Extracting Human Settlement Footprint from Historical Topographic Map Series Using Context-Based Machine Learning

Johannes H. Uhl*, Stefan Leyk*, Yao-Yi Chiang†, Weiwei Duan†, Craig A. Knoblock†

* University of Colorado Boulder, Department of Geography, Boulder (CO), United States of America
{stefan.leyk; johannes.uhl}@colorado.edu

† University of Southern California, Spatial Sciences Institute, Los Angeles (CA), United States of America
{yaoyic; weiweidu; knoblock}@usc.edu

Keywords: Map processing; Geographic information systems; Image processing; Machine learning.

Abstract

Information extraction from historical maps represents a persistent challenge due to inferior graphical quality and large data volume in digital map archives, which can hold thousands of digitized map sheets. In this paper, we describe an approach to extract human settlement symbols in United States Geological Survey (USGS) historical topographic maps using contemporary building data as the contextual spatial layer. The presence of a building in the contemporary layer indicates a high probability that the same building can be found at that location on the historical map. We describe the design of an automatic sampling approach using these contemporary data to collect thousands of graphical examples for the symbol of interest. These graphical examples are then used for robust learning to then carry out feature extraction in the entire map. We employ a Convolutional Neural Network (LeNet) for the recognition task. Results are promising and will guide the next steps in this research to provide an unsupervised approach to extracting features from historical maps.

1 Introduction

Efficient graphics recognition of historical maps is impeded to date mainly due to issues of poor graphical quality and large data volume, which is a common problem when thousands of historical map sheets are scanned and stored in map archives. To overcome the need for user intervention and manual training in a recognition system, we are developing techniques to fully automate the process of extracting geographic information from scanned historical cartographic documents. The goal of such information extraction efforts is to make the data in these documents accessible to geospatial tools and thus for spatial-temporal analysis of landscape patterns and their changes [1]. One approach to improving recognition performance is to incorporate contextual geographic layers to make use of the fact that map series represent evolutionary documents that change in cumulative ways [2]. The concept of geographic context implies the effective use of ancillary geographic information containing the feature of interest such as gazetteers or other map series for guided graphics sampling in training a recognition model [3,4,5]. For example, it can be assumed that roads in a historical map spatially overlap or are in proximity to road segments in a contemporary geographic dataset. Thus, sampling along the contemporary road segments enables a system to collect graphic examples of road symbology in historical maps.

In this paper, we present an approach to extract building footprints and urban areas from historical sheets of the USGS topographic map series. The extraction of building footprints is particularly challenging due to their small areal extent, variations in shape, size, and spatial context. Furthermore, contextual geographic layers representing building locations from different points in time are difficult to obtain.

To overcome this challenge, we need a robust approach for learning the symbols of interest, based on large numbers of training samples, such as convolutional neural networks (CNN). CNNs have recently received considerable attention in object recognition, classification, and detection tasks in general [6]. Furthermore, machine learning techniques such as deep learning have increasingly been applied for information extraction from earth observation data [7,8,9,10,11] and this naturally projects into the idea of applying such techniques to other types of geospatial data. In this paper, we examine the use of geographic contextual data to guide graphics sampling for automatically generating training images. This approach can automatically generate thousands of training samples to allow the utilization of a CNN in a robust recognition system for building symbols and urban areas in historical map sheets.

2 Data & Methodological Approach

The USGS has scanned more than 180,000 historical map sheets and stored the entire map series in a digital archive. While urban areas in these map sheets are uniquely coloured areas, building symbols are shown with small black rectangles or polygons (Fig. 1c). We test our approach on a map sheet of Boulder, Colorado (1966) at a map scale of 1:24,000 scanned at a resolution of 500 dpi in the RGB colour space.

We use contemporary integrated land parcel/building data as the contextual spatial data layers. The temporal information of when a building has been established can be derived from the parcel data attributes. Spatially refining

these parcel boundaries with high-resolution LiDAR-derived building footprints makes it possible to create snapshots of existing buildings at different points in time. Such rich training data are available for only a selected number of counties in the U.S., including Boulder County, Colorado. Thus, we first train and test the graphics recognition system for those counties for which such contextual data exist. Once successful, we will then expand the approach to extract building symbols and urban areas from map sheets of other regions as well. This approach assumes that the learning capabilities and recognition models are robust and generalizable enough to be applied to regions for which contextual data do not exist, i.e., that building symbols can be expected to be similar across map sheets.

2.1 Aligning contextual data with historical maps

We collect training data (graphics examples) using the above-described contextual geographic layer as a constraining variable, i.e., at map locations where the feature of interest can and cannot be expected (positive and negative samples). The successful use of the contextual spatial data requires a satisfactory geographical co-registration between the scanned historical map and the contemporary building footprints. However, due to the positional uncertainty in the historical map, geometric features are often misaligned, and such positional discrepancies manifest themselves by offsets or slight rotations when compared to the contextual data (Fig. 1).

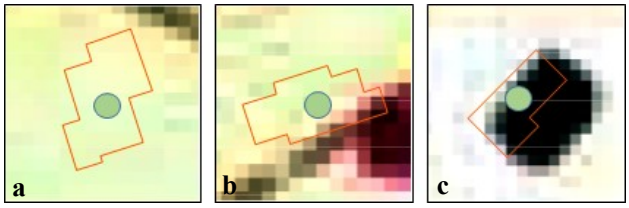

Fig. 1: Examples of positional offsets between building symbols and contextual data: (a) significant offset (map symbol outside the subimage), (b) moderate offset (partial overlap) and (c) minor offset (mostly overlapping)

Such positional uncertainty can be linked to the georeferencing process of the scanned map, distortions in the paper map document, the level of feature generalization at a given map scale, the map production process, and data acquisition techniques. The USGS preserved the coordinate pairs of ground control points (GCPs) used for georeferencing. We use this information to adjust the contextual data to reduce distortions introduced during the georeferencing process using a least squares second order polynomial transformation. Fig. 2 shows the direction and magnitude of the error vectors for an exemplary map. Fig. 3 shows the effect of a first and second order transformation on the example of alignment between railroad features. These uncertainties can be different among map sheets and thus the success of the correction may vary.

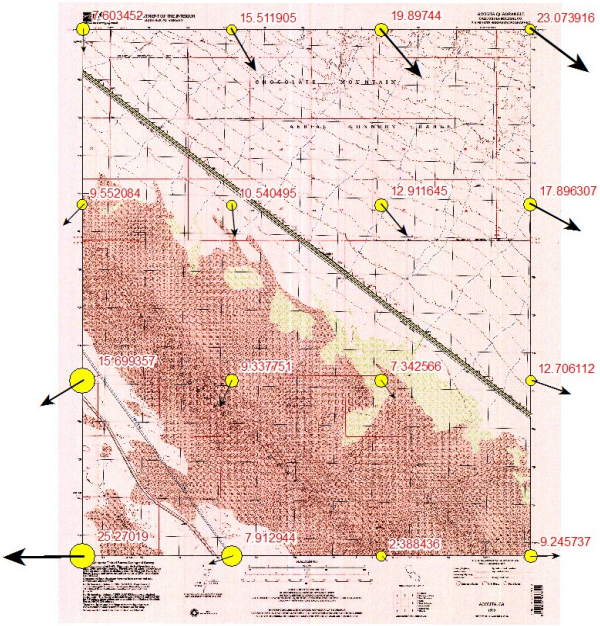

Fig. 2: Residual error vectors (direction and magnitude) for the GCPs used for georeferencing the maps obtained by least squares second order polynomial transformation.

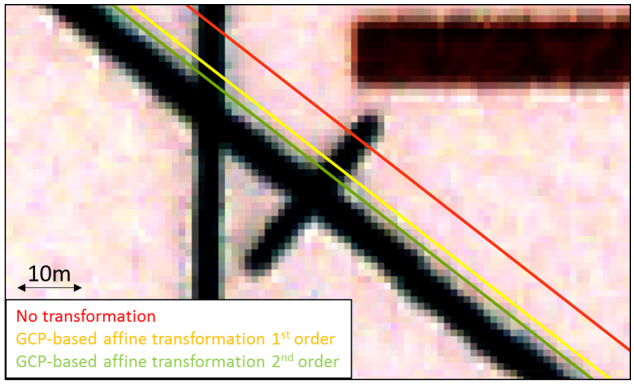

Fig. 3: Effect of the alignment of contextual data to map distortions introduced by GCP inaccuracies shown for a railroad on a USGS topographic map.

2.2 Training data creation through guided sampling

First, we clip the features in the contextual layer to the extent of a given map. Training data consists of two main classes: positive and negative. Ideally, the positive class will only contain examples of urban areas and individual buildings, which are the two subclasses of positive samples. The negative class, no buildings, consists of representative examples of anything else in the map (e.g., background, text, roads, rivers, and elevation lines). Our goal is to automatically collect positive samples from within focal windows (42x42 pixels, corresponding to 50x50 meters) centred at the centroid of each building polygon in the contextual data (by cropping sub-images of the historical map using the extents of these focal windows). However, the above-described positional

discrepancies between contextual geographic data and the map symbols result in additional challenges during training data creation and require a systematic process to identify representative and reliable training samples within the set of sub-images (Fig. 4). Therefore, we incorporate domain knowledge of the graphical representation of the subclasses (i.e., urban areas and individual buildings) in the map for the guided training process.

First, we determine training samples for the urban area positive sub-class as follows (Step 1 in Fig.4). We consider a sub-image cropped around building centroids in the contextual data a representative urban area sample if the green/red colour ratio of the dominant colour within the focal window after running a Gaussian filter and a k-means clustering colour reduction step is less than 0.8 (Fig. 5).

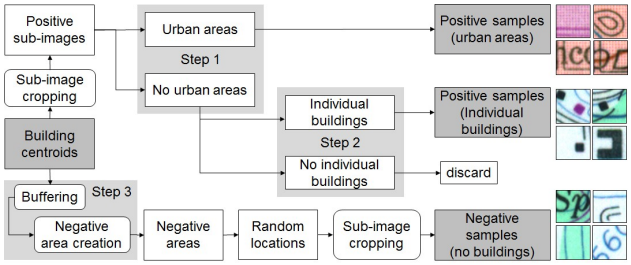

Fig. 4: Workflow of graphics sampling for training data creation.

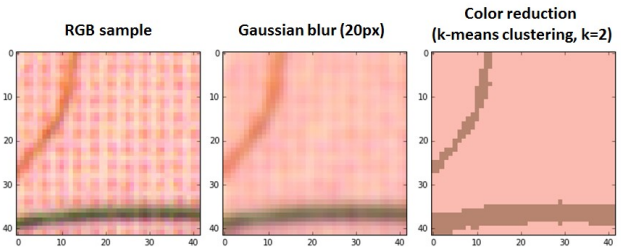

Fig. 5: Example of the image processing chain for the detection of urban training samples.

Second, we derive positive samples for the individual buildings subclass from the remaining sub-images as follows (Step 2 in Fig. 4). After applying a Gaussian filter to the grayscale sub-image to remove irrelevant noise, we emphasize dark pixels (potential building symbols) in the sub-image by inverting the colour space and using a threshold (Fig. 6). Then we use the Scale-Invariant Feature Transform (SIFT, [12]) algorithm to detect maxima in the Difference of Gaussian (DoG) scale space, which are potential keypoints. Here, we set the maximum number of potential keypoints to one. Due to the previously applied Gaussian filter (i.e., removal of sharp edges) the DoG maximum tends to be detected at the building centre. Hence, if a keypoint is detected in a given pre-processed sub-image, the system assumes the presence of a building at the keypoint location with a high confidence and labels the entire sub-image as a building training sample.

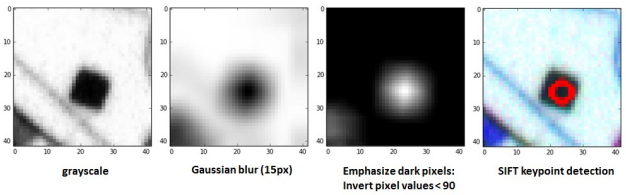

Fig. 6: Example of the processing chain for the creation of individual building training samples.

Third, the system also needs to collect negative graphics samples (Step 3 in Fig. 4). To do so, we buffer all building centroids in the contextual data by a distance of 100 meters (84 pixels). We then apply a random sampling scheme to collect negative examples by cropping sub-images at random locations within the area outside of the buffer areas.

Here, we create 10,000 negative samples (i.e., no buildings). In order to obtain a balanced set of training samples for all three classes, we oversample the graphics samples of urban area and individual buildings yielded by the training data creation process using random duplicates until an equal sample size of 10,000 is reached (see [13]).

We use t-Distributed Stochastic Neighbor Embedding (t-SNE) plots [14] to visually assess the quality of the created training data. T-SNE is a technique for reducing the dimensionality of data and is effective for the visualization of high-dimensional datasets in a 2D space arranging the features in direct neighbourhood according to their mutual similarity based on pairwise L2 distance in the feature space. In this study, we create t-SNE plots for each sub-class and rectify them using a nearest-neighbour technique to visualize labels of urban area, individual buildings, and no buildings, respectively. This results in groups or clusters of similar training samples in each of the plots (Fig. 7).

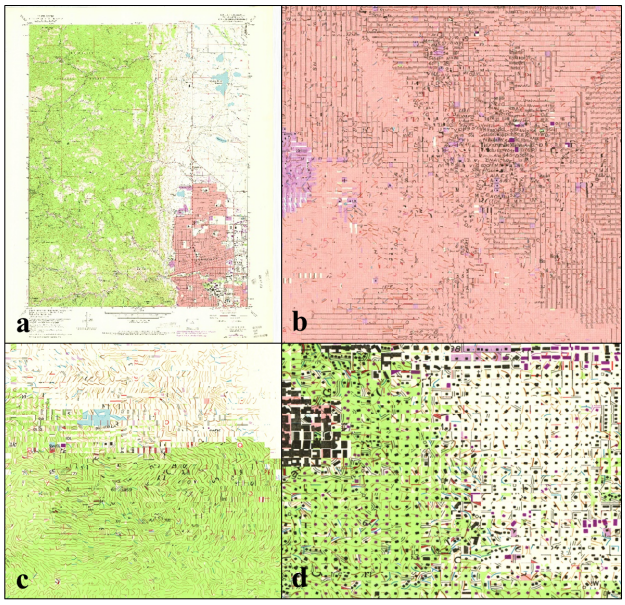

Fig. 7. a) Historic topographic map (Boulder CO, 1966), and rectified t-SNE plots of extracted training samples for b) urban area, c) no buildings, and d) individual buildings.

2.3 CNN-based feature extraction

In this study, we apply convolutional neural networks (CNN). More specifically, we use a variant of the classical LeNet architecture, which has been successfully applied for the recognition of handwritten digits ([15], see Fig. 8). Here, we apply LeNet for inference on the presence of settlements in the entire map, using the training data created in Section 2.2.

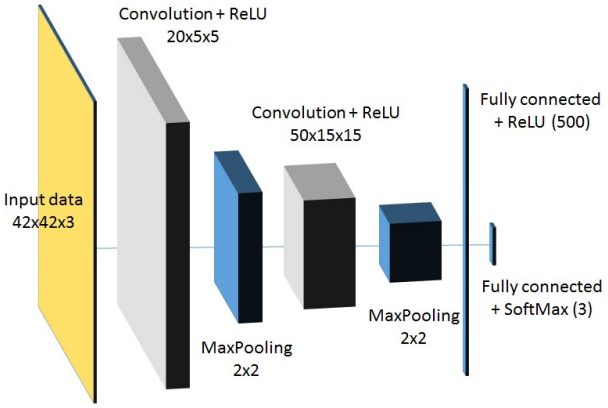

Fig. 8: Layer configuration of the LeNet CNN applied.

3 Preliminary Results

Out of the total amount of sampled data, we used 60% for model training, 20% for testing during the training phase, and held out 20% for validation. In addition, we manually digitized the features of interest within a portion of the target map to run an objective comparison during validation. Using a learning rate of 0.001 the CNN yielded an overall accuracy of 0.59 after 5,000 iterations running on a simple Intel I7 CPU with NVIDIA GeForce GT 740 GPU (training time: 14 minutes).

We used the trained CNN to predict the labels of the three classes of interest for 50x50m sub-images in the map at a given stride ***s***. As a result, the model created class score maps of spatial resolution ***s***. By assigning the class of highest score to each patch of ***s*** x ***s*** meters, we created a three-class segmentation of the map: no buildings, urban areas, and individual buildings. We processed the subset of the map for which we manually digitized reference data (Fig. 9), using a stride of two meters. Figure 10 shows the class score maps and the segmentation result.

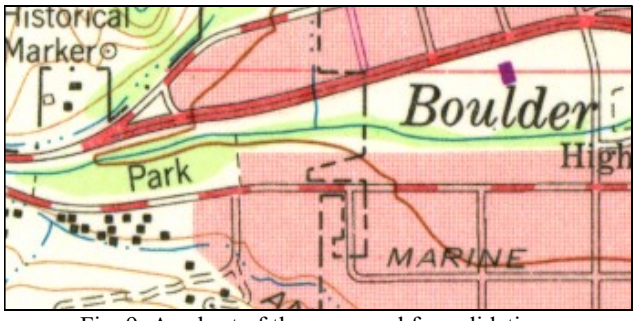

Fig. 9: A subset of the map used for validation.

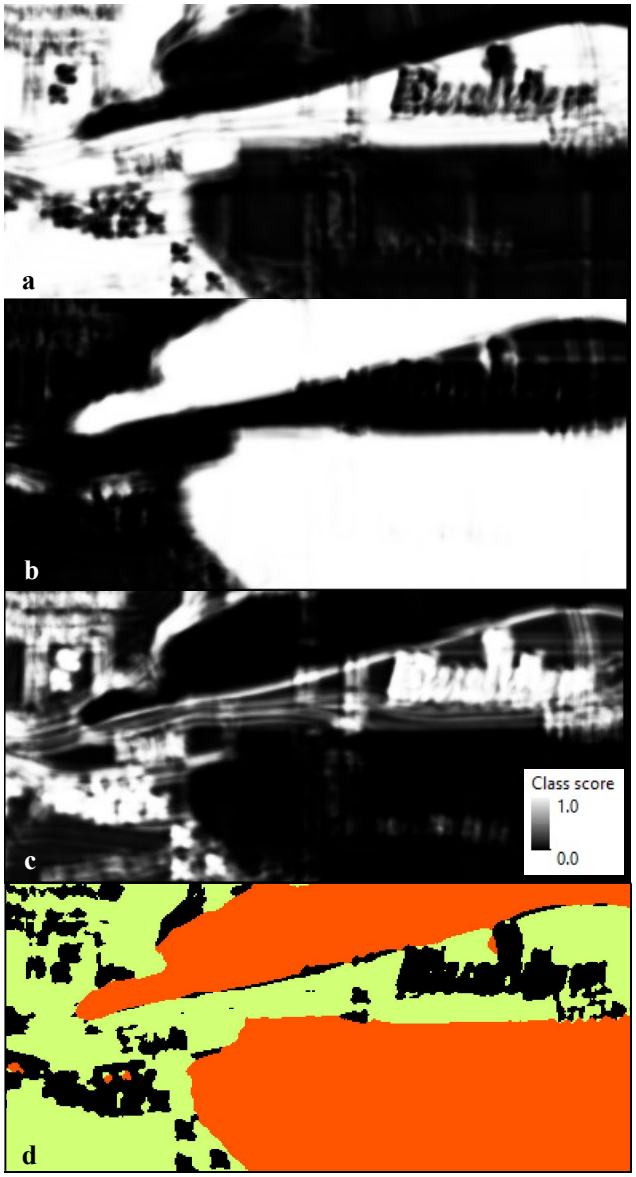

Fig. 10: Predicted score maps for (a) no buildings, (b) urban areas, (c) individual buildings, and (d) segmentation result for the map subset (stride of 2m).

As expected, the validation of the CNN classifier using the 20% of the samples that we held back during the training process provided highly biased accuracy measures, probably because the negative sample was not representative of the underlying variability of the no-buildings signature in a considered map. Therefore, we used the manually digitized reference data from the map subset (Fig. 9) as the validation base and found more objective results in calculating the confusion matrix (Fig. 11). This matrix equated to the following overall accuracy measures:

- Percentage of correctly classified (PCC) = 0.81
- Kappa index = 0.66
- Normalized Mutual Information (NMI) = 0.46

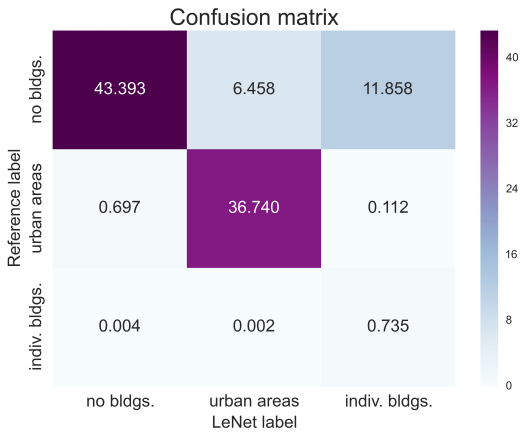

Fig. 11: Confusion matrix of the CNN validation (in %) using digitized data from a portion of the test map.

Class-specific accuracy metrics (i.e., precision and recall) for all three target classes are shown in Table 1:

Class	Precision	Recall
No buildings	0.98	0.70
Urban area	0.85	0.98
Individual buildings	0.06	0.99

Table 1: Class-specific accuracy metrics derived from the confusion matrix in Fig. 11.

As can be seen by comparing Figs. 9 and 10d, there were numerous locations of false positive labels for individual buildings, which means the precision of the individual building class was low. For example, numerous building locations were labeled within text labels and along roads and railroads. However, the high recall measure confirms that most buildings in the map were included in the predicted set of labels. These results indicate a high sensitivity to building signatures but poor discriminatory power regarding features of similar color and shape characteristics.

For the recognition task on the entire map, we predicted the labels of the three classes of interest for 50x50m sub-images at a stride of 20m (Fig. 12). Visually comparing the result with the original map (Fig. 7a), we found that the urban area (red) was extracted well by the CNN, but identified high proportions of false positives in the individual buildings class as confirmed by the confusion matrix and the map subset shown in Figs. 9 and 10.

4 Discussion

While the relatively high rate of false positives in the individual building class deserves more attention for improving the system, the high recall measure for the same class is encouraging as are all other class-specific accuracy measures. The current limitations in form of low precision for the individual building class can be linked to the sampling strategy (oversampling) and the high rate of confusion of map symbols belonging to the black map layer. The system currently fails to consider size or shape properties of the target symbol (e.g., isolation/connectedness, contiguity; part of larger object or not).

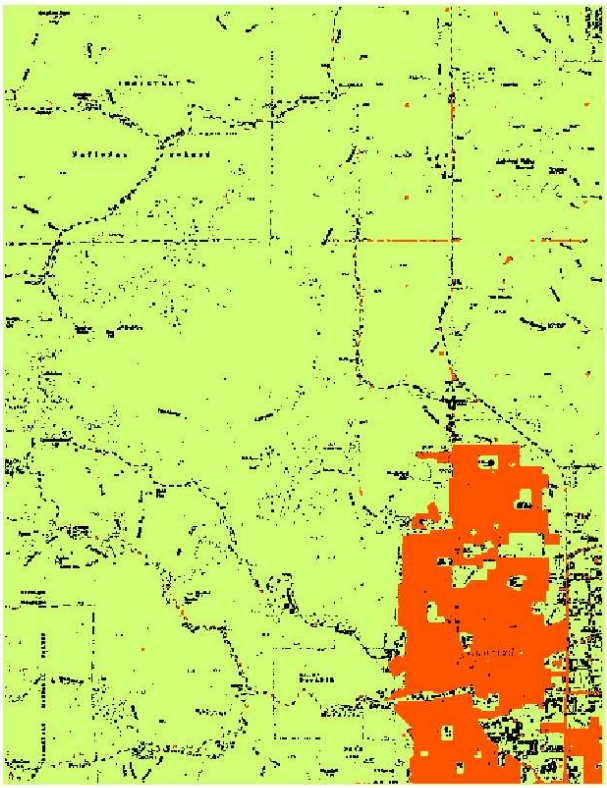

Fig. 12: Predicted settlement locations for the historical map (Boulder, Colorado, 1966): Black pixels are individual building locations with obvious high false positive rate; red pixels are urban areas.

However, the described results indicate that the segmented class of individual buildings has the benefit that most building symbols in a map page were extracted (recall = 0.99). This can be seen as a successful first extraction step based on which subsequent image processing steps can be applied for further refinement to systematically narrow down the most likely locations of building symbols and thus filter out those false positives. For example, such subsequent post-processing will consider different options including the testing of dimensional measures (length, area, number of connected components) or geometries of underlying map contents that falsely contain detected single building labels. Based on such measures, the system will be able to identify the most likely candidate labels for individual buildings among those currently classified.

In the next steps, we will apply more complex CNNs with greater depth (e.g., VGGNet) in a more performant computational environment, create more abundant and more representative negative training samples, examine the effects of the used sampling scheme, and extend this extraction task to larger amounts of map sheets. This could include the use of contextual road network data to systematically increase the proportion of road samples in the negative class. In addition to that, the CNN-based image segmentation result could be

improved by using further geographic context layers or by applying post-segmentation refinement techniques. Geographic contextual information derived from existing, remote-sensing derived land cover data products can also be used for thematic validation and calibration of the proposed method.

Longer-term problems to solve will include those of domain adaptation and transfer learning [16] in the context of applying the extraction procedure to a large number of map sheets. Training data as shown here is available for only a few counties in the U.S. Making inference for map sheets of the entire U.S. will result in differences in spectral characteristics of scanned map documents and map symbology. Thus, training and inference will take place in different data distributions due to variations in symbology and map properties, and the system needs to be robust against such differences.

Acknowledgements

This material is based on research sponsored in part by the National Science Foundation under Grant Nos. IIS 1563933 (to the University of Colorado at Boulder) and IIS 1564164 (to the University of Southern California).

References

[1] Y.-Y. Chiang, S. Leyk, and C.A. Knoblock, "A survey of digital map processing techniques", *ACM Computing Surveys (CSUR)* **47**.1: 1, (2014).

[2] S. Leyk and Y.-Y. Chiang "Information extraction of hydrographic features from historical map archives using the concept of geographic context", *Proceedings of AutoCarto 2016*, Albuquerque, NM, U.S.A., Sept. 14-16, 2016, pp. 100-110, (2016).

[3] Y.-Y. Chiang, and S. Leyk, "Exploiting online gazetteer for fully automatic extraction of cartographic symbols", *Proceedings of the 27th International Cartographic Conference ICC 2015*, Rio de Janeiro, Brazil, August 23-28, (2015).

[4] Y.-Y. Chiang, S. Leyk, N.H. Nazari, and S. Moghaddam, "Assessing impact of graphical quality on automatic text recognition in digital maps", *Computers and Geosciences* **93**, pp. 21-35, (2016).

[5] R. Yu, Z. Luo, and Y.-Y. Chiang, "Recognizing text on historical maps using maps from multiple time periods", *Proceedings of the 23rd International Conference on Pattern Recognition, IEEE*, Cancun, Mexico (to appear).

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, **521**(7553), pp. 436-444, (2015).

[7] F. Maire, L. Mejias, and A. Hodgson, "A convolutional neural network for automatic analysis of aerial imagery", In *IEEE International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2014,* pp. 1-8, (2014).

[8] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks", *arXiv preprint arXiv:1508.00092*, (2015).

[9] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using ImageNet pretrained networks", *IEEE Geoscience and Remote Sensing Letters*, **13**(1), pp. 105-109, (2016).

[10] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification", *IEEE Transactions on Geoscience and Remote Sensing*, **54**(3), pp. 1349-1362, (2016).

[11] G. J. Scott, M.R. England, W.A. Starms, R.A. Marcum, & C.H. Davis, "Training deep convolutional neural networks for land–cover classification of high-resolution imagery", *IEEE Geoscience and Remote Sensing Letters*, **14**(4), pp. 549-553, (2017).

[12] D.G. Lowe, "Object recognition from local scale-invariant features". *Proceedings of the seventh IEEE International Conference on Computer Vision* **2**, pp. 1150-1157 (1999).

[13] P. Hensman, and D. Masko, "The impact of imbalanced training data for convolutional neural networks", *Degree Project in Computer Science, KTH Royal Institute of Technology*, Stockholm, Sweden, (2015).

[14] L. V. D. Maaten, and G. Hinton, "Visualizing data using t-SNE", *Journal of Machine Learning Research*, **9**(Nov), pp. 2579-2605, (2008).

[15] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel, "Backpropagation applied to handwritten zip code recognition", *Neural Computation*, **1**(4), pp. 541-551, (1989).

[16] G.J. Maclaurin and S. Leyk, "Temporal replication of the national land cover database using active machine learning", *GIScience & Remote Sensing*, **53**:6, pp. 759-777, (2016).