CLASSIFICATION OF URBAN AERIAL DATA BASED ON PIXEL LABELLING WITH DEEP CONVOLUTIONAL NEURAL NETWORKS AND LOGISTIC REGRESSION

W.Yao [a*], P. Poleswki[a], P. Krzystek [a]
[a]Munich University of Applied sciences, 80333 Munich, Germany - (yao, poleswki, krzystek)@hm.edu

Working Group VII/4

KEY WORDS: aerial data, urban areas, evidence combination, object classification, deep feature learning

ABSTRACT:
The recent success of deep convolutional neural networks (CNN) on a large number of applications can be attributed to large amounts of available training data and increasing computing power. In this paper, a semantic pixel labelling scheme for urban areas using multi-resolution CNN and hand-crafted spatial-spectral features of airborne remotely sensed data is presented. Both CNN and hand-crafted features are applied to image/DSM patches to produce per-pixel class probabilities with a L_1-norm regularized logistical regression classifier. The evidence theory infers a degree of belief for pixel labelling from different sources to smooth regions by handling the conflicts present in the both classifiers while reducing the uncertainty. The aerial data used in this study were provided by ISPRS as benchmark datasets for 2D semantic labelling tasks in urban areas, which consists of two data sources from LiDAR and color infrared camera. The test sites are parts of a city in Germany which is assumed to consist of typical object classes including impervious surfaces, trees, buildings, low vegetation, vehicles and clutter. The evaluation is based on the computation of pixel-based confusion matrices by random sampling. The performance of the strategy with respect to scene characteristics and method combination strategies is analyzed and discussed. The competitive classification accuracy could be not only explained by the nature of input data sources: e.g. the above-ground height of nDSM highlight the vertical dimension of houses, trees even cars and the near-infrared spectrum indicates vegetation, but also attributed to decision-level fusion of CNN's texture-based approach with multi-channel spatial-spectral hand-crafted features based on the evidence combination theory.

1. INTRODUCTION

Object classification analysis is a very important topic in urban remote sensing. The results of such research are appealing for a wide range of data modeling tasks across diverse applications including city mapping, urban environment assessment and road inventory. While digital cameras are frequently adopted to characterize the urban structure and land cover distribution, aerial laser scanner (ALS) is increasingly used to directly acquire dense 3D urban and topographical information. Nowadays, due to the availability of large amounts of labeled data and powerful computers, the success of deep learning methods is already confirmed on diverse applications. Particularly, the interest for convolutional neural networks (CNN) has been growing very fast in the last few years, because of their impressive results (Krizhevsky et al., 2012; Castelluccio et al., 2015) in a series of challenging problems involving image classification and retrieval.

Several researchers have applied the image annotation scheme to overhead imagery. Early days started with predicting discrete class label for each pixel by using feature vector (Benediktsson et., 1990; Bischof et al., 1992). However, due to lack of high-resolution data most of these applications were mainly limited to land cover classification. Over recent years, with high resolution data there is an opportunity for fine-grained classification such as roads, buildings and cars using sophisticated features and machine learning algorithms [Yao et al. 2012, Niemeyer et al., 2011]. In the computer vision area, CNN features have been shown to outperform conventional hand-crafted features in visual recognition tasks such as image classification (Razavian et al., 2014) and object detection (Girshick et al., 2014), making it among the most promising architectures for vision applications. It seems that CNNs roughly mimic the nature of the mammalian visual cortex and exploit the strong spatially local correlation present in natural images. A deep CNN that consists of multiple layers of small neuron collections offers an alternative efficient approach to learn visual patterns directly from raw pixels.

Recently, Razavian et al. (2014) showed that superior results compared to highly tuned state-of-the--art systems in visual classification tasks on various data sets can be achieved by training a linear SVM classifier on CNN feature representation. Logistic regression is a widely used technique in statistics due to its close relations to other classifiers, such as Support Vector Machine and AdaBoost and great robustness. These classifiers are well-studied and have been shown to achieve good generalization capability in practice.

For semantic labelling in aerial imagery, Gerke (2014) used super-pixel features and CRFs for building detection in aerial imagery. A previous study on the ISPRS benchmark datasets of Wei et al. (2012) used hand-crafted height textures from ALS data and multi-spectral features for tree and vehicle classification by pixel-level labelling based on AdaBoost. They highlighted the most important features for classification of tree covers and vehicles, such as NDVI, LiDAR intensity, and planarity, which lead to a great accuracy to be validated and achieved. The combination of CNNs and other classifiers was previously applied to semantic labelling. In the work of Farabet et al. (2013) a multi-scale CNN model is used for dense classification and labelling in street scenes using a defined graph cut model over super-pixels. Apart from sequential excitation of multiple classifiers in order, Saeidi et al. (2014) used evidence fusion theory to extract multisensory features for land-cover mapping. The ability to generate auxiliary information like certainty, conflict, and maximum probability maps for better visual understanding of auxiliary information the decision process makes it more reliable for practice.

The major goal of this work is to perform a workflow for semantic labelling in city areas using multi-spectral aerial imagery and DSM, which is based on combining a CNNs image categorization scheme with conventional pixel-based

*Corresponding author

classification using spatial-spectral features. We propose an effective strategy to address urban scene classification based on pixel labeling. The CNNs method builds upon categorizing local image patch centered at each pixel by sliding the window forward over entire image. As a post-processing step, a decision-level fusion approach is applied to combine label probabilities obtained from both classifiers. The evidence fusion framework with Dempster–Shafer theory combines evidence from different sources and arrive at a degree of belief that takes into account all the available evidence and can be expected to boost the classification accuracy. The aerial data used in this study were provided by ISPRS as benchmark datasets for 2D semantic labelling tasks in urban areas. The evaluation is based on the computation of pixel-based confusion matrices. Based on random sampling or cross-validation the approach is evaluated to give us the possibility to discuss and conclude the performance of the strategy with respect to different scene characteristics and method coordination.

In this paper, we explore the possibilities of using a coherently combined approach for semantic labelling of overhead imagery: deep learned features complemented by hand-crafted features. We also demonstrate the utility of combining both approaches. Subsequent sections detail the experiments and our results.

2. METHOD

In this section, we present the framework for automated pixel classification in high-resolution aerial images. We first introduce the convolutional neural networks adopted for dense feature extraction and learning. We then discuss how we complement CNN features with hand-crafted features to further realize the classification. Finally, we introduce the concept of evidence fusion to refine final pixel labeling results. An overview of the proposed semantic pixel labeling framework is illustrated in Fig. 1.

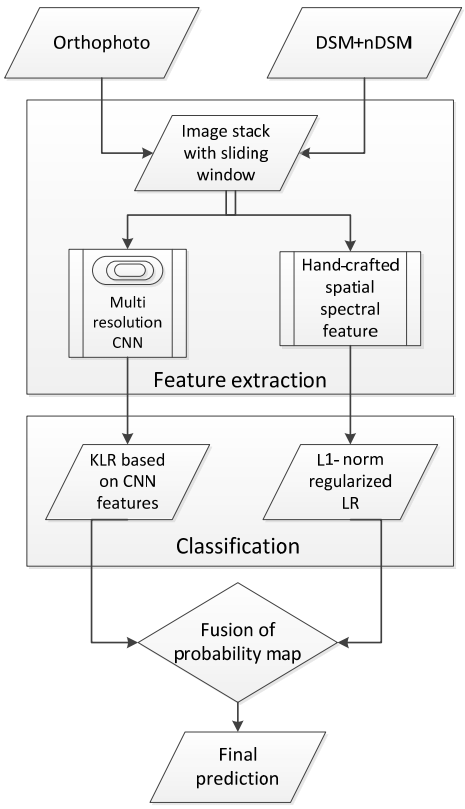

Figure 1. Workflow for the proposed pixel labelling framework.

2.1 Pixel classification with convolutional neural network

The approach consists of several steps: in the first part, the approach relies on deep learning method based on CNN feature learning and classification. We follow the work of Farabet et al. (2013) by applying a multi-class classifier on the CNNs feature representation, which is learned by supervised training of a CNN discarding the fully-connected (fc) layers. The deep CNN network usually consists of several convolutional layers, which are placed alternatively between contrast normalization and max-pooling layers. Each convolutional layer computes the convolutions between the input and a set of filters. The activation function (rectified linear unit - ReLU) performs a non-linear transformation while the max-pooling layer subsamples the output.

In fact, CNNs can also be used to directly distinguish multiple urban classes. The probability distribution over k class labels is given by feeding the output of the last fc layer to a k-way soft-max layer in a supervised way. However, recent study reported that training a margin classifier on CNN feature representation outperformed highly tuned state-of the-art systems in many vision tasks. Therefore, in this work we adopted the strategy of interpreting CNNs features by L_1-norm regularized logistic regression (LR) classifier to estimate the class labeling probability. LR was chosen as classifier due to its robustness and efficiency against high-dimensional features and multi-class task. The logistic regression solves the optimization problem:

$$\arg\min_{w_r} \sum_{i=1}^{m} \log(1+\exp(-y_i w_r^{\mathrm{T}} x_i)) \qquad (1)$$

where w_r is the weight vector to be estimated. $\{(x_i, y_i)\}^m$ is training data set, $x_i \in \mathbb{R}^d$ represents CNN feature vector, $y_i = 1$ if the class label of x_i is the same as r, m is the number of training samples, d is the dimension of feature vector x_i and $r \in \{1,2,\ldots,5\}$, which corresponds to five object class labels, respectively. We introduce L_1-norm regularization on the weight vector w_r in order to avoid over-fitting. Given a test sample x_t and the learned weight vectors $\{w_1,\ldots,w_5\}$, the probability that x_t belongs to class r is given by

$$P(y_t = r \,|\, x_t) = \frac{1}{Z}\left(\frac{1}{1+\exp(-w_t^{\mathrm{T}} x_t)}\right) \qquad (2)$$

where Z is a normalization factor used to result in a probability distribution.

Additionally, we applied a coarse-to-fine strategy to train the CNNs models in order to correctly classify image/DSM details at different scales; we train multi-resolution CNN models with input image patch of multiple sizes and concatenate the output to build the single feature vector for each center pixel. CNN models with three different image patch sizes: 21x21, 37x37 and 65x65. If we take CNN model with 64x64 pixel input patch as example, the first convolutional layer filters the 64x64x5 image patch consisting of orthophotos, DSM and nDSM with 32 kernels of size 4x4x5 with a stride of 1 pixel. Then, the second convolutional layer takes the output of the first convolutional layer as input and filters it with 64 kernels of size 4x4x32, and so on until the last convolutional layer is reached. To train the CNN models respectively with 37x37 and 65x65 pixels patch size, we simply replace the input patch to the first convolutional layer by 9x9x5 and 17x17x5, respectively. All other parameter settings remain the same as before.
Given the test image and the trained CNN model, we can extract the feature map from the last convolutional layer by vectorising CNN features and concatenating them into a single feature

vector. Then, different object classes are classified by applying logistic regression weights. For validation, we need a sliding window approach to propagate the CNN structure once to the whole test image, while the time needed for CNN feature extraction is significantly reduced due to the synchronization with validation.

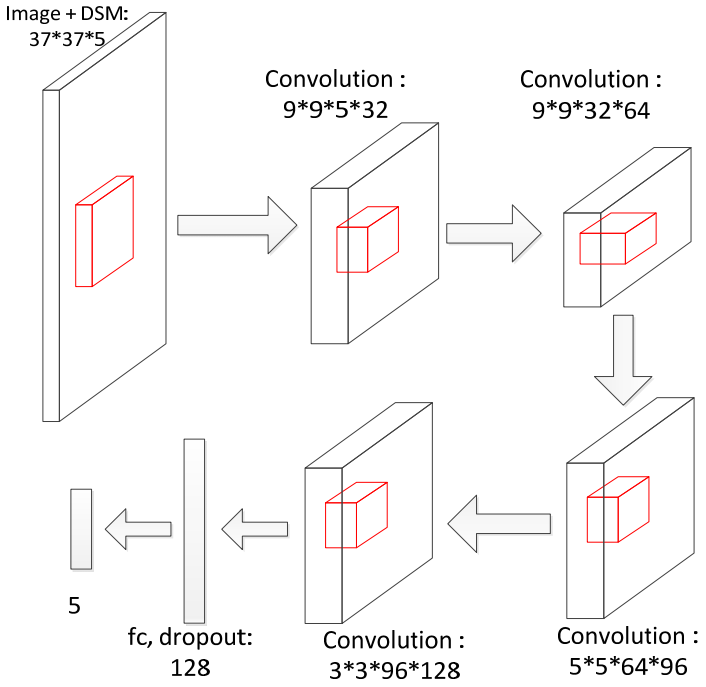

Figure 2. An illustration of the CNN architecture taking the example when the patch size is 37 pixels with all the data sources included. The network consists of five layers (four convolutional layers and one fully-connected layer) with a final 5-way soft-max loss layer.

2.2 Classification using hand-crafted features

In Yao et al. (2012), spatial contextual features around pixel together with ensemble classifier were found as effective means for discriminating various classes in the data labeling contest in urban areas. On the other hand, handcrafted spatial-spectral features were extracted in a local neighborhood to enable traditional pixel-based classification, which can be implemented here using different classifiers such as LR or random forests.

Since these features could be complementary to the patch-based texture features extracted by the CNN, a separate classifier is trained on hand-crafted features to output class probabilities which can be combined with those metrics generated by the CNN. For each pixel, a multi-dimensional feature vector is generated: NDVI, saturation, nDSM, mean of 3 imagery channels, the channel vector to indicate covariance, entropy and kurtosis of L_2 normalized histogram of normals gathered over a 21x21 neighborhood from the DSM and nDSM. Histogram is to arrange normal vectors into 2D histogram bins i.e. elevation and azimuth. Moreover, extended morphological profiles of multispectral channels were considered to enrich the feature vector as well.

2.3 Decision-level fusion

Two pixel-based classifiers provide us the evidence for semantical class labels from different sources. The key idea of the class labelling technique in this work is to apply the fusion of two independent outcomes of pixel classification schemes, which amounts to combining CNN features with hand-crafted features to further boost the classification accuracy. Since the CNN and RF are such separate approaches, we can assume they are independent, given the data and multiply their class probabilities to result in the combined probability for each class:

$$p_i^{combi} = \frac{p_i^{cnn} p_i^{hc}}{\sum_{j=1}^{C} p_j^{cnn} p_j^{hc}} \tag{3}$$

where p_{combi}, p_{cnn} and p_{rf} are the combined, CNN and hand-crafted probabilities per class. C is the set of all the labeled object classes.

An alternative to conditional probability method to combine the labelling results from separate belief sources is evidence theory for reasoning with uncertainty based on statistical inference. Dempster–Shafer (DS) evidence theory is a generalized probabilistic model that has been often used for sensor fusion. DS is defined on degrees of belief level rather than the probability to improve the accuracy and robustness of labeling. The theory is based on two ideas of obtaining degrees of belief from subjective probabilities for a related question and providing a method to combine the previous measures of evidence of different sources. It uses discrimination framework, evidence function and probability allocation (mass) function to represent and process knowledge.

Suppose that $\Theta = C_1, C_2,\ldots,C_n$ is the discrimination hypothesis space 2^{Θ} of mutually exclusive classes and n is the number of classes, therefore basic probability allocation function m is a function from 2^{Θ} to [0,1] and it satisfies the requirements of:

$$\begin{cases} m(\phi) = 0 \\ \sum_{A \subseteq \Theta} m(A) = 1 \end{cases} \tag{4}$$

If there are two or more different evidence sources (e.g. output probability of each pixels conditioned on each class from two classifiers), the orthogonal sum is used to combine those evidences. In our case, we assume that m_1, m_2 are the probability allocation functions corresponding to evidences F_1, F_2 obtained from the two classification schemes and their orthogonal sum $m = (m_1 \oplus m_2)(A)$ is:

$$m(\phi) = 0 \tag{5}$$

$$m(A) = \frac{1}{K} \sum_{\cap A_i \neq \Phi} \prod_{1 \le i \le n} m_i(A_i) \tag{6}$$

$$K = \sum_{\cap B_i \neq \Phi} \prod_{1 \le i \le n} m_i(B_i) \tag{7}$$

K is the total contradiction level of all evidences between every two mass sets $\boldsymbol{B}_i$, which do not intersect with each other, where $\boldsymbol{A}_i$, indicates the mass sets which supplement each other by sharing the common class labels from both evidence sources. This rule derives and enhances common shared belief between multiple sources and ignores or reduces the conflicting (non-shared) belief through a normalization factor K.

For semantical labeling in remote sensing images, different classifiers may generate different class labels, resulting in the generation of evidence with high contradiction, so the improved

evidence theory with ensemble approach is expected to improve the performance by reducing the limitations caused by evidence inconsistency. For combining multiple classifiers for remote sensing data, the result of each classifier can be viewed as an independent piece of evidence. Probability mass function can be represented by the classification likelihood of a specific class. For example, if a pixel is labelled with the i_{th} class in a classifier, the basic probability is $m_i(B_i) = P_i$, where P_i is the likelihood of the i_{th} class by the specific classifier. For instance, A_i = {building or car}, {building or tree}, or {building or car}, because those dual classes are always easily confused with each other and need to be particularly treated, whereas B_i = {building or car}, {building or tree}, or {building or car}, which do not share any common class labels and are used to measure the conflict between the two mass sets from two sources.

After the completion of evidence combination for each single hypothesis of the multisource model, a decision rule to decide which hypothesis is the more realistic is chosen to specify the label. Usually the class with maximum combined evidence is selected as the final result.

3. EXPERIMENTS

3.1 Dataset

The proposed method was applied to the ISPRS benchmark dataset (Gerke, 2012), which consists of multiple high resolution large image patches, each being a true orthophoto captured over Vaihingen, Germany. The dataset also contains Digital Surface Models (DSM) generated via dense image matching for each patch, which have a ground sampling distance (GSD) of 9 cm. Labelled ground truth was provided for 16 of the areas, and were made up of 6 categories: Impervious surfaces, Building, Low vegetation, Tree, Car and Clutter/background. Normalised DSM (nDSM) was provided as well, where the normalized height is computed based on the off-ground pixels. The effect of terrain is removed in nDSM compared to the regular DSM.

3.2 Experimental design and parameter setting

The experimental design of our approach is investigated. We split the labelled images into training and validation sets. The training set consists of several areas and the validation set consists of randomly selected dispersed areas, which should not intersect with training ones. The evaluation is based on the computation of pixel-based confusion matrices. For each class, we report the harmonic mean of precision and recall (F1-score). We also report the overall accuracy (Overall Acc.). As per the different experimental design compared to Paisitkriangkra et al. (2014), the training test data were equally subdivided into five subsets, each of which should represent a semantic class to be classified. We argue that in this case the classification performance can be assessed in a more objective way to avoid the biased class distribution, especially for cars whose correct classification is actually much more difficult to be achieved than other classes.

CNN:
We randomly extract 10000 patches from each class for training. The time to train the CNN model is usually under two hours using a PC with an eight core 3.4HZ CPU, which is of course still dependant on the input patch size selected, the kernel size of the convolution layers as well as the number of epochs set. The CNN model is trained with stochastic gradient decent at a learning rate of 0.001, which is reduced by a factor of ten at every 20 epochs. The momentum and weight decay parameters are set to 0.9 and 0.0005, respectively.

Hand-crafted features:
We used the same training data set consisting of 50000 examples all inclusive as used by CNN method, whose features are calculated within each image patch and used along with the corresponding ground truth pixel labels to train a LR classifier. Note that the Car class is also included within this classifier although the features are not supposed to be very discriminative for cars.

3.3 Results

The conditional probability theory and DS theory is applied to combining CNN and LR inference probabilities. The quantitative validation accuracy obtained using the base line of parameter settings is given in Table 1. Whereas the average F1-score increased by two to eight percent against the CNN and LR classification performance, the overall accuracy is improved at least by 4%, while the biggest impact on the impervious surface class. It seems that the aesthetic appeal of the labelling is not explicitly improved by the fusion, but the numeric accuracy arguably makes it worthwhile. The main improvements brought by DS fusion are to alleviate the regions labelling with ambiguous or conflicting probabilities, by removing mislabelled and fragmented regions. Figure 3 shows two labelling examples whereby the first underlines the benefit of the combination of two feature sets and classifies via DS fusion and the second one rather indicates that the fusion of two evidence sources is not always worth especially when they produced strongly inconsistent class labelling probability.

Hand-crafted features
The accuracy of the LR classifier solely using spatial-spectral features on the validation set of images is shown in Table 1. The 21x21 CNN result is also included there. The accuracy is not bad considering the relative simplicity of the features and fixed local neighbourhood size, each using only input values from single pixels. Table 1 also shows the accuracy of the combined probabilities and DS theory when used to label the pixels. The overall accuracy improves on the CNN by about 2%, indicating that the hand-crafted features do indeed contain information that is independent from the CNN features. However, CNN features show much higher representativeness and extensiveness of information content towards different object classes in urban areas.

Method	Building	Imp. surf.	Low veg.	Tree	Car	Overall Acc.
CNN	0.74	0.82	0.68	0.80	**0.82**	0.78
LR	0.61	0.89	0.65	0.84	0.59	0.72
LR+CNN[1]	0.73	0.90	0.72	0.85	0.75	0.79
LR+CNN[2]	**0.76**	**0.90**	**0.71**	**0.85**	0.78	**0.82**

Table 1. Labelling accuracy (F1-scores) of LR classifier and evidence fusion on the validation set. [1]Classfication results based on conditional probability fusion, [2]Classfication results based on DS fusion

DSM features

In this experiment, we compare the classification performance based on CNN feature with and without the ALS height model. All experimental settings are kept identical, except the number of channels of convolutional kernels in the first layer. For orthophotos, the filter size in the first layer is set to 5x 5x3x32. For orthophotos+DSM, the filter size in the first layer is

changed to 5x5x4x32. We conduct experiments with both raw DSM and normalized DSM. Table 2 shows the average F1-score and overall accuracy of the CNN approach given different input data combinations. We observe that it is beneficial to use the normalized height as it improves the overall accuracy by 3%. Opposite to a finding reported previously, we observe that the normalized height feature has the same positive impact on the detection rate of car (increase by 2%-3%) as on that of other objects. In our experiment, we achieved the best accuracy when we combine orthophoto with the raw DSM and nDSM (an improvement of 2% on the overall accuracy).

Data sources	Building	Imp. surf.	Low veg.	Tree	Car	Overall Acc.
CIR	0.74	0.81	0.67	0.79	0.81	**0.76**
CIR+DSM	0.74	0.82	0.67	0.80	0.82	**0.77**
CIR+nDSM	**0.76**	**0.83**	**0.70**	**0.81**	**0.83**	**0.79**
All	0.74	0.82	0.68	0.80	0.82	**0.78**

Table 2. Performance comparison (F1-scores for respective class) of the CNN with different input data sources

Multi-resolution CNN feature

We employed a multi-resolution deep CNN network that predicts an output based on the 21x21, 37x37 and 65x65 pixel image patches. Experimental results are reported in Table 3. For our baseline, we trained a single-resolution CNN network. Table 3 shows a significant improvement of multi-resolution CNNs over a single-resolution CNNs based on the F1-score and overall accuracy, which is supposed to mainly benefit from more complete information content in large image patches. It can be seen from the results that the larger CNN resolution sizes usually lead to significant performance improvements with respect to all the object classes. However, the training time increased rapidly as well when a bigger image patch was applied to train the CNN. All computations were performed using an eight core CPU of an Intel Xeon E3-1245 with 3.40GHz.

Method	Building	Imp. surf.	Low veg.	Tree	Car	Overall Acc.
All	**0.88**	0.92	**0.83**	0.88	0.97	**0.89**
21*21	0.74	0.82	0.68	0.80	0.82	**0.78**
37*37	0.83	0.90	0.77	0.84	0.94	**0.86**
65*65	0.87	**0.92**	0.82	**0.87**	**0.97**	**0.88**

Table 3. Performance comparison (F1-score) between single-resolution and multi-resolution

3.4 Discussion

As also highlighted in Farabet el al. (2013) for street scenes and Paisitkriangkra et al. (2014), this work demonstrates that CNNs can effectively perform dense semantic labelling of aerial imagery with the help of nDSM, especially for those with large image patch sizes, although features are learned directly from original pixels' values rather than being hand-crafted or somehow extracted based on mathematical models. Meanwhile, in contrast to the CNN approach which is usually more sophisticated and computationally time-consuming, simple pixel-level hand-crafted features achieved (even slightly for some certain object classes) worse accuracy but with significantly lower computational costs. Perhaps this is not surprising because the input hand-crafted features are explicitly designed to discriminate the target classes: the local distribution of normal vectors from DSM highlights low vegetation and trees, nDSM highlights the house and cars and infrared channel highlights vegetation. In single-channel panchromatic images these phenomenologies cannot be relied upon, and the CNN's texture-based approach would be much more accurate.

In Paisitkriangkra et al. (2014) CRF smoothing worked as a global filter to combine the two independent classifiers to generate final labels. It had a positive effect on accuracy, whereas in former work the accuracy decreased. DS provide a probabilistic framework for combining these detections from multiple belief sources for the classifier labelling as well. Further along with paper they concluded that the CRF improved the labelling visually by removing speckle from classifier output labels, which yet does not mean to necessarily lead to better results for the quantitative evaluation. Since the classification based on CNN and spatial-spectral features is applied with a sliding window, it does not have access to object-level context during the classification. Although CRFs working on nodes constructed on the different image levels could provide object-level constraints using higher-order cliques or a hierarchical approach, the DS fusion is expected to encode the essential feature at high spatial level from both classifiers via a coherent inference and judgment, if the local neighbourhood covers a reasonably large area. The performed experiments give rise to the fact that the pixel labelling based on CNNs with a single large patch can even generate perfect labelling results for all classes, without considering the time complexity incurred.

CNN and spatial-spectral features based on LR are largely complementary to each other, thus resulting in almost all categories of classification performance benefiting from considerable improvement after non-selective binding, except vehicles. Thereby the proof is obtained by experiment that DS theory can make decisions at single-point level on the two feature sets to obtain mutual enhancement, but for small targets like vehicle DS fusion leads to information loss while carrying out the judgment. The major implication is: selective decision level fusion on single-pixel tags – the relevant pixels to integrate CNN with LR outputs can balance the trade-off between classification accuracy to be achieved and computational time to be required.

4. CONCLUSION

This work demonstrated that CNNs can not only effectively perform semantic labelling of aerial imagery by learning the texture features derived directly from the data rather than being hand-crafted, but also would achieve better accuracy in multi-channel images whereby spatial-spectral features around local neighborhood cannot be always relied upon. In spite of the computational cost of the CNN approach, the pixel-level classification based on local spatial-spectral features does not necessarily mean to achieve much worse accuracy. The DS theory provides a probabilistic framework for combining the class labelling results, when both classifier outputs are consistent and even complement each other. The competitive classification accuracy is explained by the nature of input data: e.g. the DSM and near infrared channel, and also attributed to feature/decision-level fusion of CNN texture-based approach with multi-channel spatial-spectral features generating more coherent labelling probability based on evidence combination theory.

REFERENCES

Benediktsson, J., Swain, P and Ersoy. O., 1990. Neural network approaches versus statistical methods in classification of multisource remote sensing data. *IEEE Trans. on Geoscience and Remote Sensing*, 28(4), pp.540–552.

Bischof, H., Schneider, W., and Pinz., A., 1992. Multispectral classification of landsatimages using neural networks. *IEEE Trans. on Geoscience and Remote Sensing*, 30(3):482–490.

Castelluccio, M., Poggi, G., Sansone, C., and Verdoliva, L. 2015. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. arXiv preprint arXiv:1508.00092.

Farabet, C., Couprie, C., Najman, L and LeCun. Y., 2013. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Analysis. Machine. Intelligence*, 35(8), pp.1915–1929.

Friedl, M. A and Carla, E. B., 1997. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, 61(3), pp. 399-409.

Gerke, M., ISPRS 2D Semantic Labeling Contest. http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html. 1, 5 (20 August. 2015)

Gerke, M., 2015. Use of the stair vision library within the ISPRS 2D semantic labelling benchmark (Vaihingen). Technical report, University of Twente, 2015.

Girshick, R., Donahue, J., Darrell, T., and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. pp. 580-587.

Khaleghi, B., Khamis, A., Karray, F. O., Razavi, S. N., 2013. Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, 14(1), pp.28-44.

Krizhevsky, K., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in *conference on advances in neural information processing systems*, pp.1097-1105.

Liu, K., Skibbe, H., Schmidt, T., Blein, Z., Palme, K., Brox, T., Ronneberger, O., 2014. Rotation-invariant hog descriptors using fourier analysis in polar and spherical coordinates. *International Journal of Computer Vision*, 106(3), pp.342–364.

Niemeyer, J., Wegner, J., Mallet, C., Rottensteiner, F and Soergel, U., 2011. Conditional random fields for urban scene classification with full waveform lidar data. In *Photogrammetric Image Analysis*, volume, 6952, pp. 233–244.

Paisitkriangkra, S., Sherrahy, J., Janneyz, P and Van-Den Hengel, A., 2014. Report on the ISPRS 2D Semantic Labeling Contest. "Effective Semantic Pixel labelling with Convolutional Networks and Conditional Random Fields", http://www.itc.nl/external/ISPRS_WGIII4/ISPRSIII_4_Test_results/papers/DSTO__.pdf (28 May. 2015).

Saeidi, V., Pradhan, B., Idrees, M. O., and Abd Latif, Z., 2014. Fusion of airborne LiDAR with multispectral spot 5 image for enhancement of feature extraction using Dempster–Shafer theory. *IEEE Transactions on Geoscience and Remote Sensing*, 52(10), pp.6017-6025.

Razavian, A.S., Azizpour , H., Sullivan, J and Carlsson, S., 2014. CNN features off the-shelf: an astounding baseline for recognition. In Proc. IEEE Conf. Computer. Vision. Pattern Recognition, Columbus, OH, pp.512-519.

Rottensteiner, F., Trinder, J., Clode, S., Kubik, K., 2005. Using the Dempster–Shafer method for the fusion of LIDAR data and multi-spectral images for building detection. *Information fusion*, 6(4), pp.283-300.

Tabib Mahmoudi, F., Samadzadegan, F., Reinartz, P., 2015. Object recognition based on the context aware decision-level fusion in multiviews imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(1), pp.12-22.

Wei, Y., Yao, W., Wu, J., Schmitt, M., Stilla, U. 2012. Adaboost-based feature relevance assessment in fusing lidar and image data for classification of trees and vehicles in urban scenes. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1(7), pp.323-328.

Yao, W., Hinz, S., Stilla, U., 2011. Extraction and motion estimation of vehicles in single-pass airborne LiDAR data towards urban traffic analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3), pp.260-271.

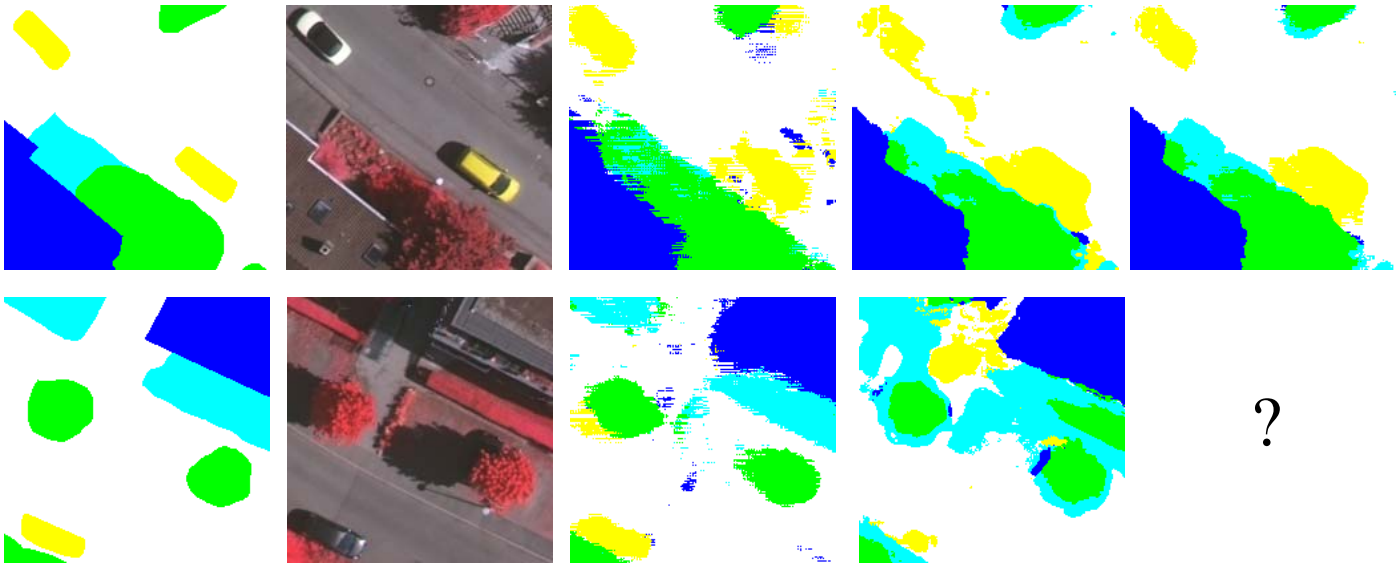

Figure 3. Two examples of classification results for complementary (top) and conflicting (bottom) scenes. From left to right: ground truth, input image CNN labelling, combined classifier labelling