SEMANTIC SEGMENTATION OF LISS-4 SATELLITE IMAGERY FOR MAPPING BUILT-UP AREAS USING DEEP LEARNING CONVOLUTIONAL NEURAL NETWORK

Pankaj Bodani[1], Kriti Rastogi[2] and Ujjwal K. Gupta[3]
Space Applications Centre, Jodhpur Tekra, Ahmedabad 380015, INDIA,
Email: [1]pankajb@sac.isro.gov.in, [2]kritirastogi@sac.isro.gov.in, [3]ujjwal_gupta@sac.isro.gov.in

KEY WORDS: urban growth modelling, ortho-imagery, deep learning, artificial intelligence

ABSTRACT: Mapping built-up areas is important for city growth monitoring and calibrating models for city growth forecasting. Aerial and satellite ortho-imagery is popularly used for this purpose. Knowledge based, semi-automated approach for this task is challenging and requires significant human intervention and application of subjective expertise for iterative refinement of rule sets used for semantic segmentation. This is due to high heterogeneity in shape, density and composition of built-up areas. Further, methods that rely on calculating tonal and textural indices fail to generalize well and require manual tuning of thresholds for different regions for accurate segmentation. This paper investigates a fully automated deep learning approach based on a convolutional neural network for this task. This eliminates the need of human expertise for defining complex rulesets for semantic segmentation and does not require subjective retuning of deterministic thresholds for different regions. Specifically, this paper discusses design, training and performance evaluation of a deep convolutional network for identifying built-up land areas in LISS-4 satellite ortho-imagery. The network was trained and evaluated on multi-spectral LISS-4 satellite ortho-imagery covering the central core and peripheral developing areas of a city.

1. INTRODUCTION

Over the years, SLEUTH model has been extensively used for forecasting urban growth patterns (Chaudhuri and Clarke, 2013). Inputs required for calibrating growth forecasting models include historical maps of built-up areas. Satellite and aerial ortho-imagery is popularly used for preparing these maps. Manual approach to mapping built-up areas using visual interpretation of satellite imagery is costly to scale up when wide geographical regions of extended historical time frame are required to be mapped. Extensive research and development has been carried out towards automating mapping of built-up areas. A direct approach to identifying built-up areas relies on normalized difference built-up index (Zha and Ni, 2003), but the reliability of this method is lowered in mapping peripheral urban areas where barren or fallow land with low moisture content is widespread. Another approach that is extensively used is application of object based classification using hand-crafted features (Zhang and Wang, 2014). This involves creation of rules for defining characteristics of built-up areas or individual built-up structures as they visually appear in the images. The characteristics that are commonly used, represent spectral properties, shape and texture. However, due to high variability in texture and spectral properties of built-up areas as show in Figure 1, it is difficult to design robust rule-sets that generalize well. Therefore, in this paper we investigate a fully automated deep learning approach based on convolutional neural network (ConvNet) for mapping built-up areas which can accurately map built-up areas even in areas with dry or fallow land. This approach has an additional advantage that it does not require feature engineering or manual tuning of deterministic thresholds for building an accurate segmentation model.

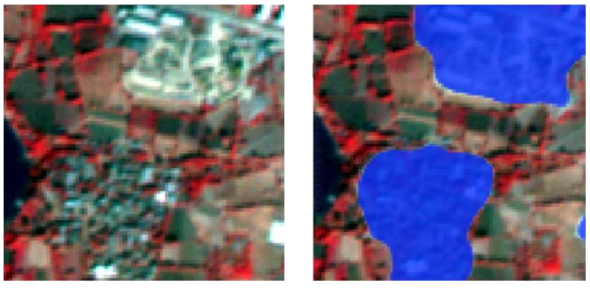


Figure 1 Image Demonstrating Visual Heterogeneity of Built-up Areas

2. LITERATURE REVIEW

A variety of ConvNet architectures (Long et al., 2014; Liu et al., 2015; Noh et al., 2015; Badrinarayanan et al., 2015) have demonstrated success in image segmentation benchmark datasets (Everingham et al.,2010), (Song et al., 2015), (Geiger et al., 2012). The success of ConvNets has also been adapted to over-head sub-meter resolution aerial imagery

and digital surface models as demonstrated by top performing models in ISPRS Vaihingen and Potsdam 2D semantic labelling challenges (Potsdam 2017; Vaihingen, 2017). The top performing ConvNets in these challenges are specifically designed trained for identifying footprints of individual objects by using very fine resolution (less than 1m) ortho-imagery. Footprint level maps can then be used to produce higher level built-up area maps. However, historical availability of such imagery is usually very limited and temporally inconsistent. This makes it difficult to use these images for generating sets of long-term historical built-up area maps that are suitable for calibrating urban growth models. Therefore, this paper discusses a direct segmentation approach using multispectral LISS-4 satellite ortho-imagery of 5.8m resolution which has consistent historical availability in Indian region.

3. ARCHITECTURE

The architecture of our network is shown in Figure 2. The architecture is based on SegNet encoder-decoder architecture (Badrinarayanan et al., 2015) and has no fully connected layers. Each layer group in the encoding part comprises two 3x3 2D convolutional layers (Conv) with rectified linear unit (ReLU) activation. Each convolution layer in the network is followed by a spatial max pooling (Pooling) layer which consists of pooling units that output the maximum value from each of 2x2 cluster of neuron activations at the prior layer. Each layer group of the encoding network thus produces feature maps in the form of 3D activation arrays that are fed to the next encoding layer group. Progressive down sampling of feature maps by each layer of the encoder network results in context aggregation properties with relatively fewer parameters to train when compared to other popular architectures (Long et al., 2014), (Liu and Berg, 2015), (Noh and Han, 2015). The layer groups in the decoding part of the network comprise a concatenation and a 2x2 up-sampling layer followed by two 3x3 convolution layers. Each layer group in the decoder part thus progressively densifies the feature maps by concatenating with the pooling indices from the corresponding input layer group to ensure localization of features. The final sigmoid layer performs pixel wise classification.

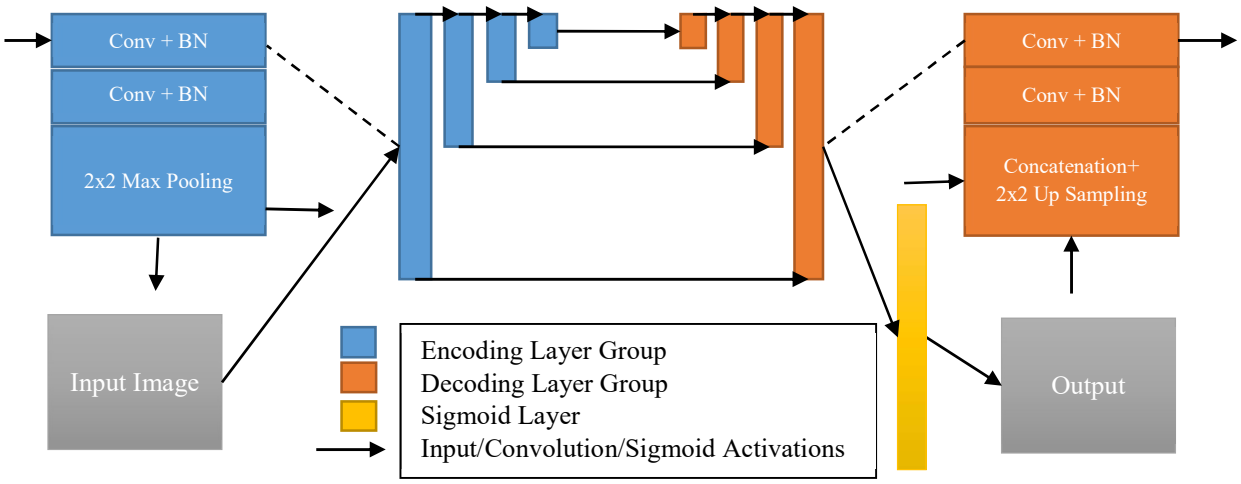


Figure 2 Network Architecture

Batch normalization (BN) (Ioffe and Szegedy, 2015) layers were used for accelerating the training of the network and improving regularization. The dimensions of the convolution window and feature map depths at different pooling levels are given in Table 1.

Table 1 Layer Parameters

Pooling Level	No pooling	First pooling	Second pooling	Third pooling	Fourth Pooling
Conv. Depth	16	32	64	128	128
Feature Map Size	1024x1024	512x512	256x256	128x128	64x64

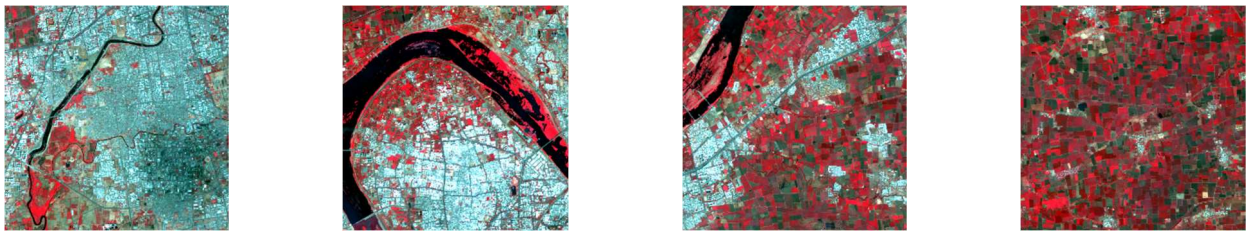
4. DATA

Training and test data consisted of manually labelled 6144x3072 pixel, 3 band, 5m resolution (resampled from 5.8m resolution) LISS-4 image each. Spectral specifications of bands are given in Table 2. Both images covered non-

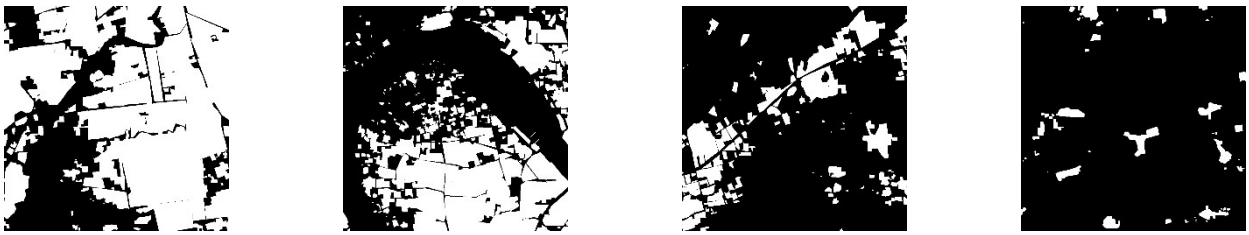
overlapping regions of Surat city and its peripheral area with suburban development. The training images were chosen such to cover entire spectrum of built-up densities from dense core of the city to outer peripheral areas with very sparse built-up area patches. Anderson's land use (man-made) and land cover (natural or semi-natural) classification system (Anderson et al., 1976) was used for manual labelling of built-up areas. Visually discernible major roads were not included in the marked areas. Boundaries of built-up areas were digitized as polygons using Quantum GIS (QGIS) software. The polygons were then rasterized using Geospatial Data Abstraction Library (GDAL) to generate a binary mask. The resulting raster mask was split into tiles of 1024x1024 pixel images. Similarly, the LISS-4 input image was aligned with rasterized mask and similarly split into 1024x1024 pixel images. Three 1024x1024 pixel images from the training data were used for cross-validation of hyper-parameters.

Table 2 LISS-4 Spectral Specification

Band Name	Band width (from – to) [µm]	Code	Maximal resolution [m]
MX Mode 2	0.52 - 0.59	GREEN	5.8
MX Mode 1	0.62 - 0.68	RED	5.8
MX Mode 4	0.77 - 0.86	NIR	5.8



a. training images



b. manually mapped built-up areas

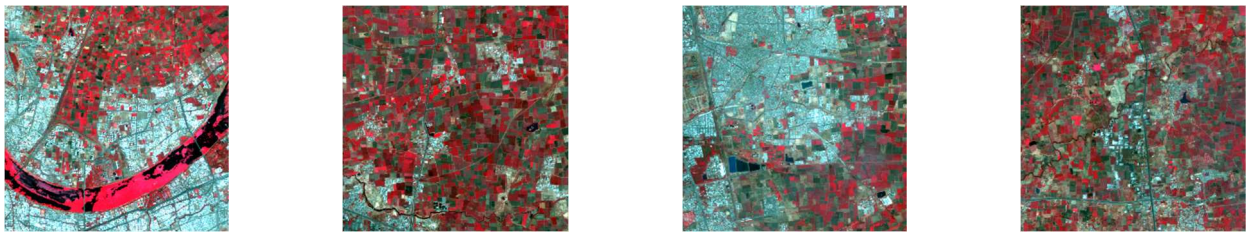
Figure 3 Training Data Samples

5. EXPERIMENT AND RESULTS

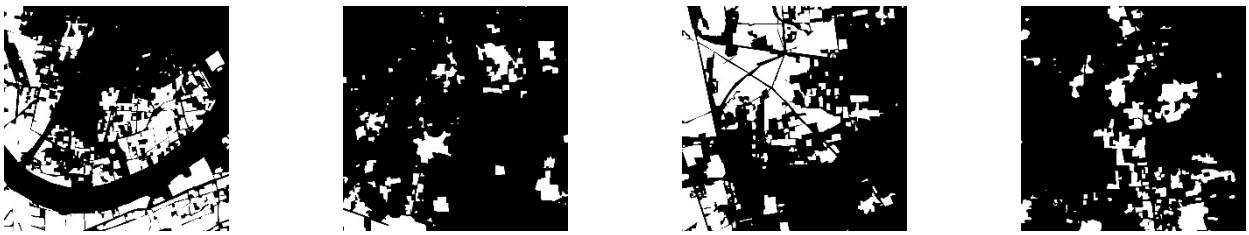
The training and test LISS-4 images were mean centered by subtracting band-wise mean across all images in the training set and dividing by band-wise standard deviation. The network was trained for 100 epochs using Adam optimizer (Kingma and Ba, 2014) with learning rate of 0.0001 and without learning rate decay. Weights that resulted in best cross-validation performance were used for performing segmentation of test images. The results of the experiments are summarized in Table 3. Comparison of trained network output and manually segmented images is shown in Figure 4. The trained network was able to accurately exclude extremely dry sandy areas (which are spectrally similar to large built-up structures) as shown in Figure 5. Also, as shown in Figure 6, the network also demonstrated very good generalization when tested on input images of Jabalpur which were not a part of training, cross-validation or test region and significantly differed from the training set in terms of spectral and textural mix and distribution.

Table 3 Results

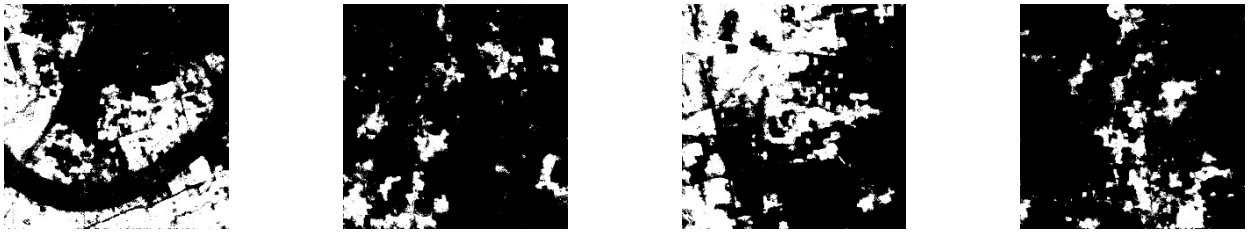
True Positive	True Negative	False Positive	False Negative	Accuracy	Kappa
40.6%	52.3%	4.7%	2.4%	92.9%	0.856



a. test images

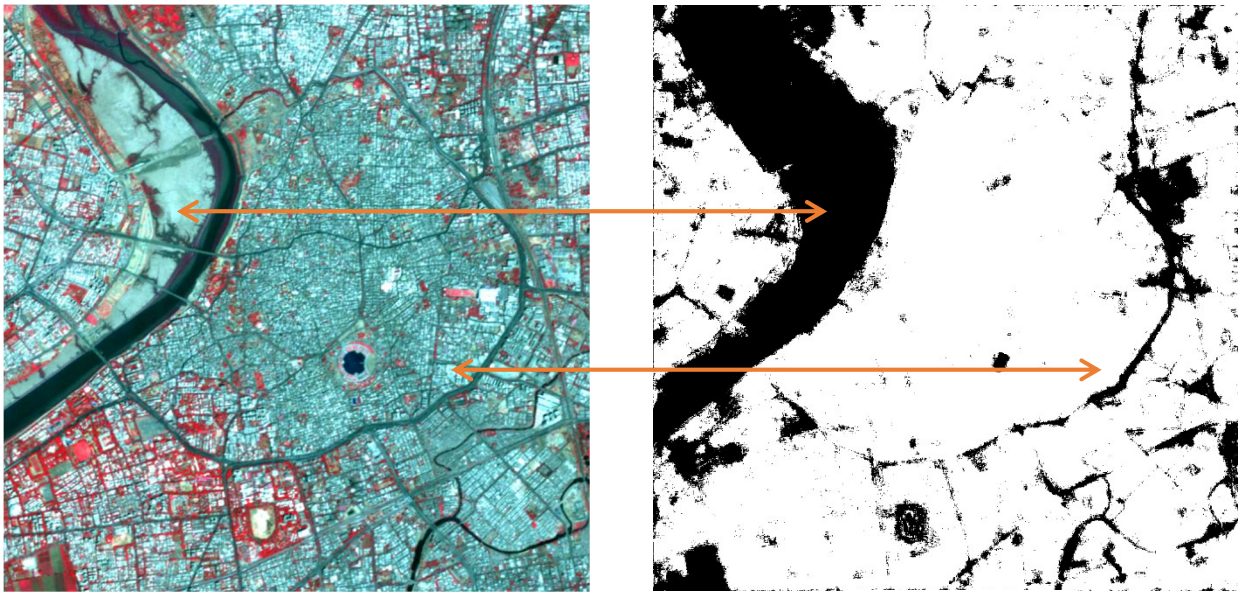


b. manually mapped built-up areas



c. network output

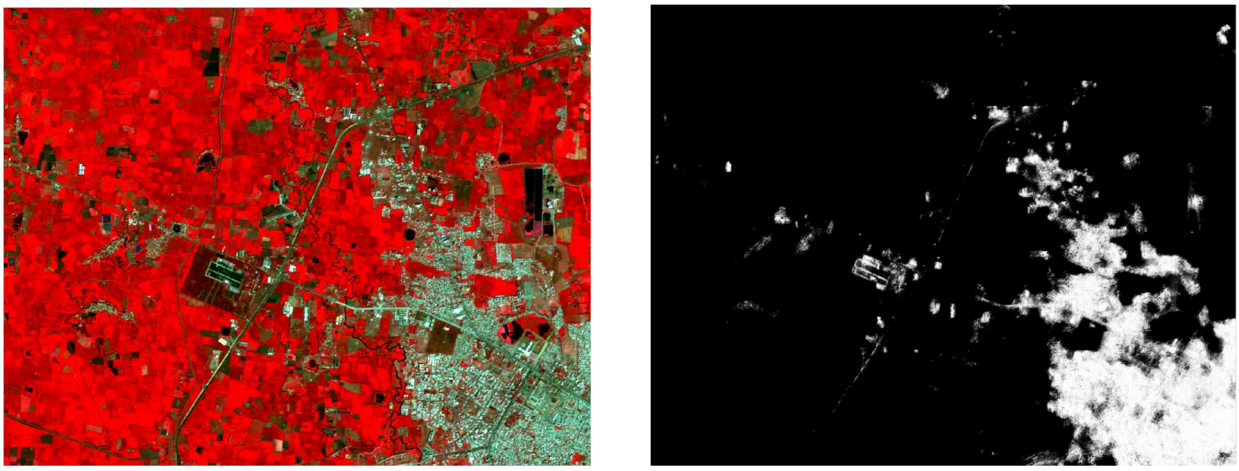
Figure 4 Segmented Maps of Built-up Areas



a. input image

b. network output

Figure 5 Discrimination of Spectrally Similar Pixels



a. input image b. network output

Figure 6 Generalization Test

ACKNOWLEDGEMENT

The authors would like to thank Shri. Tapan Misra (Director, SAC), Dr. Rajkumar (Dy. Director, EPSA), Shri. Shashikant Sharma (Group Head, VRG) and Dr. Markand Oza (Head, CGDD) at Space Applications Centre for their guidance and institutional support in carrying out this work.

REFERENCES

Anderson, J.R., Hardy, E.E.,, Roach, J.T., and Witmer, R.E., 1976. A land use and land cover classification system for use with remote sensor data. USGS Publications Warehouse, abs/1412.7062.

Badrinarayanan, V., Kendall, A., and Cipolla, R., 2015. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. CoRR, abs/1511.00561.

Barron, J.T., 2017. Continuously Differentiable Exponential Linear Units. CoRR, abs/1704.07483.

Chaudhuri, G., and Clarke, K., 2013. The SLEUTH land use change model: A review. International Journal of Environmental Resource Research, 1, pp.88–105.

Chen, L.-C., et al., 2014. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. CoRR, abs/1412.7062.

Everingham, M., et al., 2010. The Pascal Visual Object Classes (VOC) Challenge. International journal of computer vision, 88(2), pp.303–338.

Geiger, A., Lenz, P., and Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012. pp.3354–3361.

Ioffe, S., and Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. CoRR, abs/1502.03167.

Kingma, D., and Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412. 6980.

Liu, W., Rabinovich, A., and Berg, A.C., 2015. ParseNet: Looking Wider to See Better. CoRR, abs/1506.04579.

Long, J., Shelhamer, E., and Darrell, T., 2014. Fully Convolutional Networks for Semantic Segmentation. CoRR, abs/1411.4038.

Noh, H., Hong, S., and Han, B., 2015. Learning Deconvolution Network for Semantic Segmentation. CoRR, abs/1505.04366.

Potsdam: 2D Labelling challenge, Retrieved August 20, 2017, from http://www2.isprs.org/potsdam-2d-semantic-labeling-contest.html.

Song, S., Lichtenberg, S.P., and Xiao, J., 2015. SUN RGB-D: A RGB-D scene understanding benchmark suite. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp.567–576.

Vaihingen: 2D Labelling challenge, Retrieved August 20, 2017, from http://www2.isprs.org/vaihingen-2d-semantic-labeling-contest.html.

Zha, Y., Gao, J., and Ni, S., 2003. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. International journal of remote sensing, 24(3), pp.583–594.

Zhang, J., Li, P., and Wang, J., 2014. Urban Built-Up Area Extraction from Landsat TM/ETM+ Images Using Spectral Information and Multivariate Texture. Remote Sensing, 6(8), pp.7339–7359.