SPMK AND GRABCUT BASED TARGET EXTRACTION FROM HIGH RESOLUTION REMOTE SENSING IMAGES

Weihong Cui[a, b, c, *], Guofeng Wang[d], Chenyi Feng[e], Yiwei Zheng[e], Jonathan Li[c], Yi Zhang[a]

[a] School of Remote Sensing and Information Engineering, Wuhan University, 129 LuoYu Road, Wuhan, China(whcui, ivory2008)@whu.edu.cn
[b] Collaborative Innovation Center for Geospatial Technology, 129 LuoYu Road, Wuhan,China. whcui@whu.edu.cn
[c] Mobile Mapping Lab, University of Waterloo, Canada. (w28cui, junli)@uwaterloo.ca
[d] China Highway Engineering Consulting Corporation, China wgf@checsc.com.cn
[e] Xi'an University of Science and Technology, China (491016316, 276014260)@qq.com

Commission VII WG VII/4

KEY WORDS: Spatial Pyramid Matching, Bag of Visual Words, GrabCut, Segmentation, Target Extraction

ABSTRACT:

Target detection and extraction from high resolution remote sensing images is a basic and wide needed application. In this paper, to improve the efficiency of image interpretation, we propose a detection and segmentation combined method to realize semi-automatic target extraction. We introduce the dense transform color scale invariant feature transform (TC-SIFT) descriptor and the histogram of oriented gradients (HOG) & HSV descriptor to characterize the spatial structure and color information of the targets. With the k-means cluster method, we get the bag of visual words, and then, we adopt three levels' spatial pyramid (SP) to represent the target patch. After gathering lots of different kinds of target image patches from many high resolution UAV images, and using the TC-SIFT-SP and the multi-scale HOG & HSV feature, we constructed the SVM classifier to detect the target. In this paper, we take buildings as the targets. Experiment results show that the target detection accuracy of buildings can reach to above 90%. Based on the detection results which are a series of rectangle regions of the targets. We select the rectangle regions as candidates for foreground and adopt the GrabCut based and boundary regularized semi-auto interactive segmentation algorithm to get the accurate boundary of the target. Experiment results show its accuracy and efficiency. It can be an effective way for some special targets extraction.

1. INTRODUCTION

1. 1 Motivation

Land cover classification of very high resolution (VHR) imagery over urban areas is an extremely challenging task. Impervious land covers such as buildings, roads, and parking lots are spectrally too similar to be separated using only the spectral information of VHR imagery. High resolution remote images supply us with more detail information of different kinds of targets, such as texture, shape and spatial structure, etc. Image classification, detecting and extracting targets from high resolution remote sensing images are required by many practical application. Object-orient image classification is a major method which can make use of the spectral, spatial, texture and context information (Blaschke, T., 2010). In this type of method, image segmentation is the first and a critical step. Many image segmentation methods have been proposed, such as, watershed (Beucher and Meyer, 1993), graph cut (Boykov and Jolly, 2001), mean shift (Comaniciu and Meer, 2002), MST (Felzenszwalb and Huttenlocher, 2004), etc., and many method developed on them for example Hu et al. (2005), Cui and Zhang (2011), Montoya-Zegarra et al. (2015) etc., but the uncertainty of segmentation and especially the optimal segmentation scale is the common problem which is still difficult to resolve just because of the diversity of targets in large area images. Ming et al. (2008) proposed the segmentation scale selection method, but the segmentation results usually can't fit with all the targets.

* Corresponding author

The famous software eCognition has the ability of realizing multiscale segmentation, thematic map based segmentation, multi-level and semantic context based classification and target recognition. The popular professional software ENVI, ERDAS and PCI all add the object-oriented image classification model, but they all face the same problem. The operator should take much more manual work to correct the classification results which include the category and boundary. So, manual interpretation is still a practical way while it is a huge time and manual labour consuming work. With the development of machine learning, computer vision and computer technique, the increase of image data, the accuracy of target recognition is highly improved, especially the deep learning method (Krizhevsky et al., 2012). But the automatic and accurate extracting different targets both in category and boundary is still in research. (Girshick et al., 2014; Zheng et al., 2015).

Under these conditions, using these new techniques and providing an effective semi-automatic image interpretation way will be very useful in practical work. This paper was motivated by this. We combined the automatic target detection with the interactive image segmentation to improve the manual interpretation efficiency.

In this paper, we use two local shape represented features, the dense transform color SIFT spatial pyramid (TC-SIFT-SP) descriptor and the multiscale HOG (MS-HOG) features, and an HSV colour model to express the targets. Based on these features, we construct two classifiers which are trained through the SVM classification method to detect the targets respectively. We take the intersection of the two detected results as the final

detected result which is a series of rectangle region that represents the positions and blocks of the targets. Then, we adopt the GrabCut ((Rother et al., 2004)) segmentation method to get the accurate boundary of the targets.

This paper is structured as follows. Section 1.2 focuses on related work and section 1.3 highlights the contributions of our approach. Section 2 presents the methodology, whereas section 3 describes the experimental evaluation of our approach. Finally, conclusions and an outlook are given in section 4.

1.2 Related Work

The first and critical step of our target extraction work is target detection which depends on computer vision and machine learning algorithms. The performance of machine learning algorithms are heavily dependent on the choice of data representation on which they are applied. For this reason, much of work is focused on designing the features that can support effective machine learning (Bengio, et al. 2013). Colour, shape, texture, context and multiscale characters, the main clues for objects recognizing, should be described in the data representation. Extracting local patch-level descriptors from an image, such as SIFT (Lowe, 2004), HOG (Dalal et al., 2005) and color invariant descriptors (Burghouts et al., 2009, Abdel-Hakim 2006) have been very successful representation paradigm. Bag of visual words (BOVW) is one of the popular data representation method and is widely used in target recognition and detection. It has achieved the state-of-the-art performance in several databases and competitions. These approaches quantize local region descriptors using a visual dictionary usually constructed through k-means clustering. The BOVW representation usually starts from well-designed local features, such as SIFT (Csurka et al., 2004, gradient location and orientation histogram (GLOH) (Mikolajczyk and Schmid, 2005), HOG (Dalal et al., 2005) and color invariant descriptors (Burghouts et al., 2009). Spatial pyramid representation (Lazebnik et al. 2006) is one of the first works of adding the spatial information into the BOVW representation. Yi Yang and Shawn Newsam (2011) use spatial pyramid co-occurrence representation to characterize both the photometric and geometric aspects of an image, which has been used on high-resolution aerial imagery classification. To reduce the training complexity and improve image classification performance, Yang et al. (2009) proposed sparse coding based linear spatial pyramid matching methods. Zhou et al. (2013) introduce a multi-resolution bag-of-features representation which was constructed through multiple resolution images and extract local features from all the resolution images with dense regions to scene classification. Zhang et al. (2013) use spatial pyramid robust sparse coding to image classification. Generally, the BOVW framework used for image classification has five basic steps, which are extracting patches, representing patches, generating words, encoding features, pooling features, and in which, feature coding is the core component (Huang et al., 2014). In some ways, the dense SIFT descriptor is the same as the HOG descriptor, the SPM adds the spatial construction information through the combination of multi-level grid cell visual words which may lead some kind of information loss because of the generation of words. While the HOG feature vector implies the original spatial structure of the target, and at the same time, HOG feature can be extracted in any rectangles which usually are fit with the target bounding box. These two features have its own specialty, so, to improve the exactness of target, we use them to detect targets respectively and take their intersection as the final results.

The second step is the target segmentation, our main objective is to get the boundary accurately through interactive action so as to reduce the manual work of drawing the outline of the targets. Usually, the interactive segmentation problem is taken as an optimal problem. Snakes (Kass et al., 1988), active contour models, based on energy minimization. Intelligent Scissors (Mortensen et al., 1995) take the boundary detect as graph searching problem which is to find the minimum cumulative cost path between a start pixel and a goal pixel. Geometric active contours (Caselles et al., 1997) based on active contours evolving. Graph Cut (Boykov and Jolly, 2001) combines the hard constraints and the soft constraints which incorporate both boundary and region information to find the globally optimal segmentation of the image through the max-flow algorithm. Blake et al. (2004) proposed an adaptive probabilistic model, the contrast-sensitive GMMRF, for interactive segmentation based on graph cut. Bayes matting (Chuang et al., 2001) models colour distributions probabilistically to achieve full alpha mattes. GrabCut (Rother et al., 2004), an extended version of graph-cut approach, developed an iterative algorithm of the optimisation and a robust border matting algorithm. Veksler (2008) and Gulshan et al. (2010) incorporated the object shape (i.e. a shape prior) into graph cut segmentation. Ning et al. (2010) introduce a maximal similarity region merging based interactive image segmentation. Price et al. (2010) combined geodesic distance information with edge information in a graph cut optimization framework. In these interactive segmentation methods, GrabCut can segment image using a bounding box prior robustly, it's easy to apply it to our target detection results. So, we select it as our interactive segmentation method.

1.3 Contribution

The main contribution of our work is combing target detection and image segmentation to practical image interpretation work. We give a workflow of semi-auto target extraction to relieve the manual labour. In the target detection step, to improve the precision we take the intersection of the detect result from dense TC-SIFT SPMK method and multi-scale HOG method as the final result with high degree of confidence.

2. METHODOLOGY

2.1 Feature Extraction

Colour, shape, texture, context and multiscale characters are very important features for image interpretation. High resolution remote sensing images provided more detailed textures. To utilize different kinds of and different scale of features, we select the following four kinds of popular features, the SIFT descriptor, the HOG descriptor, the transformed color and HSV color model.

The SIFT descriptor has the capability of describing the spatial distribution of a window, it has been used in many target detection researches. This feature is derived from a 4x4 gradient window by using a histogram of 4x4 samples per window in 8 direction. The gradients are then Gaussian weighted around the center. This leads to a 128 dimensional feature vector. It reflects the distribution of gradients' direction. For more detail please reference to Lowe (2004).

The HOG feature descriptor (Dalal et al., 2005) is a histogram of oriented gradients, is the local object appearance and shape within an image, it has been used to detect objects in computer vision and image processing too. To calculate this feature, the image is divided into small connected regions called cells, and for the pixels within each cell, a histogram of gradient directions is compiled. The descriptor is then the concatenation

of these histograms. To improve accuracy, the local histograms can be contrast-normalized by calculating a measure of the intensity across a larger region of the image, called a block, and then using this value to normalize all cells within the block. This normalization results in better invariance to changes in illumination and shadowing.

Color is an important component for districting objects. Color invariant descriptors are proposed to increase illumination invariance and discriminative power. There are many different methods to obtain color descriptors. Van et al. (2008) compared the invariance properties and the distinctiveness of color descriptors. The transform formula is shown in expression (1).

$$\begin{pmatrix} R' \\ G' \\ B' \end{pmatrix} = \begin{pmatrix} \frac{R-\mu_R}{\sigma_R} \\ \frac{G-\mu_G}{\sigma_G} \\ \frac{B-\mu_B}{\sigma_B} \end{pmatrix} \quad (1)$$

Where, μ_C and σ_C represents the mean and the standard deviation of the distribution respectively in channel C. This yields a distribution with $\mu_C = 0$ and $\sigma_C = 1$ for every channel. By normalizing the pixel value distributions of each chanel, the transformed color distribution is scale-invariance and shift-invariance is achieved with respect to light intensity. In this paper, we choose transformed color-SIFT descriptor to describe the color and spatial structure.

HSV is a color model, which includes three components, hue (H), saturation (S) and value (V). The hue of a color refers to which pure color it resembles. The saturation of a color describes how white the color is. The value of a color, also called its lightness, describes how dark the color is. The hue is the most distinct and important color information for people to perceive, which means it's very useful to distinct the different colors.

2. 2 Feature Representation

To represent the color and spatial feature of the targets deeply and in detail, we select two types of feature representation. One is the TC-SIFT-SP, the other one is MS-HOG &HSV feature vector composition. The prerequisite of this work is that we have enough representative positive and negative samples, which are all the same size images. With the accumulation of the thematic maps, it's easy to get the samples.

2.2.1 TC-SIFT-SP Representation: The BOVW and its spatial pyramid representation (Lazebnik et al. 2006) makes it can describe spatial distribution. To make use of color information, we select the TC-SIFT-SP descriptor to represent the feature of targets. This work need four steps to complete. First, use the formula (1) to change the RGB image to its R'G'B' image. Second, to get detailed structure information, we adopt the dense grid points to calculate the SIFT descriptor in each channel of the R'G'B' image and combine the three channel's SIFT descriptor to a vector of 384 (128*3) dimensions. We select 8 pixels as the grid space and 16 pixels as the patch size which is used to calculate the SIFT descriptor. Thirdly, use k-means cluster to get the BOVW, which is also called dictionary. Lastly, build histogram of each sample and compile its spatial pyramid histogram. Here we select the three level pyramid like Lazebnik et al. (2006), more detail please reference to it. Take a building sample (128*128 pixels) as example, Figure 1 shows one of the words which is a feature vector of 384 dimensions. Through experiments, we select 500 as the number of the words. In Figure 2, the left one shows the form of spatial pyramid, in which the blue box indicates the first (original) level, the red grids indicate the second level, and the yellows indicate the third level. The middle one shows the first level's histogram which has 500 bins and the right one shows the three level spatial pyramid histogram which have 10500 (500+4*500+16*500) bins, that is the TC-SIFT-SP representation of an image. From the spatial pyramid histogram, we could find obvious difference between the two types of targets, which means it is prominent to distinguish objects.

2.2.2 MS HOG & HSV Representation: The HOG of an image is a global structure representation. The number of its dimension is determined by the cell size, block size, overlap size, orientation bins and image size, more details please reference to Dalal et al. (2005). In experience, 9 is the most optimal number of orientation bins, cells size is 4*4 pixels, block size is 16*16 pixels, and overlap between the blocks is half of the block size. . Also take the building and grassland sample (128*128 pixels) as example, Figure 3 gives the visualization of the HOG feature in three scale images. In it, we could see the main structure of the image and the difference among the scales. The three scale's HOG feature has different dimensions, 8100 dimensions in scale 1, 1764 dimensions in scale 2, 324 dimensions in scale 3, which are the sequential arrangement of the histogram of each block in the image. To characterize the color feature of the targets, we use HSV histogram (100 bins in each channel) of the image to represent its color distribution, which should be normalized. At the end, we combined these features to a vector which contains the global spatial and color attributes, its total dimension is 10488. Figure 4 shows the feature vector value of each dimension, that is the multi-scale HOG and HSV distribution representation of the image. From it, we could also find the difference between the two types of targets.

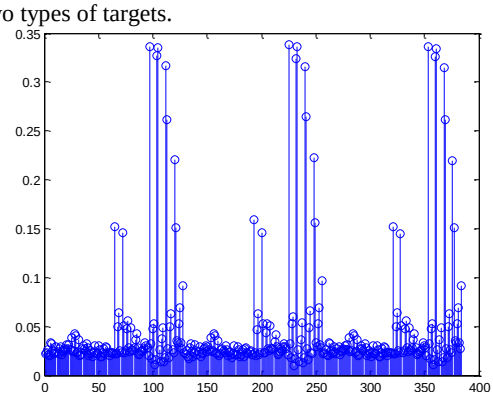
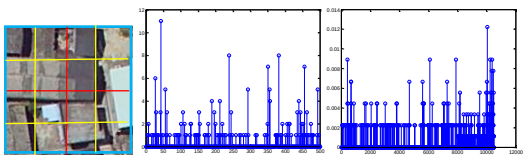
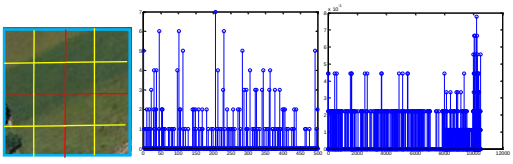


Figure 1. One of the transform color SIFT descriptor word

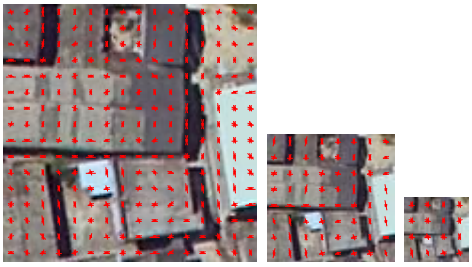


(a) TC-SIFT-SP representation of a building sample

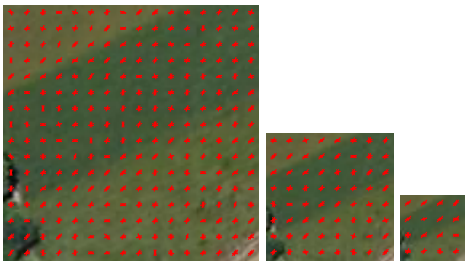


(b) TC-SIFT-SP representation of a grassland sample

Figure 2. The TC-SIFT-SP representation of two objects, (a) is of a building, (b) is of a grassland.



(a) Multi-scale HOG feature of a building sample



(b) Multi-scale HOG feature of a grassland sample

Figure 3. Visualiation of HOG features in three scale images

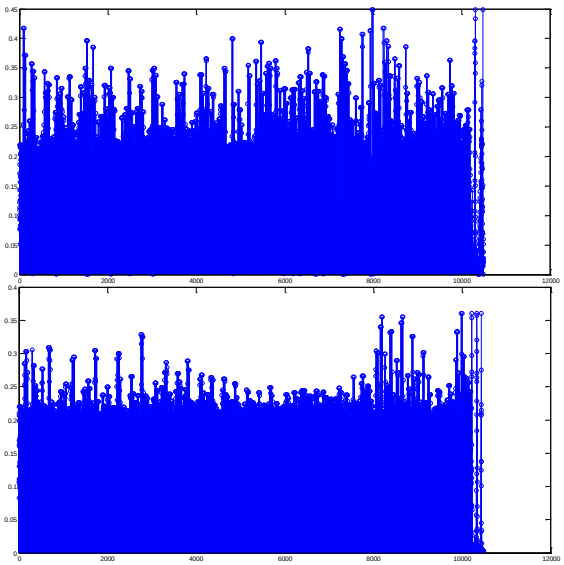


Figure 4. Value of the MS-HOG & HSV representation

2. 3 SVM Target Detection

SVM is a very popular object recognition or classification method. It has high ability in processing high dimension features. It finds the optimal separating hyperplane between two classes in an embedded space or the feature space. In this paper, the features we used are all with high dimensions, so, we select SVM to realize the target detection work. Generally, feature pre-processing and kernel selection is the first step of SVM. The kernels (such as the Gaussian RBF or polynomial) are designed to operate on a $\Re^n$ vector inputs, where each vector entry corresponds to a particular global attribute for that instance. For example, in image processing, an ordered features such as color, texture, etc. of equal length measured from the image as a whole. The MS-HOG & HSV is this kind of feature representation, which means it can be taken as an input directly. The TS-SIFT-SP is a histogram sets of different level, a disordered image representation. It has been proved that the pyramid kernel is Mercer kernel, which can be used to process unordered sets of varying sizes and works well than RBF kernel. (Grauman and Darrell, 2005; Bo et. al, 2010, 2011). Lazebnik et al. (2006) use the spatial pyramid match kernel and SVM to realize recognizing natural scene categories. So we adopt the spatial pyramid match kernel to construct the feature space too.

2. 3. 1 Spatial Pyramid Match Kernel: Feature similarity or dissimilarity is a value which can be used to distinguish the target, that is the feature matching. There are many method to measure the difference between two sets. Grauman and Darrell (2005) proposed pyramid matching to find an approximate correspondence between two sets. They gives the definition of the pyramid match kernel. Let W be the set of visual words, M is the number of visual words, that is, $\mathrm{W}=\{\mathrm{w}_1,\mathrm{w}_2,...,\mathrm{w}_M\}$. Let N represents the number of the dense points in the image where the transform color SIFT descriptor is calculated, that is, we express the image I to a set $\mathrm{I}=\{\mathrm{p}_1,\mathrm{p}_2,...,\mathrm{p}_N\}$, each p_i is a vector of 384 dimensions in this paper, assign each p_i to the most similar visual word, and then calculate the histogram of visual in I. The histogram of the visual words is denoted by $BOVW_I=\{f_1,f_2,...,f_M\}$, where f_i is the frequency of word w_i in the image. Let L represents the number of total levels of the pyramid. Each level l has $D=4^l$ cells for an image patch. The histogram of each level is denoted by $BOVW_{Ll}$ and the histogram in each cell of level l is $BOVW_{ILl}^k, k\in[1,D]$, where k represents the kth cell in level l of the image patch. Put the BOVW in each cell of each level in sequence, we'll get the spatial pyramid expression of the image, that is,

$$BOVW_I^{Ln}=[BOVW_{IL0},BOVW_{IL1},\cdots,BOVW_{ILn}]$$
$$=[BOVW_{IL0}^1,BOVW_{IL1}^1,BOVW_{IL1}^2,\cdots,BOVW_{ILn}^{4^{Ln}}]$$

The pyramid match makes use of a histogram intersection function C, which measures the "overlap" between two histograms' bin count. Let $BOVW_{I1}^l$ and $BOVW_{I2}^l$ represents the spatial pyramid representation of image I1 and I2 in level l, the pyramid match kernel of level l is:

$$C(BOVW_{I1}^l,BOVW_{I2}^l)=\sum_{k=1}^{4^l}\sum_{m=1}^{M}\min(BOVW_{I1l}^k(m),BOVW_{I2l}^k(m))$$

Where $BOVW_{I1l}^k(m)$ denotes the value of the mth bin of image I1 in cell k at level l.

Let C_i denotes the match value in level i, note that the number of matches found at level l also includes all the matches found at the finer level $l+1$. Therefore, the number of new matches found at level l is given by C_l-C_{l+1} for $l=0,1,\cdots L-1$. The weight associated with level l is set to $\frac{1}{2^{L-l}}$, which is inversely proportional to cell width at that level. The final kernel is then the sum of the separate kernels:

$$K_{SPMK}=\frac{1}{2^L}C_0+\sum_{l=1}^{L}\frac{1}{2^{L-l+1}}C_l \qquad (2)$$

This kernel reflects the similarity of two images represented by the BOVW of spatial pyramid.

2.3.2 SVM Classifier Training: We construct two types of feature representation of the samples. The difference between the representations makes us to train the SVM classifier with different inputs. For the TC-SIFT-SP representation, we get the matrix of kernel value between all pairs of training samples. The kernel's similarity values determine the samples' relative positions in the feature space. Based on this, we take the similarity matrix as the input of the SVM trainer, each row corresponds to a sample' label. For the MS-HOG & HSV representation, take the feature vector of each sample as the input. To get the optimal classification parameter, we use cross-validation to get the final classifier of each feature representation. We use the svmlib (http://www.csie.ntu.edu.tw/~cjlin/libsvm) and libsvm-faruto (http://blog.sina.com.cn/faruto) to train, predict and find the best parameters for SVM classification.

2.4 Target Segmentation

After finding out the targets in the image, we should cut the whole target out with its accurate boundary. According to this need and the characters of different segmentation algorithms, we select GrabCut (Rother et al., 2004) as the basic segmentation method to do this work. GrabCut is an interactive image segmentation method, and an iterative graph cut algorithm (Boykov and Jolly 2001), which put a light load on the user, whose interaction consists simply of dragging a rectangle around the desired object. In the process, the user indicates a region of background, and is free of any need to mark a foreground region. And it has a border matting mechanism which can reduce visible artefacts. This just meets the need of our aim of semi-auto target extract after we get the rectangle box of in the previous target detection step. That is why we select the GrabCut method to realize the target extraction and lighten the user's load. GrabCut replaced the monochrome image model for colour by a Gaussian Mixture Model (GMM) in place of histograms. It replaced the one-shot minimum cut estimation algorithm by a more powerful, iterative procedure that alternates between estimation and parameter learning. The user can interactive edit the foreground and background again, then continue perform entire iterative minimisation algorithm until convergence and the user is satisfied with the segmentation. In this paper, we take building as the target to be extracted.

After GrabCut, the contours and the main direction of each segmentation can be obtained. The direction of the eigenvector belonging to the larger of the two eigenvalues, derived from the covariance matrix of the spatial distribution of the segmentation is taken as its main direction. Use these information, we can generate a fitting and regular polygon of each building target. Firstly, we use Douglas–Peucker algorithm to simplify the contours of the target's boundary. In experience, we select 3 pixels as the distance threshold to find the simplified curves. The bigger values may lead to larger gap between the original line and the simplified one. Secondly, adjust the orientation and the position of the lines. Based on the general shape character of buildings, we need to regularize the lines according to their length, orientation and position. Find the longest line which is nearly parallel to the main direction and adjust its direction to the main direction, and then analysis the angle and distance between each line and its neighbours in sequence of the length, adjust the position and orientation of the neighbours follow the following rules. Considering some buildings are not rectangles, we divide their angle differentials into three ranges, which are 0-30, 30-60 and 60-90 degree. When the angle differentials between two lines falls in this interval, adjust the shorter line's direction to the longer one and move it near to the boundary. Finally, get the polygon of the building. Through getting the points of intersection between each two neighbour lines, and take them as the corners of the building and link them to a polygon, which is the final regularized border of the building.

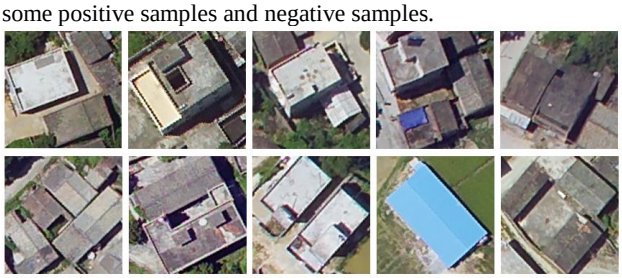
3. EXPERIMENTS

Building, roads, etc. are typical targets which are usually paid more attention to monitor their changes. Extracting them from high resolution remote sensing images is a possible and effective way. But it is still a challenging work. In this paper we select buildings as the targets and do the experiment on UAV images.

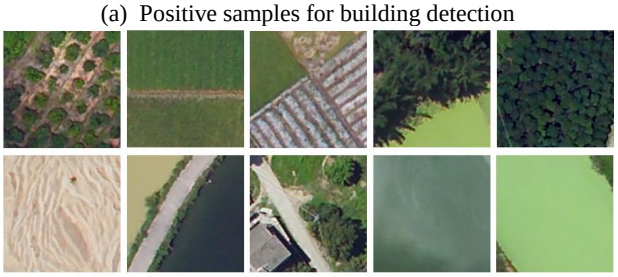
3.1 Data Prepare

The data used in the experiments are digital aerial images containing three visible bands (red, green and blue) at a spatial resolution of 0.2m. We select the aerial image of Huizhou district which is located in the middle of Guangdong province in China to extract buildings. This district is of mild climate and abundant rainfall, the major land covers of this area are forest, farms, water, roads, buildings (residential areas)..

To full reflect the character of the targets, according to the target size and image spatial resolution, we select 128*128 as the size of the positive and negative sample image patches for building detection. In our target detection experiment, we collect the positive samples and the negative samples from different original images, the negative samples include all other category objects. There are 300 positive samples and 500 negative samples for building detection training. Figure 5 gives some positive samples and negative samples.



(a) Positive samples for building detection



(b) Negative samples for building detection

Figure 5. Some examples of samples for targets detection. (a) Positive samples. (b) Negative samples

3.2 Target Detect

Using the trained SVM classifier, we do the target detection. When doing target detection process, we scan on the whole test image by the size of sample image and with the overlap of half width and half height of the sample. In detecting building experiments, we calculate its dense TC-SIFT-SP descriptor and MS-HOG & HSV feature vector of each image patch like the training periods. For TC-SIFT-SP descriptor, using SPMK to

get the similarity vector and take it as the input to predict its corresponding label. For MS-HOG & HSV feature, take the feature vector as the input directly. Through experiments, we select 500 words for building. In building detection, to improve the precision, we select the intersection of the two detection results as the final results. To improve the recall, we could choose the union set as the final results which may be a robust way in practice. Figure 6 gives the building detection results in three test images, in which, the red '*' are the predicted buildings, the yellow 'o' are the ground truth labels which are put on the centre pixel of the image patch. From the view of vision, the predicted labels are consistent with the ground truth mostly.

Figure 6. The intersection output of the building detection

We use precision and recall to evaluate the accuracy of the detection. Here we list the experiment results of different methods. Table 1 and Table 2 shows the precision and recall of different number of words in TC-SIFT-SPMK method and the SIFT-SPMK method. From them, we can find that the TC-SIFT-SPMK method has better performance than the SIFT-SPMK and 500 is the proper number of visual words. Table 3 show the results of HOG and MS-HOG&HSV method. We also can find that the MS-HOG&HSV is a bit better than HOG. Table 4 gives the intersection of TC-SIFT-SPMK and MS-HOG&HSV. From the detection results, we could find that the TC-SIFT-SPMK method and MS-HOG & HSV method can find out most of the buildings, but each of them still has its specificity. For example, the TC-SIFT-SPMK can avoid the confusion of structure cropland while the MS-HOG & HSV method can separate the road from building more robustly. Through the intersection process, the precision is improved.

Num. of Words	Test1		Test2		Test3	
	P	R	P	R	P	R
100	0.80	0.91	0.82	0.93	0.79	0.75
150	0.82	0.88	**0.83**	**0.92**	0.77	0.74
200	0.80	0.89	0.80	0.93	0.76	0.73
250	0.81	0.90	0.80	0.91	0.77	0.76
300	0.78	0.88	0.81	0.92	0.80	0.73
350	0.80	0.89	0.81	0.92	0.80	0.73
400	0.80	0.88	0.81	0.90	0.80	0.73
450	0.80	0.87	0.81	0.91	0.78	0.70
500	**0.83**	**0.88**	**0.82**	**0.92**	**0.81**	**0.70**

Table 1. Accuracy of TC-SIFT-SPMK methods

Num. of Words	Test1		Test2		Test3	
	P	R	P	R	P	R
100	0.76	0.95	0.77	0.95	0.73	0.89
200	0.74	0.93	0.74	0.96	0.64	0.89
500	**0.77**	**0.93**	**0.77**	**0.95**	**0.73**	**0.84**

Table 2. Accuracy of SIFT-SPMK methods

Feature Vector	Test1		Test2		Test3	
	P	R	P	R	P	R
HOG	0.81	0.73	0.85	0.80	0.70	0.73
MS-HOG-HSV	**0.84**	**0.74**	**0.88**	**0.79**	**0.76**	**0.72**

Table 3. Accuracy of HOG Feature methods

Method	Test1		Test2		Test3	
	P	R	P	R	P	R
TC-SIFT-SPMK	0.83	0.88	0.82	0.92	0.81	0.70
MS-HOG&HSV	0.84	0.74	0.88	0.79	0.76	0.72
Intersection	**0.92**	**0.68**	**0.92**	**0.76**	**0.97**	**0.62**

Table 4. Accuracy of intersection

3. 3 Target Segmentation

In this section, we only focus on the segmentation of buildings. After getting the label of each image patch of the test image, we can get a series of boxes which are labelled with buildings. To guide the user to do target segmentation, we first take these boxes region as masks and get the connected regions of the masks which will be taken as the candidate foregrounds, and the outside region around the foreground will be taken as the candidate background. Then, use the GrabCut algorithm to cut the targets out one by one. The user can add the foregrounds and background through interactive. Figure 8 shows the candidate foregrounds in light white color and Figure 9 gives the segmentation results.

Figure 8. The candidate foregrounds

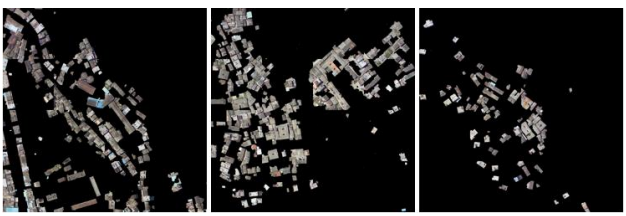


Figure 9. The segmentation results

Usually, the boundary of the segmentations are not fit very well with the true boundary of the targets, especially for buildings which have apparent lines and corners. So, regularize the segmentation results is necessary for buildings. Using the method introduced in 2.4, we get the regularized segmentations. Figure 10 gives the corresponding final extracted targets which after regularizing process. These results indicate that this work flow is feasible way.

Figure 10. The regularized results.

3. 4 Discussion

In the case of target detection, the quantitative evaluation shows that the color information has a positive influence on the detection accuracy. The TC-SIFT-SPMK method get the higher accuracy than the SIFT-SPMK method under the same condition. Similarly, the MS-HOG & HSV feature get higher accuracy than the MS-HOG. At the same time, the MS-HOG & HSV is robust to different spatial resolution images than TC-SIFT-SPMK. In TC-SIFT-SPMK, the number of the visual words influences the detection accuracy and the computational work. Experiments results show that the appropriate number of words should be selected. Through analysing the tendency of the accuracy when changing the number of words, we can find the proper number of words. Each feature representation has its own speciality, the combination of the different feature representations will increase the precision of the detection and give more confidence on the detection results. The size and the spatial resolution of the sample images influence the classifier's applicability.

In the target segmentation process, the candidates of foreground and background and the number of iterations are all have influence on the segmentation results. It really can get very good results. The user still need more experience to do it well. How to make it more effective and easy should be taken into consideration. The good news is that, through experiments, we find the GrabCut algorithm can find out some buildings which are not selected out by the target detection methods, it's a complementing for target detection.

4. CONCLUSION

The proposed workflow is a feasible semi-auto manual interactive target extraction way. It can reduce the operator's workload. But there is still some aspects need to be improved. Firstly, take the probability of the target as the classifier's output and the combination of the different classifiers give the degree of confidence. Secondly, self-organized fit the scale of the sample image to the spatial resolution of the test image is a future work. It will reduce the workload of sample collection work. Thirdly, the boundary regularizing method need to be improved because of the diversity of the dense building area.

ACKNOWLEDGEMENTS

The study was partially supported by the High-resolution Comprehensive Traffic Remote Sensing Application program under Grant No. 07-Y30B10-9001-14/16, the National Natural Science Foundation of China under Grant No. 41101410 and Foundation of Key Laboratory for National Geographic State Monitoring of National Administration of Survey, Mapping and Geoinformation under Grant No. 2014NGCM. During the experiments, Haoying Cui, an undergraduate student of University of Waterloo, helps us selecting the samples and labeling the ground truth. We gratefully thanks to her.

REFERENCES

Abdel-Hakim, A.E. and Farag, A.A., 2006. CSIFT: A SIFT descriptor with color invariant characteristics. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, Vol. 2, pp. 1978-1983. IEEE.

Bengio, Y., Courville, A. and Vincent, P., 2013. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *35*(8), pp.1798-1828.

Beucher S., Meyer F., 1993. The morphological approach to segmentation: the watershed transformation. In *Mathematical Morphology in Image Processing*, pp. 433–481,

Blake, A., Rother, C., Brown, M., Perez, P. and Torr, P., 2004. Interactive image segmentation using an adaptive GMMRF model. In *Computer Vision-ECCV 2004*, pp. 428-441. Springer Berlin Heidelberg.

Blaschke, T., 2010. Object based image analysis for remote sensing. *ISPRS journal of photogrammetry and remote sensing*, *65*(1), pp.2-16.

Bo, L., Ren, X., & Fox, D. 2010. Kernel descriptors for visual recognition. In Advances in neural information processing systems, pp. 244-252.

Bo, L., Lai, K., Ren, X., & Fox, D. 2011. Object recognition with hierarchical kernel descriptors. In *Computer Vision and Pattern Recognition (CVPR)*, 2011 IEEE Conference on, pp. 1729-1736. IEEE.

Boykov, Y.Y. and Jolly, M.P., 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, Vol. 1, pp. 105-112. IEEE.

Burghouts, G.J. and Geusebroek, J.M., 2009. Performance evaluation of local colour invariants. *Computer Vision and Image Understanding*, *113*(1), pp.48-62.

Caselles, V., Kimmel, R. and Sapiro, G., 1997. Geodesic active contours.*International journal of computer vision*, *22*(1), pp.61-79.

Chuang, Y.Y., Curless, B., Salesin, D.H. and Szeliski, R., 2001. A bayesian approach to digital matting. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, Vol. 2, pp. II-264. IEEE.

Comaniciu D, Meer P., 2002. Mean shift: a robust approach toward feature space analysis, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 24(5): 603-619.

Csurka, G., Dance, C., Fan, L., Willamowski, J. and Bray, C., 2004, May. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV* , Vol. 1, No. 1-22, pp. 1-2.

Cui, W. and Zhang, Y., 2011. An effective graph-based hierarchy image segmentation. *Intelligent Automation & Soft Computing*, *17*(7), pp.969-981.

Dalal, N., & Triggs, B., 2005. Histograms of oriented gradients for human detection. *In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1, pp. 886-893. IEEE.

Felzenszwalb, P.F. and Huttenlocher, D.P., 2004. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2), pp.167-181.

Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In*Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580-587.

Grauman, K. and Darrell, T., 2005, October. The pyramid match kernel: Discriminative classification with sets of image features. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, Vol. 2, pp. 1458-1465. IEEE.

Gulshan, V., Rother, C., Criminisi, A., Blake, A. and Zisserman, A., 2010, June. Geodesic star convexity for interactive image segmentation. In*Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 3129-3136. IEEE.

Hu, X., Tao, C. V., & Prenzel, B. 2005. Automatic segmentation of high-resolution satellite imagery by integrating texture, intensity, and color features. *Photogrammetric Engineering & Remote Sensing*, 71(12), 1399-1406.

Huang, Y., Wu, Z., Wang, L. and Tan, T., 2014. Feature coding in image classification: A comprehensive study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *36*(3), pp.493-506.

Kass, M., Witkin, A. and Terzopoulos, D., 1988. Snakes: Active contour models. *International journal of computer vision*, *1*(4), pp.321-331.

Kohli, P., Osokin, A., & Jegelka, S. 2013. A principled deep random field model for image segmentation. In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1971-1978.

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097-1105.

Lazebnik, S., Schmid, C., & Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, Vol. 2, pp. 2169-2178. IEEE.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*(2), pp.91-110.

Mikolajczyk, K., & Schmid, C. 2005. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10), 1615-1630.

Ming, D., Wang, Q. and Yang, J., 2008. Spatial Scale of Remote Sensing Image and Selection of Optimal Spatial Resolution. *Journal of Remote Sensing*, (4), pp.529-537.

Mortensen, E.N. and Barrett, W.A., 1995, September. Intelligent scissors for image composition. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pp. 191-198. ACM.

Montoya-Zegarra, J. A., Wegner, J. D., Ladický, L., & Schindler, K., 2015. Semantic segmentation of aerial images in urban areas with class-specific higher-order cliques. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2(3), 127.

Ning, J., Zhang, L., Zhang, D. and Wu, C., 2010. Interactive image segmentation by maximal similarity based region merging. *Pattern Recognition*, *43*(2), pp.445-456.

Price, B.L., Morse, B. and Cohen, S., 2010, June. Geodesic graph cut for interactive image segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 3161-3168. IEEE.

Rassem, T.H. and Khoo, B.E., 2011, May. Object class recognition using combination of color SIFT descriptors. In *Imaging Systems and Techniques (IST), 2011 IEEE International Conference on*, pp. 290-295. IEEE.

Rother, C., Kolmogorov, V. and Blake, A., 2004, August. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, Vol. 23, No. 3, pp. 309-314. ACM.

Van De Sande, K.E., Gevers, T. and Snoek, C.G., 2008, July. A comparison of color features for visual concept classification. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pp. 141-150. ACM.

Van De Sande, K.E., Gevers, T. and Snoek, C.G., 2010. Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *32*(9), pp.1582-1596.

Veksler, O., 2008. Star shape prior for graph-cut image segmentation. In*Computer Vision–ECCV 2008*, pp. 454-467. Springer Berlin Heidelberg.

Volpi, M., & Ferrari, V., 2015. Semantic segmentation of urban scenes by learning local class interactions. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1-9.

Yang, J., Yu, K., Gong, Y., & Huang, T., 2009, June. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1794-1801. IEEE.

Yang, Y. and Newsam, S., 2011, November. Spatial pyramid co-occurrence for image classification. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 1465-1472. IEEE.

Zhang, H., Li, B., Wang, Y., Zhang, J., & Xu, F., 2013, October. Bag-of-Words model based image classification and evaluation of image sample quality in remote sensing. *In TENCON 2013-2013 IEEE Region 10 Conference* (31194), pp. 1-4. IEEE.

Zhang, C., Wang, S., Huang, Q., Liu, J., Liang, C. and Tian, Q., 2013. Image classification using spatial pyramid robust sparse coding. *Pattern Recognition Letters*, *34*(9), pp.1046-1052.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C. and Torr, P.H., 2015. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1529-1537.

Zhou, L., Zhou, Z., & Hu, D., 2013. Scene classification using a multi-resolution bag-of-features model. *Pattern Recognition*, 46(1), 424-433.