AERIAL PHOTO BUILDING CLASSIFICATION BY STACKING APPEARANCE AND ELEVATION MEASUREMENTS

T.T. Nguyen, S. Kluckner*, H. Bischof, F. Leberl

Institute for Computer Graphics and Vision, Graz University of Technology, Austria – {thuy,kluckner,bischof,leberl}@icg.tugraz.at

KEY WORDS: Vision, Urban, Combination, Classification, Colour, DEM/DTM, Model, High resolution

ABSTRACT:

Remote Sensing is trending towards the use of greater detail of its source data, advancing from ever better resolving satellite imagery via decimeter-type aerial photography towards centimeter-level street-side data. It also is taking advantage of an increase in methodological sophistication, greatly supported by rapid progress of the available computing environment. The location awareness of the Internet furthermore demonstrates that large area remote sensing strives for a model of human scale detail. This paper addresses the task of mapping entire urban areas, where objects to be mapped are naturally three dimensional. Specifically we are introducing a novel approach for the segmentation and classification of buildings from aerial images at the level of pixels. Buildings are complex 3D objects which are usually represented by features of different modalities, i.e. visual information and 3D height data. The idea is to treat them in separated processes for learning and then integrate them into a unified model. This aims to exploit the discriminative power of each feature modality and to leverage the performance by fusing the classification potentials at a higher level of the trained model. First, representative features of visual information and height field data are extracted for training discriminative classifiers. We exploit powerful covariance descriptors due to the low-dimensional region representation and the capability to integrate vector-valued cues such as color or texture. Then, a stacked graphical model is constructed for each feature type based on the feature attributes and classifier's outputs. This allows to learn inter-dependencies of modalities and to integrate spatial knowledge efficiently. Finally, the classification confidences from the models are fused together to infer the object class. The proposed system provides a simple, yet efficient way to incorporate visual information and 3D data in a unified model to learn a complex object class. Learning and inference are effective and general that can be applied for many learning tasks and input sources. Experiments have been conducted extensively on real aerial images. Moreover, due to our general formulation the proposed approach also works with satellite images or aligned LIDAR data. An experimental evaluation shows an improvement of our proposed model over several traditional state-of-the-art approaches.

1. INTRODUCTION

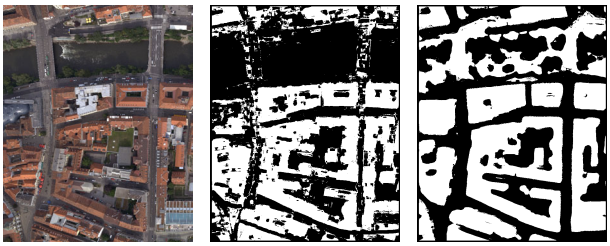
Remote sensing is rapidly moving towards half-meter satellite imagery, decimeter aerial imagery and centimeter-type street-side photography, and all of these in a multi-spectral mode. Simultaneously, large areas of the World are now being mapped at human-scale detail to support the Internet's recent appetite for location awareness. This is resulting in a new domain of large-area urban mapping, however to result not in photo-textured point clouds, but in interpreted objects from which one can build a model of the urban World. A central task is the segmentation and classification of images of buildings. This needs to be fully automated to be at sufficiently low cost so that large area mapping is feasible.

A large urban area may encompass 150 to 500 square-kilometers. Large scale aerial imagery may be at a pixel size of 10 cm. Such large urban area may be covered by 10,000 large-format aerial photographs at high overlaps. We are thus addressing a challenging task of scene interpretation and understanding. It is essential for many location-based applications, such as detailed image description (Meixner and Leberl, 2010), realistic 3D building modeling (Zebedin et al., 2008) or virtual city construction (Leberl et al., 2009). Over the years, the automated building extraction has been an active research topic. Considering a large scale processing, the problem of building classification becomes very difficult for many reasons. Buildings are complex objects with many architectural details and shape variations. Buildings are located in urban scenes that contain various objects from man-made to natural ones. Many of those are in close proximity or disturbing, such as parking lots, vehicle, street lamps, trees, etc. Some objects are covered with shadows or cluttered. These difficulties make the problem of a general building detection challenging. Figure 1 depicts typical urban scenes taken from three challenging datasets *Graz*, *Dallas* and *San Francisco* showing some of these difficulties. We therefore propose an approach which combines several feature cues such as color, texture and 3D information in order to obtain a reliable building extraction from aerial images.

Figure 1: Typical color images of complex urban scenes taken from the dataset *Graz*, *Dallas* and *San Francisco*.

With the success of the aerial imaging technology, high resolution images can be obtained cost-effectively. Multiple sources of data become available, i.e. color, infrared and panchromatic images (Zebedin et al., 2006). Furthermore, since the aerial images are taken with a high overlap from different camera viewpoints, a dense match approach (Klaus et al., 2006) can be applied to obtain range images, representing digital surface models (DSM), from neighboring images. Taking into account the DSM, 3D height information describing the real elevation of each pixel from ground can be computed. The obtained 3D information in combination with visual cues can be exploited efficiently for tasks like accurate building extraction. Figure 2 shows two classification results obtained for a scene of *Graz* by separately incorporating color and 3D information. It is obvious, that a combination of both cues will provide an improved classification results. Moreover, aerial images contain a huge amount of data, which requires efficient methods for processing. This work presents a general

* Corresponding author.



(a) (b) (c)

Figure 2: A classification result obtained by using only color (b) or 3D height information (c). In this work we apply a stacked graphical model in order to combine the strength and advantages of both pipelines.

yet efficient approach that integrates the power of discriminative feature cues.

Automatic building classification and extraction has been a very active research topic in photogrammetry and computer vision for years. The proposed approaches heavily differ in the use of data sources, extracted feature types, the applied models or the evaluation methods (Jaynes et al., 2003, Matei et al., 2008, Lafarge et al., 2008, Mueller and Zaum, 2005, Persson et al., 2005, Xie et al., 2006, Sirmacek and Unsalan, 2008).
Traditional approaches for general image classification problems are mainly based on locally extracted features and a learned classifier that discriminate the object from background. Visual information describing the appearance, such as color and texture are mixed together in a single feature vector to represent the object instance. A concatenating of multiple feature types into a single vector may cause an encountered affect, i.e. one feature type may inhibit the performance of another; besides, it may also cause the problem of over-fitting due to redundancy and correlation in the input data (Duda et al., 2001). Moreover, standard learning algorithms, such as Naive Bayes, logistic regression, support vector machines (SVM) assume that the training data is independent and identically distributed. This is inappropriate in many cases, as image pixels possess dependencies, e.g. if a pixel is labeled as building, it is likely that a neighboring pixel is also labeled as building; non-building pixels tend to be next to other non-building pixels. The spatial dependencies should be exploited properly to improve the classification performance rather than classifying each of the image sites independently.
There have been wide research interests in random field models, i.e. Markov random field (MRF), conditional random field (CRF) (Lafferty et al., 2001, Li, 2001), and their variants in the computer vision community. These models aim to incorporate contextual information to the decision of the object class for improving the performance of the classifiers. In (Korc and Foerstner, 2008), the authors employed MRFs and showed that parameter learning methods can be improved and that using the approach to interpret terrestrial images of urban scenes is feasible. In the vision community, modern approaches exploit graphical models for integrating additional information about the content of a whole scene (Shotton et al., 2006, Larlus and Jurie, 2008, Verbeek and Triggs, 2007).
Recently, Ma and Grimson (Ma and Grimson, 2008) proposed a coupled CRF model for decomposing the feature space in order to learn the object classes. Besides, there have been attempts to model contextual interactions by employing related predictions in a stacked graphical model (SGM) learning framework (Kou and Cohen, 2007). This model enables efficient learning and inference. Moreover, the concept of a relational template can be flexibly exploited to incorporate multi-modal interactions. Our work can be considered as an extension of both, the coupled CRF model (Ma and Grimson, 2008) and the SGM learning (Kou and Cohen, 2007). In this work we propose a novel approach based on an ensemble of SGM in order to integrate different data sources for building classification at the pixel level.
In contrast to the work of Matikainen et al. (Matikainen et al., 2007), where they proposed to use a DSM segmentation and a classification pipeline discriminating buildings from trees, we focus on a more direct and general approach. Our model is comprised of multiple classifiers that are learned over stages and then fused together. Each classifier is responsible for a certain feature modality and modeled as a SGM for the learning procedure. For each SGM, we propose to use a relational template which takes into account the predictions not only of related instances of a certain feature type, but also predictions from other types. This enables to learn not only spatial knowledge of object class, but also the inter-modality dependencies. The proposed system provides a simple yet efficient framework to model a complex object class such as buildings and exploit the potentials from different aspects of the object properties. Learning and inference are effective, general and straightforward, that can be easily applied for many other learning tasks.

Our paper is organized as follows: In Section 2. we introduce our novel framework. Section 3. describes the aerial imagery and the involved feature cues. Section 4. highlights the experimental evaluation. Finally, Section 5. concludes the work and discusses open issues for future work.

2. OUR FRAMEWORK

Let the observed data from an input image be $X = \{\mathbf{x}_i, 0 < i < |X|\}$, where $\mathbf{x}_i$ is the data from a site i. The problem is to find the most likely configuration of the labels $Y = \{\mathbf{y}_i\}$, where $\mathbf{y}_i \in \{c_1 \dots c_k\}$. For an image labeling, a site is a pixel location, and a class may be a car, a building, etc. For the task of the building segmentation each pixel in the aerial image, represented by a feature vector $\mathbf{x}_i$, is mapped to a bit $\mathbf{y}_i \in \{-1, +1\}$, corresponding to either building or non-building. A traditional CRF with local potentials and pairwise (spatial) dependencies can be written as

$$P(Y|X) = \frac{1}{Z(X)} \prod_{i \in S} A(\mathbf{y}_i, X) \prod_{i,j \in N_i} B(\mathbf{y}_i, \mathbf{y}_j, X), \qquad (1)$$

where $A(\mathbf{y}_i, X)$ corresponds to the local potential of $\mathbf{x}_i$ given a class label $\mathbf{y}_i$; $B(\mathbf{y}_i, \mathbf{y}_j, X)$ is the interaction potential function which encodes the dependencies between data X, labels at i and its neighbor j, based on the set of pixels in a neighbor N_i of $\mathbf{x}_i$. $Z(X)$ is a partition function and S defines a set all available image sites. Note that in the formal CRF formulations, potentials depend on the whole image X, not only on the local site $\mathbf{x}_i$. A SGM can be seen as a simplified form of the CRF, given in Equation 1, which allows a flexible structure for the interactions and provides efficient learning and inference. The general stacked model is formulated as a combination of T multiple components of conditional distribution that capture contextual information

$$P(Y|X) = \frac{1}{Z(X)} \prod_{t=1}^{T} p^t(\mathbf{y}_i|X). \qquad (2)$$

The number of components T depends on the model built for a particular application. The flexibility of a CRF formulation al-

lows to incorporate multiple aspects of data from the image, such as: local statistic of an image site, neighboring labels, or potentials from higher levels of contexts. This property will be employed in our framework, where we propose to use a two-staged approach.

2.1 The Ensemble Model - Stage 1

We decompose the input feature space into different feature cues, which represent multiple modalities of the input data. These types of features may be representative for color, texture and 3D information. Assuming that T feature types are extracted from an input image X, $X = \{X^t\}$, $t \in T$, then the CRF of the first stage can be modeled as combination of multiple sources

$$P_{(1)}(Y|X) = \frac{1}{Z(X)} \prod_{t=1}^{T} p^t_{(1)}(Y|X^t), \tag{3}$$

where each $p^t_{(1)}(\mathbf{y}|\mathbf{x})$ is a SGM. The main reasons of decomposing the input data and using an ensemble model are: First, as it has been investigated, object properties such as color, shape, texture, 3D data, etc. play different roles in distinguishing object classes (Kluckner et al., 2009, Ma and Grimson, 2008). Therefore, we treat them separately in different classification processes and combine them at later stages to infer the object classes (Nilsback and Caputo, 2004). At the first stage, we employ multiple strong classifiers learned from different feature modalities. The classifiers provide a probabilistic class assignment in terms of a likelihood. In this work, we use efficient randomized forests (RF) (Breiman, 2001) as base classifiers to generate initial yet accurate confidence maps (Kluckner et al., 2009, Shotton et al., 2008). However, any other types of classifiers, that results a class probability, e.g. boosting, SVM, etc., can be applied to our framework. In order to train the the classifiers of the ensemble, we use fast covariance matrix descriptors as feature representation as proposed by (Kluckner et al., 2009). The details for the feature representation and the base classifier are described in Section 2.3 and Section 2.4.

Since random field modeling approaches exploit contextual information to improve the detection rate of standard classifiers, it is intuitively sensible that different object's property have their own context where it is more likely to appear. This is especially true in our application, where multiple sources of aerial image data are employed, i.e. color image and height data: it may be claimed that pixels with similar color could have similar labels; however, this is not true for height data: pixels with the same height values may belong to buildings or trees. So, the ensemble model comprised of multiple SGMs, where each responses to a certain feature type, is useful to exploit potential of each feature type and its own context.

2.2 The Ensemble of a SGM - Stage 2

We are interested in a model that captures the dependencies among different kinds of feature modalities and spatial knowledge as contexts. Therefore, the second stage of the model is based on the features and the outputs of classifiers from the first stage. Again, we treat each type of features separately. At this stage, we model the dependencies between the feature types and the spatial context. This enables to handle inter-features dependencies and to learn the interactions at a higher level:

$$P_{(2)}(Y|X) = \frac{1}{Z(X)} \prod_{t=1}^{T} p^t_{(2)}(Y|X^t, p_{(1)}). \tag{4}$$

We propose a new relational template for the SGM, in which each feature vector of a certain type is expanded with predictions from its related instances. In particular, each original feature vector of a certain feature type is augmented (stacked) with the prediction confidences from its neighboring sites and confidences from other feature types, which forms a new training set. This allows to learn the spatial dependencies as well as the inter-modality relationships. We use an aggregate function to build the new training dataset: For each feature type $t \in T$, each instance $\mathbf{x}^t_i$ is expanded with the prediction confidences from its neighbor N_j and from other feature types $p^t_{N_j}$ and $p^{T\setminus t}_j$, respectively:

$$\mathbf{x}^t_{i,new} = (\mathbf{x}^t_i, p^t_{N_j}, p^{T\setminus t}_{N_j}, 0 < j < 8). \tag{5}$$

Multiple discriminative probabilistic classifiers are now learned on these new training sets. At this stage, we use a linear SVM due to efficiency and its discriminative power. Finally, the classification confidences of the classifiers are fused together for the inference of the object classes.

2.3 Feature Representation

In the vision community, covariance matrix based descriptors are widely used for detection and classification tasks (Tuzel et al., 2006) due to providing a compact and low-dimensional feature representation. A set of independent feature vectors $\mathbf{f}_i \in F$, where F is an image structure that includes the feature attributes for e.g. color, height, etc. and i defines an image site, can be represented by a sample mean μ_i and a covariance matrix Σ_i defining the first and second order statistics. Importantly, extended integral images (Tuzel et al., 2006) provide an efficient computation of covariance matrices within rectangular image regions. Since the space of covariance matrices is non-Euclidean, these descriptors can not be directly used as a feature representation for learning RFs or SVMs. Here, we exploit a derived representation based on Sigma Points (Kluckner et al., 2009) to obtain a valid feature space, that can be trained with our machine learning techniques. Please note, due to our general model, any other feature representation or classification procedure can be applied.

2.4 Randomized Forests as Base Classifier

An RF classifier (Breiman, 2001) consists of a collection of K decision trees. The nodes of each tree include fast binary decisions that give the direction of splitting left and right down the tree until a leaf node is reached. Each leaf node $l \in L$ contains a learned class distribution $P(c|L)$. By propagating single class distributions bottom-up to the root node for all K trees in a forest the resulting accumulated probabilities yields an accurate class distribution $P(c|L) = \frac{1}{K}\sum_{i=1}^{K} P(c|l_i)$. As demonstrated in (Shotton et al., 2008, Kluckner et al., 2009), RF classifiers give robust and accurate results in challenging image classification tasks. To grow each tree of the forest, node tests are learned by using only a small chosen subset of the training data X^t (Shotton et al., 2008). The learning proceeds from the root node top-down by splitting the subset at each node into tiled left and right subsets. The decisions in the nodes minimize the sample weighted information gain ratio (Shotton et al., 2008) of the class distribution in currently available subsets of the training data. Proposed decisions

in (Shotton et al., 2008) arise by reason of comparing two randomly chosen elements of a given feature vector. At evaluation time, the class distribution for each pixel is determined by parsing down the extracted feature representation in the forest. RFs provide robust probabilistic outputs and are extremely fast to train and test.

2.5 Learning and Inference

In this work, the RF classifiers provide the local class potentials for each individual feature type. After obtaining initial classification confidences at the first stage, the new training sets are constructed using Equation 5. At the second stage, linear SVMs are employed to train these new expanded datasets. We keep the same parameters for individual classifiers at each stage. Thus, there is no need for a parameter tuning in a high-dimensional space. After learning, the classifiers are applied to various test images. However, due to spectral differences in color and specified height conditions, we have to train individual models for each dataset of *Graz*, *Dallas* and *San Francisco*. We then combine the confidence maps (rather than hard output of classifiers) to infer the final object class. Our ensemble model enables to classify buildings from aerial images and to segment building's regions at the pixel level. This involves inferring a true label for each pixel, which is done by computing the most likelihood $y^* = argmax_y P(Y|X)$, given the features X and the classification function. The overall procedure for learning and inference of the model is summarized in Algorithm 1.

Algorithm 1 Learning and Inference

Learning algorithm:
Given a training set (X^t, Y), for each feature types $t \in T$.
For each feature type $t \in T$
- Stage 1: Learn the local model using an RF with (X^t, Y)
- Compute a probabilistic class assignment $p^t_{(1)}$
- Expand the dataset by stacking (Eq. 5): $x^t_{i,new} = (\mathbf{x}^t_i, p_{(1)})$
- Stage 2: Learn the SGM using SVM with (X^t_{new}, Y)

Inference:
Given a test image X, for each feature type $t \in T$
Compute $P^t_{(1)}(Y|X)$ and $P^t_{(2)}(Y|X)$
Infer final class labels: $y^* = argmax_y \prod_t P^t_{(2)}(Y|X)$

3. AERIAL IMAGERY

We perform experiments on high resolution aerial images extracted from three datasets (*Graz*, *San Francisco* and *Dallas*) showing different characteristics. The dataset *Graz* shows a colorful appearance with challenging buildings, the images of *San Francisco* have suburban occurrence in a hilly terrain and *Dallas* includes large building structures and is mainly dominated by gray valued areas. The aerial images are taken with the *Microsoft UltraCam* in highly overlapping strips, where each image has a resolution of 11500×7500 pixels with a ground sampling distance of approximately 10 cm. We use two types of image information, which are: the RGB color image and the 3D height data produced by using the DSM (Klaus et al., 2006) and a subsequently computed digital terrain model (DTM) (Champion and Boldo, 2006). By combining the DTM and DSM, we obtain an absolute elevation per pixel from ground, which is used as the 3D height information. Additionally, we exploit texture information, provided by processing the color images with first-order derivative filters. Figure 3 shows a typical scene taken from *Graz*, including the color image, the hand-labeled ground truth mask and the corresponding normalized 3D information. In our approach we exploit such ground truth map with two classes to train our classifiers in a supervised manner.

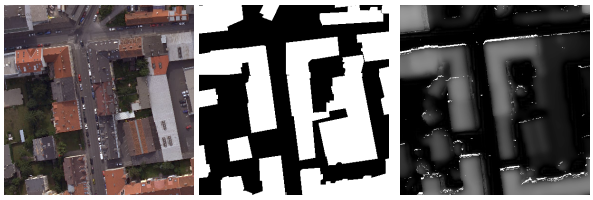


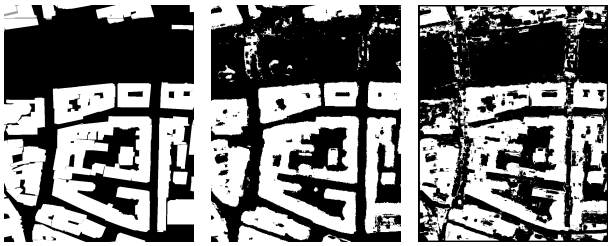
Figure 3: A scene taken from the *Graz* dataset: the color image, the hand-labeled ground truth map and the corresponding normalized 3D height information (from left to right).

Graz	Classification Accuracy (%)		
Model types	Overall	Building	Non-Build.
SVM	88.15	91.47	85.77
RF	85.42	76.95	91.46
Stacked RF model	88.39	91.45	88.39
Our SGM model	**91.65**	93.38	91.09
Dallas	Classification Accuracy (%)		
Model types	Overall	Building	Non-Build.
SVM	93.11	90.40	94.41
RF	91.76	75.86	99.39
Stacked RF model	93.31	90.94	94.63
Our SGM model	**93.76**	90.81	95.12
San Francisco	Classification Accuracy (%)		
Model types	Overall	Building	Non-Build.
SVM	87.97	81.31	96.79
RF	91.17	89.33	93.62
Stacked RF model	92.12	88.34	94.32
Our SGM model	**93.98**	94.40	93.42

Table 4: Performance evaluation of different models in terms of correctly classified pixels obtained for the datasets *Graz*, *Dallas* and *San Francisco*. We compute a global rate and the accuracy individually for each of the classes building and non-building by considering a hand-labeled ground truth map. The accuracy measurements for the building and non-building class are also referred to as completeness and correctness, respectively. The rates are given for the models integrating the visual feature cues and the 3D height information.

4. EXPERIMENTS

In this section we evaluate our proposed framework on a large amount of real world data. We compare the performance of our model to several traditional state-of-the-art approaches. The comparisons include the performances of a traditional RF and SVM classifiers, both integrating appearance and 3D height, a SGM with RFs as base classifier (in the following we call it a stacked RF), and our ensemble model also including the second stage of our approach. Each of the base RF classifiers consists of $K = 8$ trees with a maximum depth of 14. For the stacked models (including the stacked RF and our ensemble model), the cross-validation parameter is set to 4 and the relational template takes into account 8 direct neighboring pixel sites. We use a linear SVM for learning the stacked RF and our ensemble model at the second stage. The feature instances are collected on a regular image grid incorporating a small spatial neighborhood of 11 pixels in order to include important context information. The covariance feature representation based on Sigma Points comprises a compact statistical description of an image region with a dimension of $d' = d(2d + 1)$, where d denotes the number of considered feature modalities. Note, we consider each color channel of an RGB image as a single modality.



(a) (b) (c)

Figure 5: The corresponding classification results for the images presented in Figure 2 by integrating appearance and height information: (a) the ground truth image, (b) the results obtained by our SGM and (c) a traditional RF based classification. Compared to Figure 2, our SGM obtains sharp delineated building boundaries and false positive classified regions are consistently eliminated.

For training and testing the model, six representative triplets extracted from the large images are used. Each of these sub-images has a size of 4000×3200 pixels. The images cover large dense urban areas, which contain various complex objects, such as buildings of variant sizes and complex architectures, road net, parking lots, trees, shadow, water surface, etc. For the quantitative comparison we evaluate each labeled pixel with respect to the available ground truth data.

Considering Figure 2, it is obvious that a classifier, only trained on 3D features, fails in river regions, where the dense matching process provides regions with many undefined heights. In addition, tree areas are classified as buildings due to similar building height. By using only the 3D height data an RF classifier obtains a detection rate of 78.57% on this scene extracted from *Graz*. Exploiting only the visual feature modality, the raw RF classifier correctly assigns the pixels at a rate of 79.20%. However, there are significant missed detections in regions on the ground, that have similar appearance as buildings. A combination of the height and the visual features within an RF classification process significantly improves the final labeling at the pixel level to an accuracy of 85.42%. By integrating the height field data with the visual information within our proposed SGM framework, we obtain an overall pixel classification rate of more than 90% on all three datasets. The detection rates in terms of accuracy at the pixel level of different models are summarized in Table 4. The supervised segmentation of building regions obtained by a traditional RF classifier is shown in Figure 5(c), while the performance of our SGM is depicted in Figure 5(b).

The classification is given as raw outputs of each model without applying a post-processing step. However, this could be done to remove small noisy areas on the ground. Besides, our SGM obtains sharp delineated building boundaries and false positive classified regions are consistently eliminated. The improvement is obvious and results from the feature decomposition and integration at higher level with spatial context. Moreover, we obtain a very fast learning and inference thanks to the intrinsic simple model structure and the efficient relational template for the stacked graphical learning. A classification of an image with a dimension of 4000×3200 pixels can be obtained within few minutes using an unoptimized implementation. Figure 6 shows an improved performance of our approach compared to traditional state-of-the-art methods such as RF classifiers on larger scenes taken from *Dallas* and *San Francisco*, respectively.

5. CONCLUSION

We have proposed an efficient approach for learning multiple feature modalities, i.e. visual features and 3D height data. Our method decomposes an input feature space into different feature modalities in order to train individual probabilistic classifiers. In this work we used randomized forests as base classifiers, trained with various feature types, at the first stage of a stacked graphical model. Then, an ensemble of stacked models with a novel relational template has been employed for learning the dependency of different modalities. We successfully applied the proposed model to the challenging problem of the building classification task in high resolution aerial images, taken from three different datasets. Experiments have shown an improvement of our approach over several traditional state-of-the-art methods. The model is suitable for learning 3D objects like buildings from aerial imagery, but can be applied for other object classes. Due to efficiency, the proposed framework provides a promising application for large-scale computation in aerial imagery. For future work there should be more study on modeling context information for each feature type, which represent different aspects of data. Multiple kernels would be helpful in weighting the contribution of each source of information. In addition, we plan to apply our framework to various detection tasks in standard evaluation image collections.

REFERENCES

Breiman, L., 2001. Random forests. In: Machine Learning, pp. 5–32.

Champion, N. and Boldo, D., 2006. A robust algorithm for estimating digital terrain models from digital surface models in dense urban areas. In: Proceedings International Society for Photogrammetry and Remote Sensing.

Duda, R. O., Hart, P. E. and Stork, D. G., 2001. Pattern Classification. New York: Wiley.

Jaynes, C., Riseman, E. and Hanson, A., 2003. Recognition and reconstruction of buildings from multiple aerial images. Computer Vision and Image Understanding 90(1), pp. 68–98.

Klaus, A., Sormann, M. and Karner, K., 2006. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. Proceedings International Conference on Pattern Recognition.

Kluckner, S., Mauthner, T., Roth, P. M. and Bischof, H., 2009. Semantic classification in aerial imagery by integrating appearance and height information. In: Proceedings Asian Conference on Computer Vision.

Korc, F. and Foerstner, W., 2008. Interpretation terrestrial images of urban scenes using discriminative random fields. In: Proceedings International Society for Photogrammetry and Remote Sensing.

Kou, Z. and Cohen, W. W., 2007. Stacked graphical models for efficient inference in markov random fields. In: Proceedings SIAM International Conference on Data Mining.

Lafarge, F., Descombes, X., Zerubia, J. and Pierrot Deseilligny, M., 2008. Automatic building extraction from dems using an object approach and application to the 3d-city modeling. International Journal of Photogrammetry and Remote Sensing 63(3), pp. 365–381.

Lafferty, J., McCallum, A. and Pereira, F., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings International Conference on Machine Learning.

Larlus, D. and Jurie, F., 2008. Combining appearance models and markov random fields for category level object segmentation. In: Proceedings Computer Vision and Pattern Recognition.

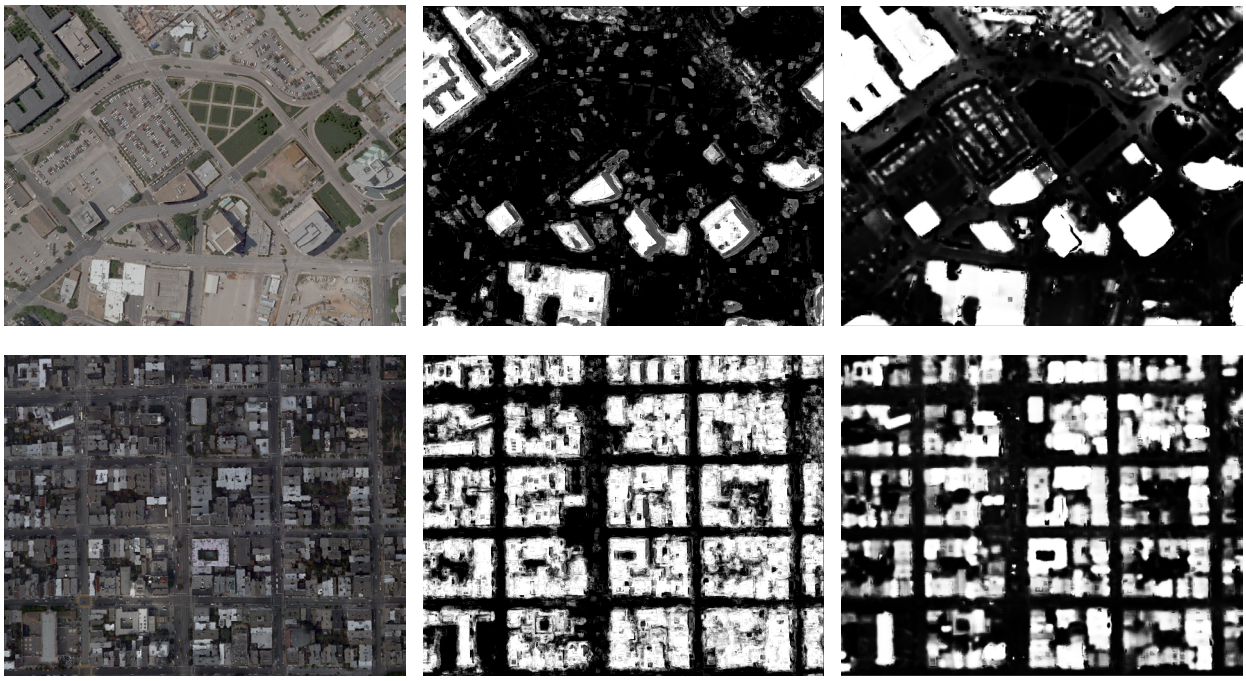


Figure 6: A visual comparison between the classification results of an RF classifier (second column) and our ensemble model (third column). The first row shows the building classification, given in terms of SVM output confidence values, for a scene of *Dallas*, while the second row depicts the results for a part of *San Francisco*.

Leberl, F., Kluckner, S. and Bischof, H., 2009. Collection, processing and augmentation of vr cities. In: Proceedings 52. Photogrammetric Week.

Li, S. Z., 2001. Markov random field modeling in image analysis. Springer-Verlag New York, Inc.

Ma, X. and Grimson, W., 2008. Learning coupled conditional random field for image decomposition with application on object categorization. In: Proceedings Computer Vision and Pattern Recognition.

Matei, B., Sawhney, H., Samarasekera, S., Kim, J. and Kumar, R., 2008. Building segmentation for densely built urban regions using aerial lidar data. In: Proceedings Computer Vision and Pattern Recognition.

Matikainen, L., Kaartinen, K. and Hyyppae, 2007. Classification tree based building detection from laser scanner and aerial image data. In: Proceedings International Society for Photogrammetry and Remote Sensing, Workshop Laser Scanning.

Meixner, P. and Leberl, F., 2010. Describing buildings by 3-dimensional details found in aerial photography. In: Symposium "100 Years ISPRS - Advancing Remote Sensing Science".

Mueller, S. and Zaum, D. W., 2005. Robust building detection in aerial images. In: Proceedings International Society for Photogrammetry and Remote Sensing, Workshop CMRT.

Nilsback, M. and Caputo, B., 2004. Cue integration through discriminative accumulation. In: Proceedings Computer Vision and Pattern Recognition.

Persson, M., Sandvall, M. and Duckett, T., 2005. Automatic building detection from aerial images for mobile robot mapping. In: Proceedings Symposium on Computational Intelligence in Robotics & Automation.

Shotton, J., Johnson, M. and Cipolla, R., 2008. Semantic texton forests for image categorization and segmentation. In: Proceedings Computer Vision and Pattern Recognition.

Shotton, J., Winn, J., Rother, C. and Crimininsi, A., 2006. TextonBoost: Joint Appearance, Shape and Conext Modeling for Muli-class object Recognition and Segmentation. In: Proceedings European Conference on Computer vision.

Sirmacek, B. and Unsalan, C., 2008. Building detection from aerial images using invariant color features and shadow information. In: Proceedings International Symposium on Computer and Information Sciences.

Tuzel, O., Porikli, F. and Meer, P., 2006. Region covariance: A fast descriptor for detection and classification. In: Proceedings European Conference on Computer Vision.

Verbeek, J. and Triggs, B., 2007. Region classification with markov field aspect models. In: Proceedings Conference on Computer Vision Pattern Recognition.

Xie, M., Fu, K. and Wu, Y., 2006. Building recognition and reconstruction from aerial imagery and lidar data. In: International Conference on Radar.

Zebedin, L., Bauer, J., Karner, K. F. and Bischof, H., 2008. Fusion of feature- and area-based information for urban buildings modeling from aerial imagery. In: Proceedings European Conference on Computer Vision.

Zebedin, L., Klaus, A., Gruber-Geymayer, B. and Karner, K., 2006. Towards 3d map generation from digital aerial images. International Journal of Photogrammetry and Remote Sensing 60(6), pp. 413–427.

ACKNOWLEDGEMENTS

This work was financed by the FFG Project APAFA (813397) and the Austrian Science Fund Project W1209 under the doctoral program Confluence of Vision and Graphics.