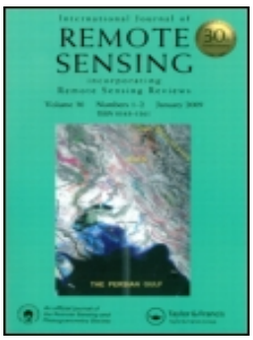

International Journal of Remote Sensing

ISSN: 0143-1161 (Print) 1366-5901 (Online) Journal homepage: http://www.tandfonline.com/loi/tres20

Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification

Wen Shao, Wen Yang & Gui-Song Xia

To cite this article: Wen Shao, Wen Yang & Gui-Song Xia (2013) Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification, International Journal of Remote Sensing, 34:23, 8588-8602, DOI: 10.1080/01431161.2013.845925

To link to this article: http://dx.doi.org/10.1080/01431161.2013.845925

International Journal of Remote Sensing, 2013
Vol. 34, No. 23, 8588–8602, http://dx.doi.org/10.1080/01431161.2013.845925

Taylor & Francis
Taylor & Francis Group

Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification

Wen Shao[a], Wen Yang[a,b]*, and Gui-Song Xia[b]

[a]*School of Electronic Information, Wuhan University, Wuhan 430072, China;* [b]*Key State Laboratory LIESMARS, Wuhan University, Wuhan 430079, China*

(*Received 4 March 2013; accepted 17 June 2013*)

This article presents a hierarchical classification method for high-resolution satellite imagery incorporating extreme value theory (EVT)-based normalization to calibrate multiple-feature scores. First, a simple linear iterative clustering algorithm is used to over-segment an image to build a superpixel representation of the scene. Then, each superpixel is characterized by three different visual descriptors. Finally, a two-phase classification model is proposed for achieving classification of the scene: (1) in the first phase, a support vector machine (SVM) with histogram intersection kernel is applied to each feature channel to obtain raw soft probability; and (2) in the second phase, the derived soft outputs are multiplied to build a product space for score-level fusion. The fused scores are subsequently further calibrated using the EVT and fed to an L1-regularized L2-loss SVM to obtain the final result. Experimental analysis on high-resolution satellite scenes shows that the proposed method achieves promising classification results and outperforms other competitive methods.

1. Introduction

Scene classification has attracted considerable attention in remote sensing. For a high-resolution satellite scene, the classifier must determine all of the categories it contains and the exact label of each pixel. This problem is one of the fundamental challenges in scene classification, especially in the presence of large intra-class variations and other external factors such as illumination variation, shadow, and partial occlusions between objects and objects (e.g. for roads, the occlusions are mainly due to vehicles, trees, shadows of buildings, etc.). Much effort has been devoted to exploiting different sources of information in the satellite image to boost classification performance (Dai and Yang 2011; Mills 2011; Johnson 2012; Stathakis and Vasilakos 2006).

Within a satellite scene, information is delivered in many forms such as structure, texture, and colour. Each individual cue reveals only one aspect of the scene, but in isolation it will not suffice. Given that the information from each cue is ambiguous and incomplete, how to integrate these diverse and complementary features becomes a crucial question. Multiple-kernel learning (MKL) (Lanckriet et al. 2004) is a prominent type of kernel fusion that makes use of kernel functions defining a measure of similarity between pairs of instances. MKL associates a kernel with each image feature and approximates the optimal feature's kernel as the relative weight for a specified task. However, although the MKL

*Corresponding author. Email: yangwen@whu.edu.cn

solution is sparse for every class in isolation, it is not sparse in a multi-class situation. Because even the simple baseline methods 'average' and 'product' are highly competitive with MKL, it may be concluded that the performance of MKL might have been overestimated in the past (Gehler and Nowozin 2009). One widely used approach is score-level fusion, in which scores from different feature channels are fused, opening up new prospects for multi-stage classification. A prominent representative of this approach is a two-stage linear support vector machine (SVM) classification scheme using sparse coding-based multiple-feature combinations (Sheng et al. 2012). In the second stage, the approach simply used score concatenation to combine the intermediate probabilities obtained from the corresponding feature channels in the first stage. Although the approach turned out to work surprisingly well, some improved probability fusion styles (Lu et al. 2011) ought also to be considered.

Another difficulty is choosing a robust score-normalization technique because of the disparate characteristics of the underlying score distributions for different data sources. Since score distributions vary as a function of the classification algorithm, one must normalize the score data before combining them in score-level fusion. *Z*-scores (Poh and Bengio 2005) are adaptive normalization techniques that are easy to compute; however, these *Z*-scores are not robust (sensitive to outliers) and are easily impacted by recognition algorithm failure (if one classification algorithm involved in the fusion process cannot produce a correct matching result, it will strongly impact the final result of fusion). 'tanh' estimators (Hampel et al. 1986) are fixed-score normalizations that are considered robust to noise, but are far more difficult to compute than adaptive *Z*-scores. Ideally, a score-normalization method would not only be robust to failure, but would also not be dominated by complex parameter estimation procedures with variation in score distribution. Normalization using *W*-scores is preferable. A statistical extreme value theory (EVT) normalization technique (Scheirer et al. 2012) has been used to construct normalized 'multi-attribute spaces' from raw classifier outputs. This method calibrates each raw score to the probability that an image exhibits a given attribute, an approach that has significant applications in multi-attribute searches and target-attribute similarity searches. EVT normalization draws the probabilities from the cumulative distribution function (CDF) of a Weibull distribution (hence the term '*W*-score'). The *W*-score renormalizes the data based on the formal probability of each point being an outlier in the extreme value 'non-match' model, thus enhancing the chance of achieving a successful classification. The resulting probabilities are the normalized *W*-scores, which can be fused together to produce an overall probability or used to perform other tasks. A significant benefit of the approach is that the calibration is done after-the-fact, requiring neither modification to the attribute classifier nor the reference attribute annotations.

The main objective of this article is the classification of high-resolution optical remote-sensing data. To this end, a robust two-phase classification model was developed by discriminatively fusing and normalizing the scores from multiple-feature channels. The model begins with a simple linear iterative clustering (SLIC; Achanta et al. 2012) over-segmentation to generate superpixel patches for a satellite scene, an approach that demonstrates satisfactory boundary recall performance (i.e. very few true edges are missed). Considering that different types of classifier have varying degrees of inductive bias (the inductive bias of a learning algorithm is the set of assumptions that the learner uses to predict outputs given inputs that it has not encountered) (Witten and Frank 2005), each classifier is likely to be proficient in a different part of the task. Therefore, a SVM with histogram intersection kernel (Maji, Berg, and Malik 2008) and an L1-regularized L2-loss

SVM (Yuan et al. 2010) are used in different classification phases. Finally, EVT calibration is performed to normalize the fused scores to improve classification accuracy further.

2. SLIC superpixels for scene segmentation

Over-segmentation of an image into superpixels is a common preprocessing step for image-parsing algorithms. Here, over-segmentation means that image regions are segmented into smaller regions, each segment referred to as a 'superpixel'. Ideally, every pixel within each superpixel region will belong to the same real-world object. Operating on superpixels instead of pixels can better retain the boundary information and can speed up existing pixel-based algorithms, such as mean-shift and TurboPixels (Fulkerson, Vedaldi, and Soatto 2009). Superpixel algorithms can be broadly classified into graph-based and gradient ascent-based methods. SLIC (Achanta et al. 2012) is a new superpixel algorithm that produces compact, nearly uniform superpixels by clustering pixels based on their colour similarity and spatial proximity in the image plane. Unlike many other superpixel algorithms, SLIC implicitly enforces connectivity and is simple and memory-efficient in practice, the only parameter needed being the desired number of superpixels. In particular, SLIC has been shown to yield state-of-the-art adherence to image boundaries and to perform better than existing methods when used for segmentation on certain data sets (Martin et al. 2001).

SLIC generates a local clustering of pixels in the five-dimensional $[l, a, b, x, y]^{\mathrm{T}}$ space, where $[l, a, b]^{\mathrm{T}}$ is the pixel colour vector in CIELAB (CIE 1976 L*, a*, b*, dimension L for lightness, a and b for the colour-opponent dimensions, based on nonlinearly compressed CIE XYZ colour space coordinates) colour space and the $[x, y]^{\mathrm{T}}$ is the pixel coordinate. Specifically, for an image with N pixels in the CIELAB colour space, in the initialization step, K superpixel cluster centres $C_k = [l_k, a_k, b_k, x_k, y_k]^{\mathrm{T}}, k \in [1, K]$ are sampled at regular grid steps, and the centres are moved to seed locations at the lowest gradient position in a 3×3 neighbourhood. To produce superpixels of roughly equal size, the grid interval S is set to $\sqrt{N/K}$. This is done to avoid placing superpixels at an edge and to reduce the chance of choosing a noisy pixel. Next, in the assignment step, each pixel k is associated with the nearest cluster centre whose search region overlaps this pixel. After all of the pixels have been associated with the nearest cluster centre, a new centre is computed as the mean vector $[l, a, b, x, y]^{\mathrm{T}}$ of all of the pixels belonging to the cluster. The process of associating pixels with the nearest centre and updating the cluster centre can be repeated iteratively until convergence. In the final step, a post-processing step enforces connectivity by relabelling disjoint pixels with the labels of the largest neighbouring cluster.

3. Multiple-feature extraction

Because a high-resolution satellite scene usually consists of several kinds of information cues, capturing these is very helpful in recognizing and distinguishing different categories. In this paper, we characterize superpixels using three types of feature: SIFT (scale invariant feature transform) descriptors for structural cues, combined scattering for textural cues, and a bag of colours for colour cues.

3.1. SIFT descriptors and BoF representation

SIFT extracts image gradient orientation histograms within a support region. A SIFT descriptor is invariant to image scaling and rotation, and partially invariant to changes

in illumination and the 3D camera viewpoint (Lowe 2004). For each of eight orientation planes, the gradient image is sampled over a 4×4 grid, resulting in a 128-dimensional feature vector for each region. A codebook-based model is usually adopted to summarize the descriptors, termed a bag of features (BoF) (Csurka et al. 2004). This method quantizes local feature descriptors using a visual dictionary typically constructed through k-means clustering. The final representation of an image is the histogram of the quantized SIFT features.

3.2. *Combined scattering*

Combined scattering has been proven highly discriminative for texture classification (Sifre and Mallat 2012). Scattering transform builds invariant, stable, and informative representations through a non-linear, unitary transform, which delocalizes signal information into scattering decomposition paths. A combined scattering is computed using two nested cascades of wavelet transforms, and a complex modulus is used along spatial and rotational variables. The first layer computes statistics for stationary textures through a cascade of wavelet-modulus operators, along a convolutional network (Lecun, Kavukvuoglu, and Farabet 2010). A second layer is built using a similar algorithm, but with convolutions computed along the rotational parameters. Angular information is split in several combined paths that are averaged along the rotational parameter to achieve rotation invariance. The resulting decomposition has the near-complete properties of Fourier spectral methods and the stability of general averaging algorithms.

3.3. *Bag of colours*

Wengert, Douze, and Jégou (2011) proposed a simple yet efficient method that introduces the concept of a Lab colour palette to extract a colour signature, called a bag of colours. This signature is extracted either from the whole image or from patches of the image, producing a global descriptor and a set of local colour descriptors, respectively. In order to collect typical and unique colours from a set of real-world images, a colour codebook $\mathrm{C} = \{c_1, \ldots, c_i, \ldots, c_{k_c}\}$ (c_i represents the i-th colour) referred to as a k_c Lab colour palette, where k_c equals 64 or 512, is first learned. The empirical distribution (i.e. histogram) of colours in an image is then computed according to the fixed colour codebook C. Similar to BoF, the histogram components are updated by multiplying the frequency component with its corresponding colour weight. Finally, to increase the discriminative power of the features, a 'power-law' transformation and an L1 vector normalization are sequentially performed to regularize the contribution of each colour in the final bag-of-colours features.

4. Proposed approach

As mentioned in Section 3, it is necessary to integrate multiple cues to capture all of the information from the image statistics. Accordingly, the ability to fuse multiple cues brings significant improvement in overall classification performance. Using extensive experiments with different configurations, a discriminative two-phase classification framework with EVT calibration after score fusion was designed and is illustrated in Figure 1. Observe that our method is also flexible to other extensions, such as new features, classifiers, and fusion styles.

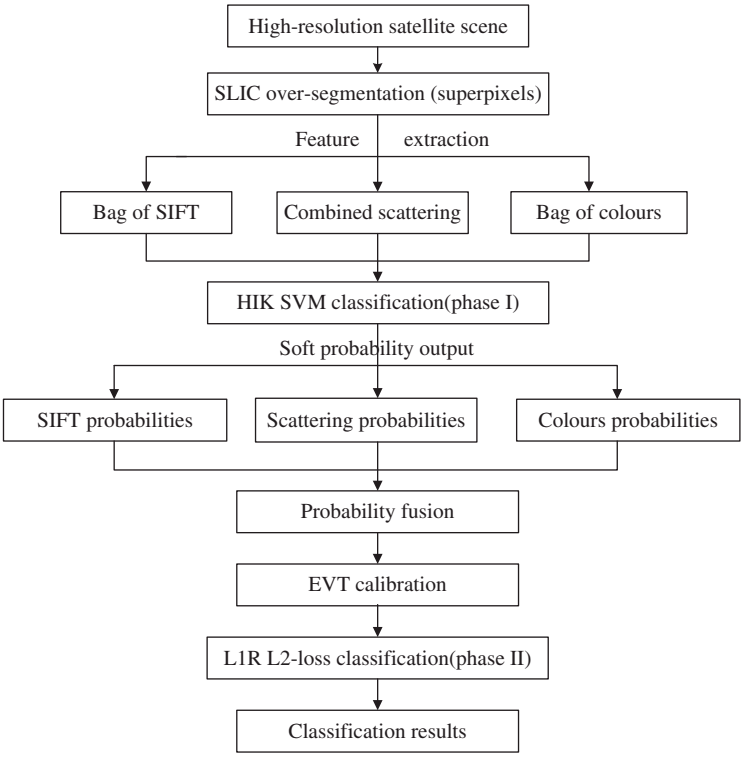

Figure 1. Flow chart of the proposed hierarchical classification model.

4.1. Classification model of phase I

The SVM with histogram intersection kernel (HIK SVM) (Maji, Berg, and Malik 2008) remains an extremely popular alternative for histogram feature classification in terms of both predictive power and running time. Moreover, the HIK SVM is able to provide probabilistic output for the test image. After feature extraction, an HIK SVM was used as the base classifier in the first phase. Thus, for three feature channels of an image I, three soft *posteriori* probability vectors $\boldsymbol{P}_m(I)$ ($m = 1,2,3$) can be estimated in parallel using the HIK SVM.

The choice of a suitable fusion method is extremely important for this work. Several score fusion methods have been reported in the literature (Lu et al. 2011; Sheng et al. 2012), among which score multiplication is one of the most feasible score-level fusion styles because multiplying the scores from all feature channels together would tend to favour categories where all features are somewhat likely, but none is particularly low. In addition, a multi-attribute space is usually a product space formed from different attribute functions (Scheirer et al. 2012). Therefore, multiplicative fusion was performed to combine intermediate soft probabilities from three feature channels:

$$\boldsymbol{M}(I) = \prod\nolimits_{m=1}^{3} \boldsymbol{P}_m(I), \tag{1}$$

where $\boldsymbol{M}(I)$ is a product space formed from well normalized probability vectors, and the length of $\boldsymbol{P}_m(I)$ is equivalent to the number of categories within a scene.

4.2. Calibration of fused scores using EVT

Score normalization before fusion has attracted much attention from many researchers (Poh 2010; Scheirer et al. 2012). It is also essential to perform an appropriate score normalization after fusion because of the irregular distribution of fused scores. An optimal score-normalization method should be robust to model assumptions, modelling errors, and parameter estimation errors, as well as to algorithm failure. Similar to multi-attribute space calibration, a statistical EVT calibration was used to extract the probabilities from the CDF of a Weibull distribution for the feature channel under consideration.

4.2.1. Extreme value theory

Suppose X_i, $i = 1, \ldots, n$ are independent random variables from the same probability distribution $F(x)$ (independent and identically distributed, *i.i.d.*), and $M_n = \max\{X_1, X_2, \ldots, X_n\}$ denotes the maximum of the first n observations, then we can say that M_n converges to an extreme value distribution. Under certain circumstances, it can be shown that there exist normalizing constants $a_n (a_n > 0)$, b_n such that

$$\lim_{x\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{x\to\infty} F^n(a_n x + b_n) = H(x). \tag{2}$$

Here, P represents probability distribution and x denotes the extreme value of maxima. Then if H is a non-degenerate distribution function, it belongs to one of three extreme value distributions: Gumbel, Frechet, and Weibull (Scheirer et al. 2010). The three types may be combined into a single generalized extreme value (GEV) distribution, represented here by '$G(x)$':

$$G(x) = \begin{cases} \frac{1}{\lambda} e^{-v^{-1/\xi}} v^{-(1/\xi+1)} & \xi \neq 0 \\ \frac{1}{\lambda} e^{-(z+e^{-z})} & \xi = 0, \end{cases} \tag{3}$$

where $z = \frac{x-\tau}{\lambda}$, $v = (1 + \xi z)$; τ, $\lambda(\lambda > 0)$, and ξ are the location, scale, and shape parameters, respectively. Meanwhile, $\xi = 0$ corresponds to the Gumbel distribution, $\xi > 0$ to the Frechet distribution, and $\xi < 0$ to the reversed Weibull distribution. Gumbel and Frechet are for unbounded distributions, and Weibull for bounded distribution. The EVT is analogous to a central-limit theorem, but with maxima (or minima) over a large collection of random observations from an arbitrary distribution. Gumbel (1954) proved that if a system/part has multiple failure modes, the failure can be better modelled by the Weibull distribution.

4.2.2. EVT calibration

In a multi-class categorization scenario, a single input and a set of output scores we want to normalize, one for each class, are given. The distribution of score values around 1 accounts for a large proportion, and yet these values are the least informative. However, the distribution of scores around 0 is much more informative and also more constrained. According to EVT, the sampling of the top-n scores always results in a Weibull distribution. The EVT fit shows that with the probability of a score being an outlier (the 'non-match distribution'), a correct match can be robustly estimated from only the top-n values, not surpassing half of

the total scores. This is the tail of the non-match distribution, called its 'extreme values'. If the top score is an outlier (match is correct), then excluding it does not impact the fitting. If the top score is not a match, excluding it in the fitting will produce higher probability scores for the correct match and most of the non-matches. Furthermore, it is important to choose the appropriate tail size n used for fitting, which is the parameter needed to be estimated for the Weibull score. Many works (Kotz and Nadarajah 2001) have reported that 3–5 is a very common fitting size range for Weibulls. Once the fitting has taken place, we have all of the information required to achieve the calibration.

Specifically, given a shape parameter $\xi > 0$ and a scale parameter $\lambda > 0$, the calibration process first applies a transform T that flips and shifts each raw SVM score s_d as necessary to satisfy the conditions given by a CDF:

$$\mathrm{CDF}(x; \xi, \lambda) = 1 - e^{-\left(\frac{x}{\lambda}\right)^{\xi}}, \tag{4}$$

where it must be guaranteed that the data x are always positive and that the extreme values are the peak scores among the set of decision scores $S = \{s_d\}, d = 1, \ldots, D$. The next steps are to apply a Weibull distribution $W(\xi, \lambda)$ to the transformed scores and finally to normalize each score using its CDF: $\mathrm{CDF}(T(s_d); W)$. The whole EVT normalization algorithm is shown in Table 1.

Based on the work described above, the combined scores $\boldsymbol{M}(I)$ can be further calibrated by fitting a Weibull distribution W to the extreme values of the non-match distribution. In this case, the scores represent probabilities. Given the Weibull fit to the data, it is possible to determine whether the top score is an outlier by considering the magnitude of the CDF. The CDF of this distribution is used to generate the normalized attribute W-scores. Thus, the final score function can be rewritten as

$$\boldsymbol{S}(I) = \mathrm{CDF}(T(\boldsymbol{M}(I)); W). \tag{5}$$

During EVT normalization, it is necessary to choose an appropriate tail size n, in our case $n = 8$. The parameter configuration of ξ and λ also greatly affects the distribution of the calibrated values. In this research, these two parameters were selected by a linear grid search and, as shown in Section 5, 0–2 was a proper size range for ξ and λ of different features and combinations, which produced good normalizations.

Table 1. EVT score-normalization algorithm.

EVT normalization
Input : A collection of decision scores $S = \{s_d\}$, of length D.
Output : Normalized scores $S' = \{s'_d\}$.
1. Flip and shift each raw SVM score s_d, so extrema are the largest;
2. Sort and retain the top-n scores: $s_1, \ldots, s_n \in \boldsymbol{S}$;
3. Fit a Weibull distribution W to $s_1, \ldots, s_n$, skipping the hypothesized outlier;
4. **while** $d < D$ **do**
5. $s'_d = \mathrm{CDF}(T(s_d); W)$
6. d $\leftarrow$ d + 1
7. **end while**
8. Return normalized decision scores $\{s'_d\}$.

4.3. Classification model of phase II

To maintain data consistency, most of the multi-phase classification task used multiple classifiers of the same type (Lu et al. 2011; Sheng et al. 2012). However, different algorithms used as base classifiers may be more capable of helping each other by focusing on different aspects of the data and seldom making the same mistakes (Koprinska, Deng, and Feger 2006). Therefore, an L1-regularized L2-loss SVM (L1R L2-loss SVM) (Yuan et al. 2010) was used in the second phase. This choice was made for two reasons. On the one hand, L1-regularization provides for robustness in the case that no categories match all feature attributes of the image. Furthermore, an L1-regularization term also has the advantage of avoiding over-fitting when a large number of parameters must be learned. On the other hand, unlike the L2-regularization which restricts large values, the L1-regularization term penalizes all factors equally. Given a set of image–label pairs $(\boldsymbol{x}_l, y_l)$, $l = 1, \ldots, L$, $y_l \in \{-1, +1\}$, L1-regularization together with L2-loss that assumes the differentiability of the loss functions can solve the following primal optimization problem:

$$\min_{\boldsymbol{\omega}} ||\boldsymbol{\omega}||_1 + C \sum_{l=1}^{L} (\max(0, 1 - y_l \boldsymbol{\omega}^{\mathrm{T}} \boldsymbol{x}_l)), \tag{6}$$

where $||\cdot||_1$ denotes the L1-norm, $C > 0$ is a penalty parameter, and $\boldsymbol{\omega}$ is the weight vector we wish to learn. For multi-class problems, we implement the one-*versus*-the-rest strategy by using a set of binary classifiers and taking the majority vote. Here, the actual input that is provided to the phase-II classifier is the combination of three soft probability vectors corresponding to the three feature channels from the phase-I HIK SVM classification.

5. Experiments and results

5.1. Data set and set-up

Experiments were performed on a high-resolution satellite view with a size of 4000 pixels × 4000 pixels, as shown in Figure 2. The view was captured by the GeoEye-1 satellite on 21 November 2009 at Majuqiao town, which lies to the southwest of Tongzhou, southeast of Beijing. The ground sampling distance is approximately 0.5 m. The view includes mainly eight categories of image: bare land, factories, farmland, green space, high buildings, waters, roads, and low buildings. The reference image shown in Figure 5(*b*) was manually labelled using associated geographic information. Additionally, some un-interpretable pixels were labelled as void (white), which can be ignored, such as the interval between one green space and its neighbours.

In the experimental setup, the segmented superpixels using SLIC were approximately 64 pixels in size. Figure 3 shows an enlarged superpixel segmentation result using the SLIC algorithm. These irregular superpixels were extracted using their minimum bounding rectangles and normalized to 80 pixels × 80 pixels for ease of processing. For each superpixel, the class occupying the largest proportion of the patch was chosen as its category, resulting in 1180 bare-land areas, 306 factory areas, 1436 farmland areas, 319 green-space areas, 243 high-building areas, 86 water areas, 256 road areas, and 136 low-building areas. Figure 4 shows two examples of each class from the eight-class satellite scene. To obtain reliable results, we ran the overall experiments on different proportions of training/test samples, and found that classification accuracy gradually increased with an increasing number of training samples. To avoid the problem that training and test samples are too spatially correlated, 10% of each class was chosen for training with the remainder used for testing.

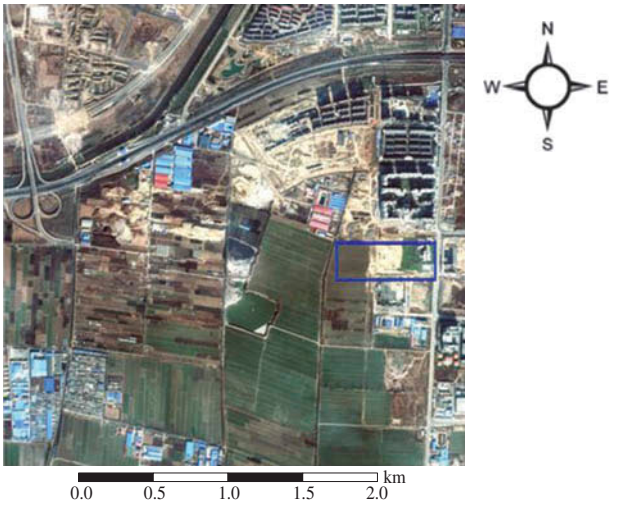

Figure 2. High-resolution satellite scene, captured by the GeoEye-1 satellite on 21 November 2009 at Majuqiao town, which lies in the southwest of Tongzhou, southeast of Beijing, where the latitudes and longitudes at the upper left and lower right corners are 39° 44′ N, 116° 30′ E, and 39° 43′ N, 116° 32′ E, respectively. Band assignment: red for band 3, green for band 2, and blue for band 1. See Figure 3 for a detailed view of the region in the red rectangle.

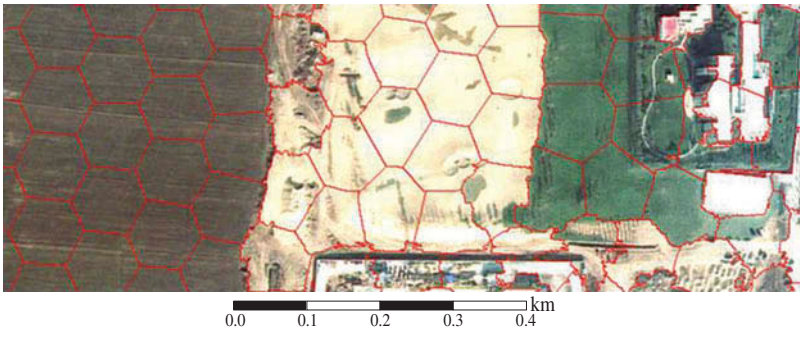

Figure 3. The superpixel segmentation result obtained by the SLIC algorithm for the area corresponding to the blue rectangle marked in Figure 2.

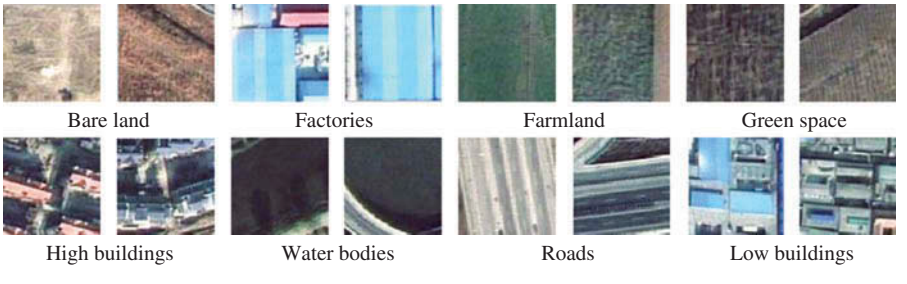

Figure 4. Examples of each class in the eight-class satellite scene.

Meanwhile, we redid the two-phase classification experiments with training data independent of test data (i.e. training data were not selected from the same high-resolution satellite scene), which also obtained comparable performances. Since the 10% proportion of training samples was not too large in our set-up, the problem that the classifiers were learning biases within this particular image was not obvious and thus would not overly inflate classification performance. The experiments were repeated over five random training/test data splits, and the final result was reported as the mean and standard deviation of the results from the individual runs.

5.2. *Experimental results*

In the present implementation, three descriptors were used: bag of SIFT, combined scattering, and bag of colours. The parameter configuration was selected by a linear search process. SIFT descriptors from 16 pixel × 16 pixel patches were densely extracted from each superpixel on a grid with a spacing of five pixels. During BoF processing we varied the visual dictionary and, with linearly increasing size, accuracy gradually increased and finally tended to convergence. The same adjustments were also made for the bag of colours. Tables 2 and 3 summarize the results of these configurations for bag and SIFT and bag of colours, respectively. In order to create a trade-off between classification accuracy and computational cost, we set codebook sizes of 1024 and 512 separately for bag of SIFT and bag of colours. In addition, a dimension of 595 was kept for the combined scattering to achieve satisfactory classification accuracy.

Table 4 summarizes the classification accuracy for each feature channel of one-phase classification using HIK SVM without EVT calibration. All of the feature channels worked

Table 2. Performance comparison of different dictionary sizes for bag of SIFT (one-phase classification: HIK SVM).

Dictionary size	128	256	512	1024	2048
Accuracy (%)	68.95 ± 0.73	70.22 ± 0.93	71.97 ± 0.82	72.83 ± 0.96	72.73 ± 0.83

Table 3. Performance comparison of different dictionary sizes for bag of colours (one-phase classification: HIK SVM).

Dictionary size	128	256	512	1024
Accuracy (%)	73.21 ± 0.78	73.82 ± 0.70	74.26 ± 0.81	73.96 ± 0.62

Table 4. Best classification results for different types of feature without EVT calibration (one-phase classification: HIK SVM).

Feature	Dimension	Accuracy (%)
Bag of SIFT	1024	72.83 ± 0.96
Combined scattering	595	65.66 ± 0.69
Bag of colours	512	74.26 ± 0.81
Concatenation	2131	77.69 ± 0.47

Table 5. The best classification results for different types of feature with EVT calibration (two-phase classifications: HIK SVM (phase I), L1-regularized L2-loss (L1R L2-loss) SVM (phase II)).

Feature for EVT	Dimension	ξ	λ	Accuracy (%)
Bag of SIFT	8	2	0.5	72.99 ± 0.70
Combined scattering	8	1	1.8	65.78 ± 0.81
Bag of colours	8	0.8	1	74.62 ± 0.43
Concatenation	24	—	—	78.32 ± 0.59

very well, but the bag-of-colours method obtained the best result. Simultaneously, the *posteriori* probability from each channel was normalized by EVT calibration and fed into an L1R L2-loss SVM for further classification; the classification results and the corresponding values of ξ and λ are listed in Table 5, where feature concatenation after EVT calibration is also included. Given that each element of the probability scores that denoted the probability of one superpixel belonging to one class was indispensable, the tail size used for fitting for all EVT calibration experiments was 8. It is clear that the accuracy of results for each channel increases after calibration, which can be explained by the observation that EVT calibration cuts off the noise on the tail of the score distributions.

To boost classification performance, it is of crucial importance to combine these three features in a reasonable way; feature concatenation is a general approach that offers no obvious improvement for this data set. For purposes of comparison, four probability fusion strategies (maximum, sum, concatenation, and multiplication) were proposed to combine the intermediate probabilities from the three channels. It is clear from Table 6 that probability multiplication fusion shows a certain advantage over other fusion strategies. The reasons are summarized as follows. On the one hand, in order to capture the comprehensive source of image information, we utilized three representative features separately representing structure, texture, and colour attributes, and these features are relatively independent, which exactly meets the requirement of multiplication fusion. On the other hand, the choice of L1R L2-loss SVM is important since L1-norm provides for robustness in the case where no categories match all feature attributes of the image. In contrast, multiplying probabilities from all feature channels would tend to favour the category where all features are somewhat likely, but none is particularly low. However, for max, which computes the maximum of the probabilities from all feature channels, it is probably preferable to return the category that has high probabilities for $m - 1$ or $m - 2$ ($m = 3$) features, and perhaps not as high for the remaining one or two. For sum and concatenation, since the advantages of high probability can make up for the disadvantages of low probability, if the probability from one or two features is particularly low, it will have no noticeable impact on the final classification accuracy.

Table 6. Performance comparison of different probability fusion strategies (two-phase classifications: HIK SVM (phase I), L1R L2-loss SVM (phase II)).

Probability fusion method	Dimension	Accuracy (%)
Max	8	77.88 ± 0.68
Sum	8	78.97 ± 0.35
Concatenation	24	78.32 ± 0.59
Multiplication	8	79.31 ± 0.62

Table 7. Performance comparison of EVT calibration for probability multiplication fusion (PMF) in different phases (two-phase classifications: HIK SVM (phase I), L1R L2-loss SVM (phase II)).

EVT calibration type	Dimension	Accuracy (%)
EVT before PMF	8	79.02 ± 0.45
EVT after PMF	8	80.57 ± 0.64
Two EVTs (simultaneous)	8	79.73 ± 0.42

In the following, the EVT process was performed under probability multiplication fusion, simplified as PMF, unless stated otherwise. Unlike multi-attribute space calibration (Scheirer et al. 2012), which only uses one EVT calibration separately for each feature channel before PMF, the present research also performed EVT calibration after PMF (with $\xi = 0.5$ and $\lambda = 1.5$), as well as two EVT calibrations simultaneously before and after PMF (the values of ξ and λ were optimally selected each time as mentioned above) to normalize the results. Table 7 compares the classification performance of EVT calibration for PMF in different phases, clearly demonstrating the superiority of EVT after PMF. Because EVT calibration cut off the tail of the score distributions so that some constrained score information around zero was lost, this information was already lost when performing PMF, while there is no need to worry about information losses for EVT calibration after PMF. In other words, it was better to perform EVT calibration before the last classification phase.

In addition, two additional experiments using state-of-the-art methods were conducted on the eight-class satellite scene, one being 'logistic regression-based feature fusion (LRFF)' (Fernando et al. 2012) and the other 'sparse coding-based multiple feature combination (SCMF)' (Sheng et al. 2012). Here, we used their features as well as classification methods with the optimal parameter configuration. Specifically, in the LRFF method, the descriptors 'SIFT + Hue + Colour Names (CN) + Opponent.SIFT (Opp.SIFT)' had sizes of 1000 + 300 + 300 + 2000. In the SCMF method, SIFT, colour histrogram (CH), and local ternary pattern histogram Fourier (LTP-HF) descriptors were processed with the same dictionary size of 512. Table 8 shows the performance comparison of these representative methods.

In particular, the proposed method with EVT after PMF achieved the highest accuracy of 81.31%. Figures 5(*a*) and (*b*) compare the best classified image with the ground-truth image for the satellite scene, and the corresponding confusion matrix is presented in Table 9. Note that the values on the last line represent the total number of superpixels in each class, the diagonal values reported are the actual superpixel numbers assigned to the correct class, and the rest are misclassified. Not surprisingly, confusion was more likely to occur between the low-building and factory classes, because low buildings often contain patterns such as dense houses and horizontal and vertical lines, which are also characteristics of factories; and some superpixels that should belong to green space were misclassified as bare land or farmland, mainly because these categories share similar patterns and components. However, confusions between some classes were difficult to explain. For instance,

Table 8. Performance comparison of state-of-the art methods.

Method	LRFF	SCMF	Our EVT after PMF
Accuracy (%)	79.68 ± 0.49	78.97 ± 0.27	80.57 ± 0.64

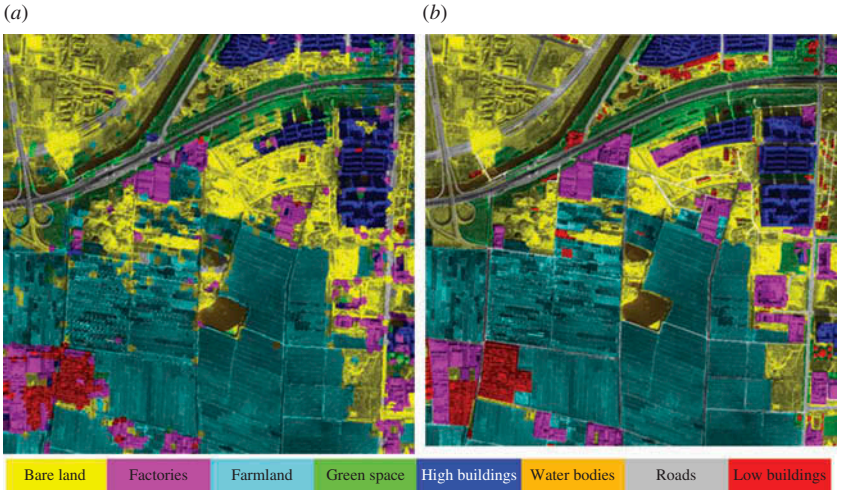

Figure 5. Classification results for the eight-class satellite scene. (*a*) The best classified image using our method. (*b*) Hand-labelled reference data.

Table 9. Confusion matrix of the best classification results for the high-resolution eight-class satellite scene, where the diagonal values are the actual numbers of superpixels assigned to the correct class and the remainder are misclassified.

	Reference data							
Mapped data	Bare land	Factories	Farmland	Green space	High buildings	Water bodies	Roads	Low buildings
Bare land	914	45	88	46	18	8	49	21
Factories	25	183	8	7	10	1	8	33
Farmland	54	5	1182	37	1	2	6	3
Green space	35	5	10	182	5	7	5	5
High buildings	18	11	1	11	180	0	11	3
Water bodies	3	1	1	2	0	56	0	0
Roads	12	6	2	2	3	3	151	8
Low buildings	1	19	0	0	1	0	0	49
Total	1062	275	1292	287	218	77	230	122

some factories, roads, and low buildings were classified as bare land. The occurrence of these ambiguities and obscurities may have been caused by image corruption introduced by over-segmented superpixels and the unknown classes they contain. Therefore, future work will focus on the mining and use of the saliency features and context information of the terrain categories (Hu et al. 2013).

6. Conclusions

In this article, an EVT-based score-level fusion and normalization approach was proposed for high-resolution satellite scene classification; in particular, a two-phase classification model was introduced. Extensive experiments on a given satellite scene demonstrate that with the optimal configuration of feature descriptors and fusion style, our hierarchical classification model with EVT calibration after probability multiplication fusion can yield

promising results. However, the current method is limited in the sense that it works on individual patches without exploiting context elements such as the statistical relations between the appearance of superpixels and their location. In future work, the authors plan to introduce conditional random fields to take advantage of multi-scale context information.

Funding

The research was supported in part by the National Key Basic Research and Development Programme of China under Contract 2013CB733404; the Chinese National Natural Sciences Foundation (NSFC) [grant number 61271401, 61331016].

References

Achanta, R., A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. 2012. "SLIC Superpixels Compared to Start-of-the-Art Superpixels Methods." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 34: 2274–2282.

Csurka, G., C. Dance, L. Fan, J. Willamowski, and C. Bray. 2004. "Visual Categorization with Bags of Keypoints." In *Proceedings of ECCV 2004 Workshop: Statistical Learning in Computer Vision*, May 15, Prague, 59–74.

Dai, D.-X., and W. Yang. 2011. "Satellite Image Classification via Two-Layer Sparse Coding with Biased Image Representation." *IEEE Geoscience and Remote Sensing Letters* 8: 173–176.

Fernando, B., E. Fromont, D. Muselet, and M. Sebban. 2012. "Discriminative Feature Fusion for Image Classification." In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Providence, RI, June 16–21, 3434–3441.

Fulkerson, B., A. Vedaldi, and S. Soatto. 2009. "Class Segmentation and Object Localization with Superpixel Neighbourhoods." In *Proceedings of IEEE International Conference on Computer Vision*, Kyoto, September 29–October 2, 670–677.

Gehler, P., and S. Nowozin. 2009. "On Feature Combination Methods for Multiclass Object Classification." In *Proceedings of IEEE International Conference on Computer Vision*, Kyoto, September 29–October 2, 221–228.

Gumbel, E. 1954. "Statistical Theory of Extreme Values and Some Practical Applications." Number National Bureau of Standards Applied Mathematics in 33. Washington, DC: US GPO.

Hampel, F., P. Rousseeuw, E. Ronchetti, and W. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Hu, F., W. Yang, J.-Y. Chen, and H. Sun. 2013. "Tile-Level Annotation of Satellite Images Using Multi-Level Max-Margin Discriminative Random Field." *Remote Sensing* 5: 2275–2291.

Johnson, B. A. 2012. "High-Resolution Urban Land-Cover Classification Using a Competitive Multi-Scale Object-Based Approach." *Remote Sensing Letters* 3: 737–746.

Koprinska, I., D. Deng, and F. Feger. 2006. "Image Classification Using Labelled and Unlabelled Data." In *Proceedings of the 14th European Signal Processing Conference* (EUSIPCO 2006), Florence, September 4–8.

Kotz, S., and S. Nadarajah. 2001. *Extreme Value Distributions: Theory and Applications*. 1st ed. Singapore: World Scientific Publishing.

Lanckriet, G. R. G., N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. 2004. "Learning the Kernel Matrix with Semi Definite Programming." *Journal of Machine Learning Research* 5: 27–72.

Lecun, Y., K. Kavukvuoglu, and C. Farabet. 2010. "Convolutional Networks and Applications Invision." In *Proceedings of IEEE International Symposium on Circuits and Systems* (ISCAS), Paris, May 30–June 2, 253–256.

Lowe, D. G. 2004. "Distinctive Image Features From Scale-Invariant Keypoints." *International Journal of Computer Vision* 60: 91–110.

Lu, F.-X., X.-K. Yang, W.-Y. Lin, R. Zhang, and S.-Y. Yu. 2011. "Image Classification with Multiple Feature Channels." *Optical Engineering* 50: 057210–057210-9.

Maji, S., A. C. Berg, and J. Malik. 2008. "Classification Using Intersection Kernel Support Vector Machines is Efficient." In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, June 24–26, 1–8.

Martin, D., C. Fowlkes, D. Tal, and J. Malik. 2001. "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics." In *Proceedings of IEEE International Conference on Computer Vision*, Vol. 2, Vancouver, BC, July 7–14, 416–423.

Mills, P. 2011. "Efficient Statistical Classification of Satellite Measurements." *International Journal of Remote Sensing* 32: 6109–6132.

Poh, N. 2010. "User-Specific Score Normalization and Fusion for Biometric Person Recognition." *In Advanced Topics in Biometrics* 16: 401–418.

Poh, N., and S. Bengio. 2005. "How Do Correlation and Variance of Base Classifiers Affect Fusion in Biometric Authentication Tasks?" *IEEE Transactions on Signal Processing* 53: 4384–4396.

Scheirer, W. J., N. Kumar, P. N. Belhumeur, and T. E. Boult. 2012. "Multi-attribute Spaces Calibration for Attribute Fusion and Similarity Search." In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Providence, RI, June 16–21, 2933–2940.

Scheirer, W. J., A. Rocha, R. Micheals, and T. E. Boult. 2010. "Robust Fusion: Extreme Value Theory for Recognition Score Normalization." In *Proceedings of European Conference on Computer Vision*, Heraklion, September 5–11, 1–14.

Sheng, G.-F., W. Yang, T. Xu, and H. Sun. 2012. "High-Resolution Satellite Scene Classification Using Sparse Coding Based Multiple Features Combination." *International Journal of Remote Sensing* 33: 2395–2412.

Sifre, L., and S. Mallat. 2012. "Combined Scattering for Rotation Invariant Texture Analysis." In *Proceedings of European Symposium on Artificial Neural Networks*, *Computational Intelligence and Machine Learning* (ESANN), Bruges, April 27–29, 127–132.

Stathakis, D., and A. Vasilakos. 2006. "Satellite Image Classification Using Granular Neural Networks." *International Journal of Remote Sensing* 27: 3991–4003.

Wengert, C., M. Douze, and H. Jégou. 2011. "Bag-Of-Colours for Improved Image Search." In *Proceedings of the 19th ACM international conference on Multimedia*, Scottsdale, AZ, November 28–December 1, 1437–1440.

Witten, I., and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools* and *Techniques*. 3rd ed. Burlington: The Morgan Kaufmann Series in Data Management Systems.

Yuan, G.-X., K.-W. Chang, C. J. Hsieh, and C. J. Lin. 2010. "A Comparison of Optimization Methods and Software for Large-Scale L1-Regularized Linear Classification." *Journal of Machine Learning* 11: 3183–3234.