULC classification typologies, such as forest vegetation, recent work shows that classification of land area types has changed between different map types and over time (Svenningsen et al., 2015). These aspects of the HMB maps are thoroughly discussed by Levin et al. (2020). However, those considerations do not have a decisive impact upon the general principle of the technique for improved colouration segmentation described here.

The presented approach to the mapping of the heath and wetland target classes from the HMB data represents a robust integration of modern machine learning methods in a broader image analysis workflow. Therefore, the overall workflow is not reliant on CNN method alone to map the target classes, and indeed, does not require that the CNN method is highly efficient at mapping the target class symbol, since filters are applied to the initial CNN symbol mapping result. Therefore the approach might be regarded as “soft machine learning”.

There are at least two further ways that the CNN symbol mapping can contribute, in a soft way, to an image object based workflow, such as the one developed and applied in this work. (1) Actual target class extent that are otherwise unmapped as target class objects (i.e. false negative errors) but having strong indication by the symbol’s CNN heatmap of symbol presence, can form seeds for additional target class image objects. (2) Wholesale absence of symbol CNN heatmap symbol indications from larger target class image objects can indicate cases of over-mapping (false positive errors).

The approach presented in this paper is dependent upon there being a sufficient number (at least around 10) of target class symbols present and detected by the CNN model. If the target class is only minimally present in the image, this condition may not be met. That situation is flagged for in the workflow, requiring then use of a manually determined threshold for that map sheet. Pre- and post-hoc visual examination of map sheets and the resulting maps for anomalous conditions is part of the full workflow.

Whilst machine learning methods such as CNN are making increasingly marked in-roads into many areas of applied image work, often with impressive results (Ball et al. 2017), it is still prudent to develop hybrid approaches that also utilize other image analysis methods. Most forms of machine learning are still dependent on training data, which can represent a manual, time consuming component of the application. In principle, these methods can also self-learn as they are applied to more and more image examples. However, production of perfect CNN models should not become a goal for its own sake, unless it can be objectively expected that they will deliver the required mapping in an almost perfect form. A machine learning solution for production of perfect geospatial machine-readable information from historical maps, including extensive spatial features, is yet to be demonstrated, and it is therefore prudent to acknowledge the limitations of current machine learning and develop the degree to which machine learning can assist other methods.

The methods described here have been developed and applied within a commercial software (Trimble eCognition, <https://geospatial.trimble.com/products-and-solutions/ecognition>). However, all of the steps could just as well be applied outside of that software, via various free and open-source programming languages and programming environments. The advantages of working within a commercial software are that the architecture of the component operations is efficiently programmed for fast processing, the different components are well integrated and the software’s graphical user interface enables fast development of new ideas. The disadvantage, other than the cost, is that the developer is limited to the set of and implementation of operations as they have been programmed within the software, some parts of which may be just barely documented.

Conclusion

A method has been described and demonstrated whereby basic CNN machine learning map symbol detection results assist image object methods to make improved mapping of map legend areal units associated with distinct map colourations and associated map symbols.

References

- Baatz, M. and A. Schäpe (2000). Multiresolution Segmentation – an optimization approach for high quality multi-scale image segmentation. In: J. Strobl, T. Blaschke and G. Griesebner (eds.) *Angewandte Geographische Informationsverarbeitung XII*. Heidelberg: Wichmann-Verlag, 12-23.
- Ball, J.A., D. Anderson and C.S. Chan (2017). Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of Applied Remote Sensing* 11(4): 042609 <https://doi.org/10.1117/1.JRS.11.042609>
- Blaschke, T., G. J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R., Queiroz Feitose, R., F. van der Meer, H. van der Werff, F. van Coillie and D. Tiede (2014). Geographic Object-Based Image Analysis – Towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing* 87(0): 180-191. <https://doi.org/10.1016/j.isprsjprs.2013.09.014>
- Brandt, J. (2004) Anvendelse af topografiske kort til monitoring af biotoper. *Geoforum Perspektiv* 3(5): 53–61.
- Fritzbøger, B. and O. H. Caspersen (2002) Long Term Landscape Dynamics - a 300-Years, Case Study from Denmark. *Danish Journal of Geography* Special Issue 3: 13–27
- Chen, G. and Q. Weng (2018). Special issue: Remote sensing of our changing landscapes with Geographic Object-based Image Analysis (GEOBIA). *GIScience & Remote Sensing* 55(2): 155-158. <https://doi.org/10.1080/15481603.2018.1436953>
- Chen, G., Q. Weng, G. J. Hay, and Y. He (2018). Geographic object-based image analysis (GEOBIA): emerging trends and future opportunities. *GIScience & Remote Sensing* 55(2): 159-182. <https://doi.org/10.1080/15481603.2018.1426092>
- Chiang, Y.-Y., S. Leyk, N. H. Nazari, S. Moghaddam and T.X. Tan (2016) Assessing the impact of graphical quality on automatic text recognition in digital maps. *Computers & Geosciences* 93, 21–35. <https://doi.org/10.1016/j.cageo.2016.04.013>
- Dam, P. (2005). Videnskabernes Selskabs kort 1768-1805 - En introduktion til såvel analog som digital brug. *HISKIS' Digitale årsskrift* 2003-2005: 36–44
- Korsgaard, P. (2006). *Kort som kilde : en håndbog om historiske kort og deres anvendelse*. Vejle (DK): Dansk Historisk Fællesråd : Sammenslutningen af Lokalarkiver.
- Jensen, R. H. and K.M. Jensen (1979). Kulturlandskabet i Borris og Sdr. Felding — En kortbladsanalyse af et Vestjysk landbrugssamfund og en dokumentation for dets udvikling. *Danish Journal of Geography* 78-79: 61–99. <https://doi.org/10.1080/00167223.1979.10649107>
- Kristensen, S.B.P., A. Reenberg and J. J. D. Peña (2009) Exploring Local Rural Landscape Changes in Denmark: A Human-Environmental Timeline Perspective. *Danish Journal of Geography* 109 (1): 47–67. <https://doi.org/10.1080/00167223.2009.10649595>
- Levin, G., G. B. Groom, S. R. Svenningsen and M. Perner (In prep) *Automated production of spatial datasets for land categories from historical maps - Method development and results for a pilot study of Danish late-1800s topographical maps*. Aarhus (DK): Aarhus University, DCE - Danish Centre for Environment and Energy.
- Leyk, S. and R. Boesch (2009) Extracting Composite Cartographic Area Features in Low-Quality Maps, *Cartography and Geographic Information Science*, 36: 71-79. <https://doi.org/10.1559/152304009787340115>

- Liu, T., P. Xu and S. Zhang (2019) A review of recent advances in scanned topographic map processing. *Neurocomputing* 328: 5–87. <https://doi.org/10.1016/j.neucom.2018.02.102>
- Lladós, J, E. Valveny, G. Sanchez and E. Marti (2002) Symbol recognition: Current advances and perspectives. In: D. Blostein and Y.-B. Kwon (eds.) *Graphics Recognition Algorithms and Applications* (Lecture Notes in Computer Science, Vol. 2390). Berlin: Springer, 104–128.
https://doi.org/10.1007/3-540-45868-9_9
- Münier, B., 2009. Landskabets udvikling siden midten af 1800-tallet. In: B. Odgaard and J. R. Rømer (eds.), *Danske Landbrugslandskaber Gennem 2000 År: Fra Digevoldinger Til Støtteordninger*. Aarhus (DK): Aarhus Universitetsforlag, 81–98.
- Quan, Y., Y. Shi, Q. Miao, and Y. Qi (2018). A Combinatorial Solution to Point Symbol Recognition. *Sensors* 18(10): E3403. <https://doi.org/10.3390/s18103403>
- Svenningsen, S.R., G. Levin and M. R. Jepsen (2015) Decrease in Danish semi-natural grassland – a social construct or a real-world change? *Danish Journal of Geography*, 115 (2): 157-166.
<https://doi.org/10.1080/00167223.2015.1070276>
- Tobler, W. (1988). Resolution, Resampling, and All That. In: H. Mounsey and R. Tomlinson (eds.), *Building Data Bases for Global Science*, London: Taylor and Francis, 129-137.