BUILDING DETECTION USING AERIAL IMAGES AND DIGITAL SURFACE MODELS

Jia Mu[a], Shiyong Cui[b], Peter Reinartz[b]

[a] Elektronische Fahrwerksysteme GmbH, Dr.-Ludwig-Kraus-Strae 6, 85080 Gaimersheim - (jia.mu.bewerbung)@googlemail.com
[b] Remote Sensing Technology Institute, German Aerospace Center (DLR), 82234 Wessling, Germany, (shiyong.cui, peter.reinartz)@dlr.de

Commission VI, WG VI/4

KEY WORDS: Building detection, variational inference, logistic regression, Bag-of-Words (BoW), conditional random fields, aerial images, classification

ABSTRACT:

In this paper a method for building detection in aerial images based on variational inference of logistic regression is proposed. It consists of three steps. In order to characterize the appearances of buildings in aerial images, an effective bag-of-Words (BoW) method is applied for feature extraction in the first step. In the second step, a classifier of logistic regression is learned using these local features. The logistic regression can be trained using different methods. In this paper we adopt a fully Bayesian treatment for learning the classifier, which has a number of obvious advantages over other learning methods. Due to the presence of hyper prior in the probabilistic model of logistic regression, approximate inference methods have to be applied for prediction. In order to speed up the inference, a variational inference method based on mean field instead of stochastic approximation such as Markov Chain Monte Carlo is applied. After the prediction, a probabilistic map is obtained. In the third step, a fully connected conditional random field model is formulated and the probabilistic map is used as the data term in the model. A mean field inference is utilized in order to obtain a binary building mask. A benchmark data set consisting of aerial images and digital surfaced model (DSM) released by ISPRS for 2D semantic labeling is used for performance evaluation. The results demonstrate the effectiveness of the proposed method.

1. INTRODUCTION

Building detection from aerial and satellite images has been a main research issue for decades and is of great interest since it plays a key role in building model generation, map updating, urban planning and reconstruction (Davydova et al., 2016). Various methods have been developed and difference data sources such as aerial images, digital surface/eleviation models, LIDAR data, multi-spectral images, synthetic aperture radar images, have been used for building detection. In this section we briefly review relevant methods in the literature on building detection. Decades ago the initial endeavor for building detection was relying on grouping of low level image features such as edge/line segments and/or corners to form building hypotheses (Ok, 2013). For instance, a generic model of the shapes of building was adopted in (Huertas and Nevatia, 1988) and shadows cast by buildings were used to confirm building hypotheses and to estimate their height. A computational techniques for utilizing the relationship between shadows and man-made structures to aid in the automatic extraction of man-made structures from aerial imagery is described in (Irvin and McKeown, 1989). An approach to perceptual grouping for detecting and describing 3-D objects in complex images was proposed in (Mohan and Nevatia, 1989) and was illustrated by applying it to the task of detecting and describing complex buildings in aerial images. The vertical and horizontal lines identified using image orientation information and vanishing point calculation were used in (McGlone and Shufelt, 1994) to constrain the set of possible building hypotheses, and vertical lines are extracted at corners to estimate structure height and permit the generation of three dimensional building models from monocular views. Due to the neglected performance evaluation in building detection, a comprehensive comparative analysis of four building extraction systems was presented in (Shufelt, 1999) and he concluded that none of the developed systems were capable of handling all of the challenges in building detection. Most of these initial methods rely heavily on the adopted low level features and assumption of a specific type of building hypothesis. However due to the uncertainty in low level features, their performance is not likely to be perfect.

Due to the popularity of space-borne VHR sensors in a wide variety of remote-sensing-related applications, multi-spectral information have motivated new approaches based on machine learning techniques for building detection. For instance, multi-spectral classification and texture filtering were combined in (Zhang, 1999) within a two-level framework to optimize building detection in satellite images. In the first level, a fused image was classified using ISODATA clustering. A filtering method based on a modified co-occurrence matrix was applied in the second level to improve the classification results of the first level. Later on, morphological transformations to build a differential morphological profile was proposed for building detection in (Pesaresi and Benediktsson, 2001) and (Benediktsson et al., 2003). A combined fuzzy pixel-based and object-based approach for classification of urban land cover from high-resolution multi-spectral image data was proposed in (Shackelford and Davis, 2003). This method was demonstrated using pan-sharpened multispectral IKONOS imagery from dense urban areas. A system was developed in (Ünsalan and Boye, 2005) to detect houses and residential street networks in multispectral satellite images, which produced a highly successful detection rate (94.8%) for house detection. However, because of the assumptions on buildings, it is applicable only to the buildings in North America. Without assumptions on building structure, a generic method for the detection of man-made objects in high resolution optical remote sensing images was developed in (Inglada, 2007) by SVM classification of geometric image features such as geometric invariants and FourierMellin descriptors. Nevertheless, this approach

is not designed to detecting building regions instead the patches of a particular size corresponding to building. A novel decision fusion approach to building detection in VHR optical satellite images is proposed in (Senaras et al., 2013). The method combines the detection results of multiple classifiers under a hierarchical architecture, called Fuzzy Stacked Generalization (FSG). However, this method assumes statistical stability of the training and test data. A new approach based on an adaptive fuzzy-genetic algorithm was proposed in (Sumer and Turker, 2013) for building detection using high-resolution satellite imagery. This approach combines a hybrid system of evolutionary techniques with a traditional classification method (Fishers linear discriminant) and an adaptive fuzzy logic component. Nevertheless it is by no means determinist that genetic algorithms can find a global optimum solution.

Methods based on graphical models are quite popular for building detection as well. An MRF model was used in (Krishnamachari and Chellappa, 1996) to group these lines to delineate buildings in aerial images. First straight lines are extracted from images by using an edge detector followed by a line extractor. Then an MRF model is formulated on these extracted lines with a suitable neighborhood. The probabilistic model is chosen to support the properties of the shapes of buildings. In the end, the energy function associated with the MRF is minimized, resulting in the grouping of lines. However, no quantitative results were provided for evaluation. Similarly, a stochastic image interpretation model, which combines both 2-D and 3-D contextual information of the imaged scene, was proposed in (Katartzis and Sahli, 2008) for the identification of building rooftops. However, the approach is only applicable to building rooftops with low inclination. A graph-based approach was developed in (Kim and Muller, 1999) for building detection. The whole process of building detection is divided into four small stages of line extraction, line-relation-graph generation, building hypothesis generation, and building hypothesis verification. This method is yet limited to certain building shapes. Another method for building detection based a graphical method was proposed by (Sirmacek and Unsalan, 2009), where the vertices in the graph are SIFT Keypoints. A multiple subgraph matching method was applied to detection individual building by matching graphs corresponding to a template and a test image. Nevertheless this method is applicable only to urban areas with well-separated buildings. A different method based on graphical modeling of buildings was proposed by in (Cui et al., 2012), where the vertices in the graph are the intersections of line segments. The graph was then adapted to the buildings by filtering the edges by considering the region properties. Then all cycles are search by an algorithm and the most probable cycle were considered as the best building candidate. A novel system was developed by (Izadi and Saeedi, 2012) for automatic detection and height estimation of buildings with polygonal shape roofs in singular satellite images and it employs image primitives such as lines, and line intersections, and examines their relationships with each other using a graph-based search to establish a set of rooftop hypotheses. The height of each rooftop hypothesis is estimated using shadows and acquisition geometry.

Another category of methods for building detection is based on active contour model. For example, a method based snake that combines the regional features of an image with context was proposed in (Peng and Liu, 2005) using the direction of the cast shadows. However, it is not applicable to complex buildings in urban areas. Similarly a modified ChanVese model based level set method was proposed in (Cao and Yang, 2007) for detecting man-made objects in aerial images. A three-stage level set evolution strategy was used to minimize the proposed model energy with a fractal error and texture edge descriptor. Unfortunately the method extracts only the boundaries of the man-made regions instead of the building outlines. A variational framework for building detection was proposed in (Karantzalos and Paragios, 2009) by an integration of multiple competing shape priors that is pose/affine invariant through an explicit estimation of the transformation. However, this method is limited to prior building shape models. A new model, based on level set formulation, is introduced in (Ahmadi et al., 2010) to detect buildings in aerial images using active contour models. All building boundaries are detected by introducing certain points in the buildings vicinity. However, the number of building and background classes must be precisely known a priori to achieve the best results.

There are still a large number of relevant works to this paper. The above introduction is by no means comprehensive. In this paper, we introduce a method for building detection by classifying the pixels comprising building through a logistic regression classifier that is learned by a Bayesian method. In the following sections, the steps comprising the entire method are presented in detail and the method is evaluated using a benchmark data set consisting of aerial images and digital surfaced model (DSM) released by ISPRS for 2D semantic labeling.

2. METHOD

2.1 Overview of the method

The overview of the proposed method is shown in Figure 1. It consists of three step: Bag-of-Words (BoW) feature extraction, Bayesian learning via variational inference, and pixel labeling via a fully connected CRF model. For the purpose of building characterization, an effective bag-of-Words (BoW) method is applied for feature extraction in the first step. After that, a classifier of logistic regression is learned in this feature space via a Bayesian method. Due to the presence of hyper prior in the probabilistic model of logistic regression, approximate inference methods have to be applied for prediction. In order to be fast, a variational inference method based the mean field theory is applied for inference. In the end, a fully connected CRF model is applied for labeling the building pixels, where the inference is performed by the mean field method.

2.2 BoW feature extraction

The Bag-of-Words (BoW) method has a large number of successful applications in various fields. In this paper we follow our previous work (Cui et al., 2015) on BoW method for image classification. The framework of BoW feature extraction is composed of four steps as shown in the second column in Figure 1, namely local feature extraction, dictionary learning, feature coding, and feature pooling. Assume we have a data set of N images $I_i, i = 1, ..., N$, the first step is to sample a collection of patches from the images in the database. This can be done by dense sampling or sparse detection. The second step is to extract local descriptor vectors $\mathbf{x}_i^j \in \mathbb{R}^D, j = 1, ..., M$ from all patches. The third one is learning a dictionary $\mathbf{D} = (\mathbf{d}_1, ..., \mathbf{d}_K) \in \mathbb{R}^{D \times K}$ with K words using all local features. Normally, this is done by a time consuming unsupervised learning method, such as k-means clustering or a Gaussian mixture model. The elements $\mathbf{d}_i$ in a dictionary are the centers of the clusters. The next step is to find a dictionary-based representation $\mathbf{v} = [v_1, ..., v_K]$ for each previously extracted local descriptor $\mathbf{x}$. This can be done using

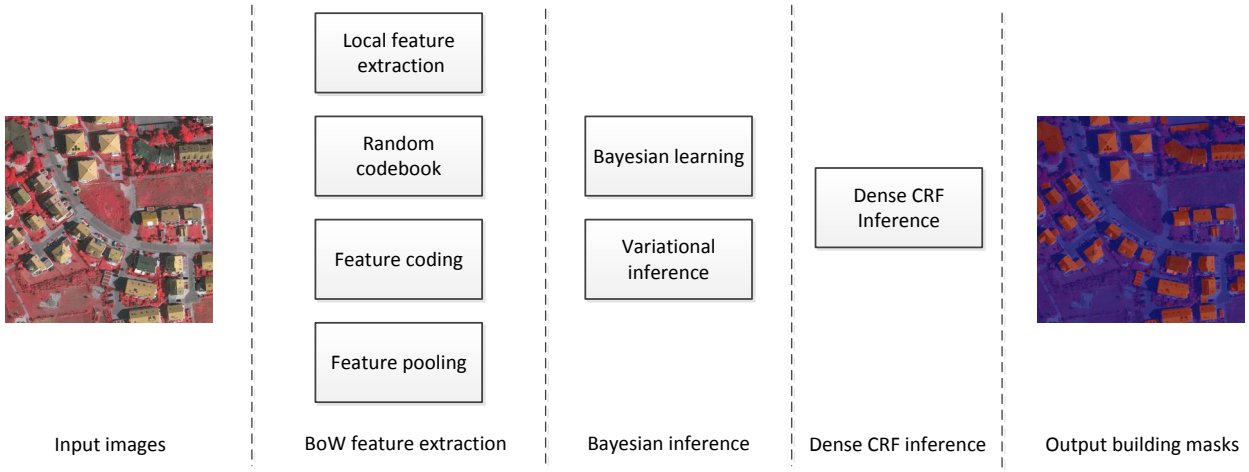


Figure 1. The workflow of the proposed method for building detection. It consists of three steps: BoW feature extraction, Bayesian learning, and Dense CRF for building pixel labeling.

hard feature assignment or soft assignment. Hard assignment assigns a single label, i.e., the index of the nearest neighbor in the dictionary, to each local descriptor $\mathbf{x}$. Formally, it is defined as:

$$v_k(\mathbf{x}) = \begin{cases} 1 & \text{if } k = \text{argmin}_i \|\mathbf{x} - \mathbf{d}_i\|^2 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Thus, the final descriptor representation $\mathbf{v} = [v_1, ..., v_K]$ has only one non-zero element. The last step is to do the sum-pooling [1] of all local descriptors extracted from one image $\mathbf{v}_i = \text{sum}(\mathbf{v}_i^j, ..., \mathbf{v}_i^j)$.

This method for image classification can be easily extended and applied to pixel classification that is the goal of this paper. To this purpose, we first extract local features as described in our previous work (Cui et al., 2015). Then we calculate a BoW feature representation of a local neighborhood surrounding each pixel. These BoW feature vectors are considered as the a characterization of building pixels and used in the next steps for learning a logistic regression.

2.3 Bayesian logistic regression

Logistic regression is a widely used statistical model for the approximation to the underlying functional relation between a feature vector $\boldsymbol{x}$ and a binary response variable $y \in \{-1, 1\}$. Formally the probabilistic model is defined as

$$p(y = 1|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1 + \exp(-\boldsymbol{w}^T\boldsymbol{x})} = \sigma(\boldsymbol{w}^T\boldsymbol{x}) \quad (2)$$

where $\sigma(\cdot)$ is the logistic sigmoid function. $\boldsymbol{w}$ is the model parameters that is going to be inferred by the Bayesian method. Consequently, the probability that $y = -1$ is $p(y = -1|\boldsymbol{x}, \boldsymbol{w}) = 1 - p(y = 1|\boldsymbol{x}, \boldsymbol{w})$. Thus, the model can be universally written as $p(y|\boldsymbol{x}, \boldsymbol{w}) = \sigma(y\boldsymbol{w}^T\boldsymbol{x})$.

Given a training data $\mathcal{D}$ consisting of inputs $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ and their corresponding target values $\boldsymbol{y} = \{y_1, ..., y_N\}$, the data likelihood is

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = \prod_{i=1}^{N} \sigma(y_i\boldsymbol{w}^T\boldsymbol{x}_i). \quad (3)$$

[1] Sum-pooling is equivalent to computing the histogram in the case of hard feature assignment.

The model is explicitly conditioned on the input data $\boldsymbol{X}$ although they are not random variables. The prior on the model parameters $\boldsymbol{w}$ is assumed to be an isotropic gaussian

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) = (\frac{\alpha}{2\pi})^{D/2} \exp\Big(-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\Big) \quad (4)$$

α is a hyper-parameter in the prior. To be a fully Bayesian treatment, all parameters including unknown quantities of interest and nuisance parameters are considered as random variables and are assumed to be following certain distributions. Therefore we assume a hyper-prior Gamma distribution $p(\alpha) = \Gamma(\alpha|c, d)$ on α. Thus, the joint distribution of all the parameters $\{\boldsymbol{w}, \alpha\}$ and the data $\boldsymbol{X} = \{\boldsymbol{x}_i, y_i\}$ is

$$p(\boldsymbol{w}, \alpha, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) \times p(\boldsymbol{w}|\alpha) \times p(\alpha) \quad (5)$$

which can be verified through the graphical model shown in Figure2. Bayesian inference in this case revolves around two

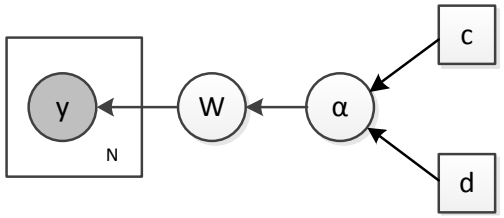


Figure 2. Graphical model for Bayesian logistic regression. Circles denote random variables and shaded circles represent observations while squares denote deterministic variables.

steps: the computation of the marginal posterior distribution $p(\boldsymbol{w}|\boldsymbol{y})$ and the computation of predicative distribution $p(y_*|\boldsymbol{y})$ for a new test data $\boldsymbol{x}_*$ based on the posterior distribution. In principle, the marginal posterior distribution $p(\boldsymbol{w}|\boldsymbol{y})$ can be computed by integrating out α in the full posterior distribution $p(\boldsymbol{w}, \alpha|\boldsymbol{y})$ that is yet hard to compute because of the model evidence $p(\boldsymbol{y})$. Accordingly the predicative distribution can be computed by integrating over the model parameters $\boldsymbol{w}$, as given in

$$p(y|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) = \int p(y|\boldsymbol{x}_*, \boldsymbol{w}) \times p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) d\boldsymbol{w} \quad (6)$$

Unfortunately, both integral involved in computing the marginal

posterior and the predicative distribution are not trackable. Therefore, one has to resort to approximation to the posterior, which can be solved mainly in two ways: deterministic and stochastic approximation. The stochastic methods, mainly referring to various Markov Chanin Monte Carlo (MCMC) method, rely on a large number of samples drawn indecently from the posterior distribution $p(\boldsymbol{w}, \alpha|\boldsymbol{y})$. In most cases, this is computationally very intensive and it is hard to ensure the independence between samples. Thus, in this paper, we concentrate on deterministic variational approximation based on the mean field theory.

2.4 Variational inference

The goal of variational inference (Blei et al., 2016) is to approximate a (posterior) distribution. The key idea is to solve this problem with optimization. A family of distributions over the latent variables, parameterized by free variational parameters, is selected. The optimization finds the member of this family, i.e., the setting of the parameters, that is closest in the Kullback-Leibler divergence to the conditional distribution of interest. The fitted variational distribution then serves as a proxy for the exact conditional distribution. All inferences involved in prediction are computed using the variational distribution instead of the posterior distribution. One widely used family of distributions is the factorized distributions, which leads to the well-known mean field inference. Thus we assume that the posterior distribution can be approximated as $q(\boldsymbol{w}, \alpha) = q(\boldsymbol{w}) \times q(\alpha)$.

The variational distribution are found by maximizing the variational lower bound

$$\mathcal{L}(q) = \iint q(\boldsymbol{w}, \alpha) \ln \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) \times p(\boldsymbol{w}|\alpha) \times p(\alpha)}{q(\boldsymbol{w}, \alpha)} d\boldsymbol{w} d\alpha \tag{7}$$

However, the data log likelihood does not have a conjugate prior in the exponential family and will be approximated by the use of a lower bound of the logistic sigmoid function (Jaakkola and Jordan, 2000), which is

$$\begin{aligned} &\sigma(x) \geq \sigma(\xi) \exp[(x-\xi)/2 - \lambda(\xi)(x^2 - \xi^2)], \\ &\lambda(\xi) = -\frac{1}{2\xi}[\sigma(\xi) - \frac{1}{2}] \end{aligned} \tag{8}$$

with one additional parameter ξ for each observation. Thus, by replacing the sigmoid function in the likelihood in (3) with this lower bound, we can obtain a lower bound in (9) of the data log-likelihood.

$$\begin{aligned} &\ln p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) \geq \ln h(\boldsymbol{w}, \boldsymbol{\xi}) \\ &= \sum_{i=1}^{N} \ln \sigma(\xi_i) + y_i \boldsymbol{w}^T \boldsymbol{x}_i / 2 - \xi_i / 2 - \lambda(\xi_i)([\boldsymbol{w}^T \boldsymbol{x}_i]^2 - \xi_i^2) \end{aligned} \tag{9}$$

Substituting the new lower bound as the data log-likelihood in (9), one can obtain a new lower bound of the variational lower bound in (7).

$$\mathcal{L}(q, \boldsymbol{\xi}) = \iint q(\boldsymbol{w}, \alpha) \ln \frac{h(\boldsymbol{w}, \boldsymbol{\xi}) \times p(\boldsymbol{w}|\alpha) \times p(\alpha)}{q(\boldsymbol{w}, \alpha)} d\boldsymbol{w} d\alpha \tag{10}$$

This lower bound is going to be maximized to seek for the variational distributions $q(\boldsymbol{w}, \alpha)$. Nevertheless it contains the parameters $\boldsymbol{\xi}$. Here we adopt an alternating optimization by maximizing w.r.t either $\boldsymbol{\xi}$ or $q(\boldsymbol{w}, \alpha)$ while fixing the other. While fixing $\boldsymbol{\xi}$, the variational distributions can be computed by standard variational methods for factorized distributions (Bishop, 2006). The variational distribution for $\boldsymbol{w}$ is given by

$$\begin{aligned} \ln q(\boldsymbol{w}) &= \ln h(\boldsymbol{w}, \boldsymbol{\xi}) + \mathbb{E}_\alpha[\ln p(\boldsymbol{w}|\alpha)] + \text{const} \\ &= \ln \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w) \end{aligned} \tag{11}$$

where

$$\boldsymbol{\Sigma}_w^{-1} = \mathbb{E}_\alpha(\alpha)\boldsymbol{I} + 2\sum_{i=1}^{N} \lambda(\xi_i)\boldsymbol{x}_i \boldsymbol{x}_i^T, \quad \boldsymbol{\mu}_w = \boldsymbol{\Sigma}_w \sum_{i=1}^{N} \frac{y_i}{2} \boldsymbol{x}_i \tag{12}$$

Similarly the variational distribution for α is

$$\begin{aligned} \ln q(\alpha) &= \ln p(\alpha) + \mathbb{E}_{\boldsymbol{w}}[\ln p(\boldsymbol{w}|\alpha)] + \text{const} \\ &= \ln \Gamma(\alpha|a, b) \end{aligned} \tag{13}$$

with

$$a = c + \frac{D}{2} \quad b = d + \frac{1}{2}\mathbb{E}_{\boldsymbol{w}}(\boldsymbol{w}^T \boldsymbol{w}) \tag{14}$$

Thus, the involved expectations are

$$\mathbb{E}_\alpha(\alpha) = \frac{a}{b}, \quad \mathbb{E}_{\boldsymbol{w}}(\boldsymbol{w}^T \boldsymbol{w}) = \boldsymbol{\mu}_w^T \boldsymbol{\mu}_w + Tr(\boldsymbol{\Sigma}_w). \tag{15}$$

While fixing the variational distributions $q(\alpha)$ and $q(\boldsymbol{w})$, the parameters $\boldsymbol{\xi}$ can be obtained by setting the derivative of (10) w.r.t $\boldsymbol{\xi}$ to zero (Bishop, 2006), giving

$$(\xi_i)^2 = \boldsymbol{x}_i^T (\boldsymbol{\Sigma}_w + \boldsymbol{\mu}_w \boldsymbol{\mu}_w^T) \boldsymbol{x}_i. \tag{16}$$

After obtaining the variational distribution $q(\boldsymbol{w})$, it can be used as a proxy for computing the predictive distribution. Then, the predictive distribution in (6) can be approximated as

$$\begin{aligned} &p(y = 1|\boldsymbol{x}_*, \boldsymbol{y}) = \int p(y = 1|\boldsymbol{x}_*, \boldsymbol{w}) \times p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) d\boldsymbol{w} \\ &\approx \int p(y = 1|\boldsymbol{x}_*, \boldsymbol{w}) \times q(\boldsymbol{w}) d\boldsymbol{w} \\ &\geq \int \sigma(\xi) \exp\left(\frac{\boldsymbol{w}^T \boldsymbol{x} - \xi}{2} + \lambda(\xi)\xi^2 - \lambda(\xi)[\boldsymbol{w}^T \boldsymbol{x}]^2\right) \times q(\boldsymbol{w}) d\boldsymbol{w} \end{aligned} \tag{17}$$

It is worth noting that the integrand is quadratic in $\boldsymbol{w}$. Thus by complete the square in exponential function, the integral can be written as

$$\begin{aligned} p(y = 1|\boldsymbol{x}_*, \boldsymbol{y}) = &\frac{1}{2} \ln \frac{|\hat{\boldsymbol{\Sigma}}|}{|\boldsymbol{\Sigma}_w|} - \frac{1}{2}\boldsymbol{\mu}_w^T \boldsymbol{\Sigma}_w^{-1} \boldsymbol{\mu}_w + \frac{1}{2}\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \\ &+ \ln \sigma(\xi) - \frac{\xi}{2} + \lambda(\xi)\xi^2 \end{aligned} \tag{18}$$

with

$$\hat{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{\Sigma}_w^{-1} + 2\lambda(\xi)\boldsymbol{x}_* \boldsymbol{x}_*^T, \quad \hat{\mu} = \hat{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}_w^{-1} \boldsymbol{\mu}_w + \frac{\boldsymbol{x}_*}{2}) \tag{19}$$

The bound parameter ξ that maximizes $\ln p(y = 1|\boldsymbol{x}_*, \boldsymbol{y})$ is given by $\xi^2 = \boldsymbol{x}_*^T(\hat{\boldsymbol{\Sigma}} + \hat{\mu}\hat{\mu}^T)\boldsymbol{x}_*^T$. Therefore, the predictive density can be computed by iterating over the updates for $\hat{\boldsymbol{\Sigma}}$, $\hat{\mu}$, ξ until $p(y = 1|\boldsymbol{x}_*, \boldsymbol{y})$ converges (Drugowitsch, 2014).

2.5 Dense CRF inference

After completing the variational inference, one obtains a probability map with each pixel having two probabilities $p(y = 1|\boldsymbol{x}_*, \boldsymbol{y})$ and $p(y = -1|\boldsymbol{x}_*, \boldsymbol{y})$. The label of each pixel can be determined by assigning the one with maximum probability. However, there will be a lot of isolated small groups of pixels

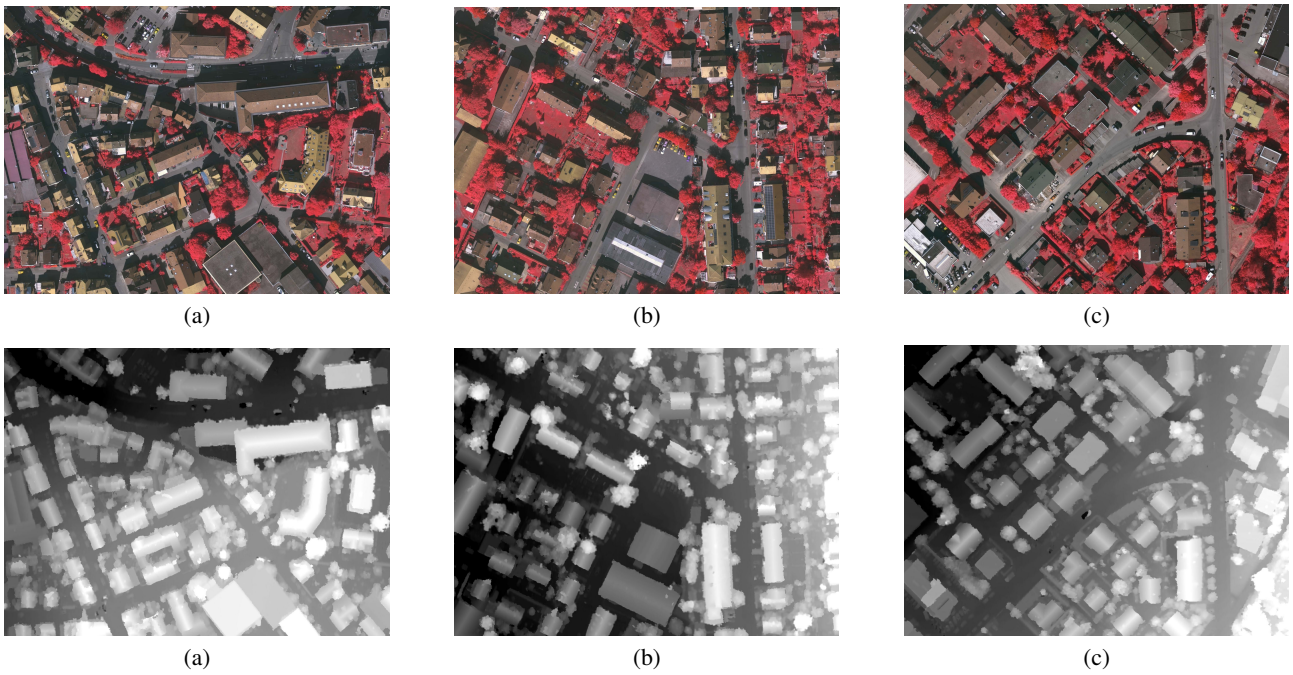


Figure 3. Example images and the corresponding DSMs used for performance evaluation.

that are wrongly labeled. Thus, a smoothness constraint should be considered when labeling the pixels based on the probability map. Usually a random filed models is applied. In this paper a fully connected CRF model (Krähenbühl and Koltun, 2011) is employed. Basic CRF models are composed of unary potentials on individual pixels or image patches and pairwise potentials on neighboring pixels or patches. The fully connected CRF uses a different model structure that establishes pairwise potentials on all pairs of pixels in the image, enabling greatly refined segmentation and labeling.

Formally, we consider a random field $\boldsymbol{X}$ defined over a set of variables $\{X_1, X_2, ... ,X_N\}$, where each variable takes value from {-1, 1}. In the fully connected pairwise CRF model, the energy function is defined as

$$E(\boldsymbol{x}) = \sum_i \phi_\mu(x_i) + \sum_{i<j} \phi_p(x_i, x_j). \tag{20}$$

where $\boldsymbol{x}$ is a particular labeling of $\boldsymbol{X}$ and i, j range from 1 to N. The unary potential $\phi_\mu(x_i)$ is computed independently for each pixel by the previous variational inference of logistic regression, which is actually a probabilistic distribution over the label assignment. The pairwise potentials in the model is defined as the following line combination of a set of Gaussian similarity functions

$$\phi_p(x_i, x_j) = \mu(x_i, x_j)\Big[w_1 \exp\Big(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\Big)] + w_2 \exp\Big(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\Big)\Big] \tag{21}$$

where p_i, p_j are the color vectors and I_i, I_j are the position vectors for pixel i and j and w_1, w_2 are the weights. There are five parameters $w_1, w_2, \theta_\alpha, \theta_\beta, \theta_\gamma$ that should be set or learned. The maximum a posteriori (MAP) labeling of the random field $\boldsymbol{X}$ is given by $\boldsymbol{x}^* = \operatorname{argmin} E(\boldsymbol{x})$. An effective inference method based on the mean field theory using high-dimensional filtering was proposed in (Krähenbühl and Koltun, 2011), which is used in this paper for labelling the building pixels.

3. EXPERIMENTS AND EVALUATION

3.1 Data set and setup

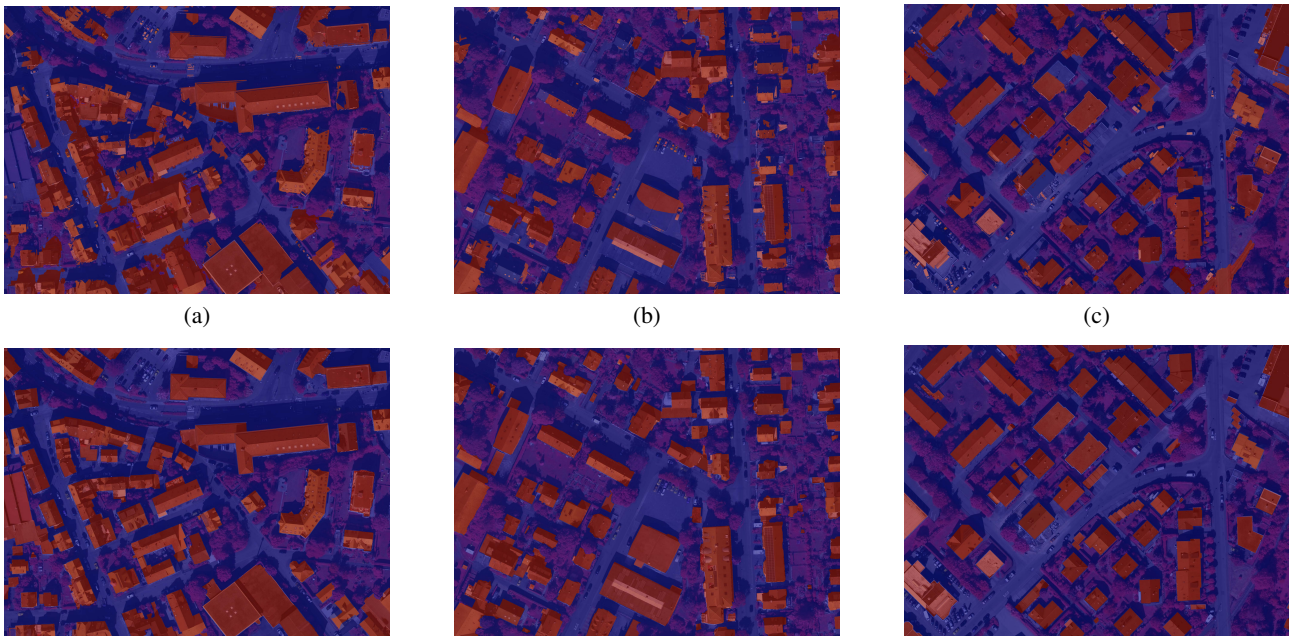
The data that we used for performance evaluation is a benchmark data set consisting of aerial images and digital surfaced model (DSM) released by ISPRS for 2D semantic labeling [2]. It has 16 arial images and corresponding DSM in addition to the ground truth of buldings. Example images and DSM in this data are shown in Figure 3.

The BoW features that we used are computed as described in section 2.2. The local features are the vectorized pixel values in a 3×3 neighborhood (Cui et al., 2015). The codebook used for feature encoding is a random dictionary, in which the entries are uniform randomly selected from the entire local feature space. The size of the codebook is empirically set to 500. 50,000 BoW features are randomly selected from each image for learning the logistic regression model by variational inference. The parameters in hyperprior are set to small values, i.e., 1e-4. The weights in (21) are set to 1.0 and the standard deviations are set to 30. All these parameters are set by try-and-error. The accuracy measures used for performance comparison are precision, recall and F1 score.

3.2 Results and discussion

The results for the 16 images are shown in table 1. The detection results using the images shown in Figure 3 are shown in the first row in Figure 4. Compared with the ground truth shown in the second row in Figure 4, the detection results are quite good. Visually there is no isolated noisy building pixels, which is the

[2] http://www2.isprs.org/commissions/comm3/wg4/tests.html



(a) (b) (c)

(d) (e) (f)

Figure 4. Example results of building detection: (a)(b)(c) detection results on the images shown in Figure 3; (d)(e)(f) ground truths.

Table 1. Performance evaluation in terms of precision and recall.

Image#	logistic regression + dense CRF		
	Precision	Recall	F1 score
1	0.74	0.86	0.80
2	0.73	0.86	0.79
3	0.91	0.79	0.84
4	0.75	0.86	0.80
5	0.59	0.88	0.71
6	0.71	0.91	0.80
7	0.76	0.82	0.79
8	0.72	0.95	0.82
9	0.58	0.84	0.69
10	0.69	0.84	0.76
11	0.92	0.63	0.75
12	0.80	0.81	0.81
13	0.76	0.88	0.82
14	0.81	0.86	0.83
15	0.85	0.85	0.85
16	0.67	0.94	0.78
aver.	0.75	0.85	0.79

desired effect by applying the fully connected CRF model. From the quantitative results shown in table 1, the average precision, recall and F1 score are respectively 75%, 85%,and 79%.

4. CONCLUSION

In this paper a method for building detection based Bayesian logistic regress and a fully connected CRF model is proposed and demonstrated. It consists of three step: Bag-of-Words (BoW) feature extraction, Bayesian learning via variational inference, and pixel labeling via a fully connected CRF model. Due to the presence of hyper prior in the probabilistic model of logistic regression, a variational inference method based on the mean field theory is applied for learning the model. A benchmark data set released by ISPRS for 2D semantic labeling is used for performance evaluation. The results demonstrate the effectiveness of the proposed method.

ACKNOWLEDGEMENTS

Acknowledgements of the data provision by the ISPRS WG II/4.

REFERENCES

Ahmadi, S., Zoej, M. V., Ebadi, H., Moghaddam, H. A. and Mohammadzadeh, A., 2010. Automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours. *International Journal of Applied Earth Observation and Geoinformation* 12(3), pp. 150 – 157.

Benediktsson, J. A., Pesaresi, M. and Amason, K., 2003. Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. *IEEE Transactions on Geoscience and Remote Sensing* 41(9), pp. 1940–1949.

Bishop, C. M., 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D., 2016. Variational Inference: A Review for Statisticians. *arXiv preprint arXiv:1601.00670*.

Cao, G. and Yang, X., 2007. Man-made object detection in aerial images using multi-stage level set evolution. *International Journal of Remote Sensing* 28(8), pp. 17471757.

Cui, S., Schwarz, G. and Datcu, M., 2015. Remote sensing image classification: No features, no clustering. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8(11), pp. 5158–5170.

Cui, S., Yan, Q. and Reinartz, P., 2012. Complex building description and extraction based on hough transformation and cycle detection. *Remote Sensing Letters* 3(2), pp. 151–159.

Davydova, K., Cui, S. and Reinartz, P., 2016. Building footprint extraction from digital surface models using neural networks. Vol. 10004, pp. 100040J–100040J–10.

Drugowitsch, J., 2014. Variational Bayesian inference for linear and logistic regression. *arXiv preprint arXiv:1310.5438*.

Huertas, A. and Nevatia, R., 1988. Detecting buildings in aerial images. *Computer Vision, Graphics, and Image Processing* 41(2), pp. 131 – 152.

Inglada, J., 2007. Automatic recognition of man-made objects in high resolution optical remote sensing images by svm classification of geometric image features. *ISPRS Journal of Photogrammetry and Remote Sensing* 62(3), pp. 236 – 248.

Irvin, R. B. and McKeown, D. M., 1989. Methods for exploiting the relationship between buildings and their shadows in aerial imagery. *IEEE Transactions on Systems, Man, and Cybernetics* 19(6), pp. 1564–1575.

Izadi, M. and Saeedi, P., 2012. Three-dimensional polygonal building model estimation from single satellite images. *IEEE Transactions on Geoscience and Remote Sensing* 50(6), pp. 2254–2272.

Jaakkola, T. S. and Jordan, M. I., 2000. Bayesian parameter estimation via variational methods. *Statistics and Computing* 10(1), pp. 25–37.

Karantzalos, K. and Paragios, N., 2009. Recognition-driven two-dimensional competing priors toward automatic and accurate building detection. *IEEE Transactions on Geoscience and Remote Sensing* 47(1), pp. 133–144.

Katartzis, A. and Sahli, H., 2008. A stochastic framework for the identification of building rooftops using a single remote sensing image. *IEEE Transactions on Geoscience and Remote Sensing* 46(1), pp. 259–271.

Kim, T. and Muller, J.-P., 1999. Development of a graph-based approach for building detection. *Image and Vision Computing* 17(1), pp. 3 – 14.

Krähenbühl, P. and Koltun, V., 2011. Efficient inference in fully connected crfs with gaussian edge potentials. In: J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger (eds), *Advances in Neural Information Processing Systems 24*, Curran Associates, Inc., pp. 109–117.

Krishnamachari, S. and Chellappa, R., 1996. Delineating buildings by grouping lines with mrfs. *IEEE Transactions on Image Processing* 5(1), pp. 164–168.

McGlone, J. C. and Shufelt, J. A., 1994. Projective and object space geometry for monocular building extraction. In: *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Mohan, R. and Nevatia, R., 1989. Using perceptual organization to extract 3d structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11(11), pp. 1121–1139.

Ok, A. O., 2013. Automated detection of buildings from single vhr multispectral images using shadow information and graph cuts. *ISPRS Journal of Photogrammetry and Remote Sensing* 86, pp. 21 – 40.

Peng, J. and Liu, Y. C., 2005. Model and context-driven building extraction in dense urban aerial images. *International Journal of Remote Sensing* 26(7), pp. 1289–1307.

Pesaresi, M. and Benediktsson, J. A., 2001. A new approach for the morphological segmentation of high-resolution satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing* 39(2), pp. 309–320.

Senaras, C., Ozay, M. and Vural, F. T. Y., 2013. Building detection with decision fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6(3), pp. 1295–1304.

Shackelford, A. K. and Davis, C. H., 2003. A combined fuzzy pixel-based and object-based approach for classification of high-resolution multispectral data over urban areas. *IEEE Transactions on Geoscience and Remote Sensing* 41(10), pp. 2354–2363.

Shufelt, J. A., 1999. Performance evaluation and analysis of monocular building extraction from aerial imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(4), pp. 311–326.

Sirmacek, B. and Unsalan, C., 2009. Urban-area and building detection using sift keypoints and graph theory. *IEEE Transactions on Geoscience and Remote Sensing* 47(4), pp. 1156–1167.

Sumer, E. and Turker, M., 2013. An adaptive fuzzy-genetic algorithm approach for building detection using high-resolution satellite images. *Computers, Environment and Urban Systems* 39, pp. 48 – 62.

Ünsalan, C. and Boye, K. L., 2005. A system to detect houses and residential street networks in multispectral satellite images. *Computer Vision and Image Understanding* 98(3), pp. 423 – 461.

Zhang, Y., 1999. Optimisation of building detection in satellite images by combining multispectral classification and texture filtering. *ISPRS Journal of Photogrammetry and Remote Sensing* 54(1), pp. 50 – 60.