A New Approach to Terrain Classification Using Three-Dimensional Features [*]

Xiaoguang Wang, Frank Stolle, Howard Schultz,
Edward M. Riseman, and Allen R. Hanson

Department of Computer Science
University of Massachusetts
Amherst, MA 01003
Email: {xwang, stolle, hschultz, hanson, riseman}@cs.umass.edu

Abstract

Texture has long been regarded as spatial distributions of image gray-level variation, and texture analysis has generally been confined to the 2-D image domain. Introducing the concept of "3-D world texture", this paper considers texture as a function of 3-D structures and proposes a set of "3-D textural features". The proposed 3-D features appear to have a great potential in terrain classification. Experiments have been carried out to compare the 3-D features with a popular traditional 2-D feature set. The results show that the 3-D features significantly outperform the 2-D features in terms of classification accuracy and training data reliability.

Keywords: texture analysis, terrain classification, image segmentation, 3-D world texture, 3-D textural feature

[*] This work was funded in part under DARPA/Army TEC contract DACA76-92-C-0041, DARPA/ARL contract DAAL02-91-K-0047, NSF grant CDA-8922572, and ARO DURIP grant DAAH04-95-1-0068.

1 Introduction

Texture analysis is an important area in computer vision and has been extensively studied (e.g. [1, 2, 3, 4]), although it is impossible to review the extensive literature here. While it is widely accepted that *texture* has no formal and precise definition [2, 4, 5], it is generally presumed to be a spatial distributions of gray-level variations, or regular structural "patterns", in the image. This presumption has dominated texture analysis for decades. Tamura et al. [1] distinguished six attributes of texture: coarseness, contrast, directionality, line-likeness, regularity, and roughness. These attributes describe how gray values change in the 2-D image space, and many texture analysis algorithms have been proposed [2, 3, 4] based on these and similar attributes. Tuceryan and Jain [4] identified four basic approaches to texture analysis: statistical, geometrical, model-based, and signal processing methods. Although the methodologies vary substantially from one algorithm to another, they generally assume that texture is a part of the 2-D image and analysis is performed in image space. Typically, the gray-level patterns are characterized as textural features that are extracted in a localized image region. Recent studies, such as local frequency [5], Gabor elementary functions [6], fractal dimension [7], and combinations of multichannel filtering and neural networks [8], are all based on this approach.

The limitation of this traditional approach is that the 3-D properties of real world textures are not used directly. Consider the concepts *3-D world texture* and *2-D image texture* that we will use in this paper. World textures are reoccurring patterns caused by physical coarseness, roughness, and other characteristics on surfaces of objects in the real world, and generally have 3-D structures. For example, forests and grass covers possess different roughness because trees and grass have different 3-D structures. On the other hand, image textures are 2-D optical patterns in the image, reflecting perspective projection of object surfaces and world textures under particular sensor and illumination configurations. Attempting to characterize image textures by image features does not focus upon the underlying physical properties of the surfaces or objects creating the texture.

The traditional meaning of texture is equivalent to image texture defined here. All the feature extraction techniques in literature [2, 3, 4] are based on the concept of image texture, attempting to describe and discriminate 2-D gray-level patterns in the image. We call them *2-D features* in this paper. In contrast, a feature is called a *3-D feature* if it is a measurement of some 3-D structural characteristics of a world texture. An accurate and thorough recovery of object surface structure is usually a goal of 3-D reconstruction, although usually quite difficult. However, for classification such accuracy is unnecessary. A significant amount of 3-D structural information relating to world texture can be obtained from multiple views of an object [9].

As an example, consider the pair of oblique views of rural terrain of Ft. Hood, Texas, shown in Fig. 1(a). They are 2k×2k images taken with an aerial survey camera from an altitude of 2600m at tilts of 53° and 36°, respectively, off of vertical (nadir), with a camera separation of 3998m. Fig. 1 (b) and (c) show the details of two small areas in the terrain after an epipolar resampling from the source images. Some significant differences occur between the image patches in Fig. 1(b) from the two views. Because the trees are off the ground plane, different parts of the trees are seen from different viewpoints. In addition, the crowns of the trees stretch out and occlude parts of the ground when seen from an oblique angle. However, the ground and road in Fig. 1(c) look quite similar in the two views, because they are almost flat, producing little perspective distortion. Generally speaking, a complicated (i.e. highly varying) 3-D structure such as vegetation tends to cause more variations in the different views than simple objects such as roads with simpler 3-D structures (i.e. less variation in depth). Hence, a measure of similarity in the image patches of the same location from different views reveals some information of the texture types at the location, and potentially can be used as

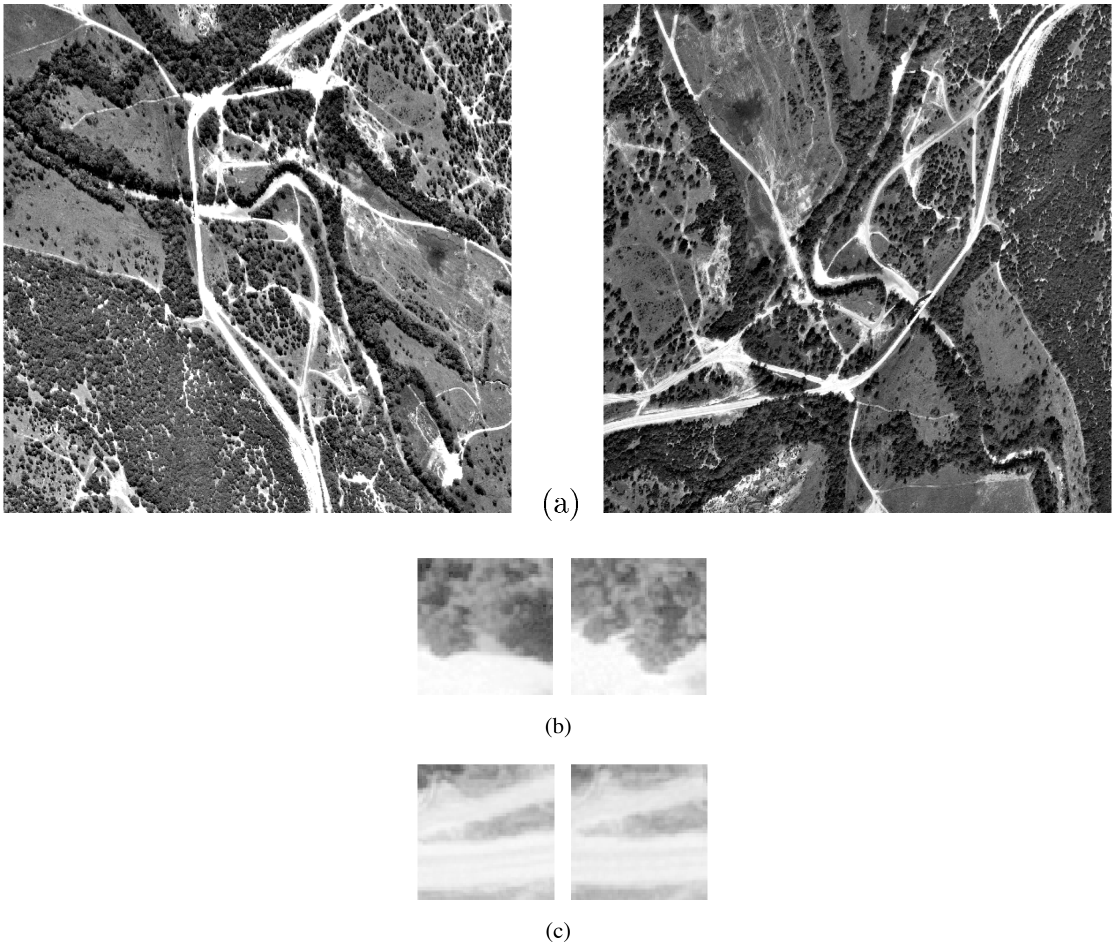


Figure 1: Sample source images from Ft. Hood Image Set

(a) two original source images

(b)(c) epipolar resamplings of two small areas from the two original images

a feature for texture classification. In summary, this information is directly a function of the 3-D variation in the scene.

It is worth noting that these new kinds of features have inherently different physical meanings from the traditional 2-D features. 2-D features typically characterize the relationship between a local area with its neighborhood in the image. The new features, however, characterize the relationship between multiple views of a local area, and hence reveal some 3-D characteristics of the world texture.

We develop a set of 3-D textural features in Section 2. A set of experiments has been designed to compare the performance of the proposed 3-D features and a set of well-known 2-D features. Section 3 describes the set-up of the experiments, such as classifiers, the comparison 2-D features, and training data. Experimental results are shown and analyzed in Section 4 and 5. In Section 6 we discuss future work.

2 Three Dimensional Features

In this section we propose a set of 3-D textural features that will be used in terrain classification. Four types of features are generated, namely, match score (MS), correlation curvature (CSF), neighborhood variation of match score (NVMS), and neighborhood density of well-defined curvature (NDC).

2.1 Similarity function

The proposed 3-D features are derived from a *similarity function* using a multi-view stereo terrain reconstruction algorithm by Schultz [10]. Input to this algorithm are two aerial or satellite images, denoted as I^F and I^G, of the same region on the surface. The algorithm utilizes a hierarchical correlation matching scheme to find pixel-wise correspondences and elevation estimates from the input images. It is mainly designed for image pairs where the cameras have a relatively large baseline to height ratio, i.e. the cameras are far apart. Fig. 2 conceptually shows a similarity function of a 2-D point in the left image in (a). This 2-D point is correlated along its epipolar line on the right image, and the similarity function, shown in (b), indicates how the point is correlated with the points on the epipolar line. The point with the best match – the highest correlation – on the epipolar line tends to be the true correspondence of the point in the left image.

Using known intrinsic camera parameters and relative exterior orientation between the

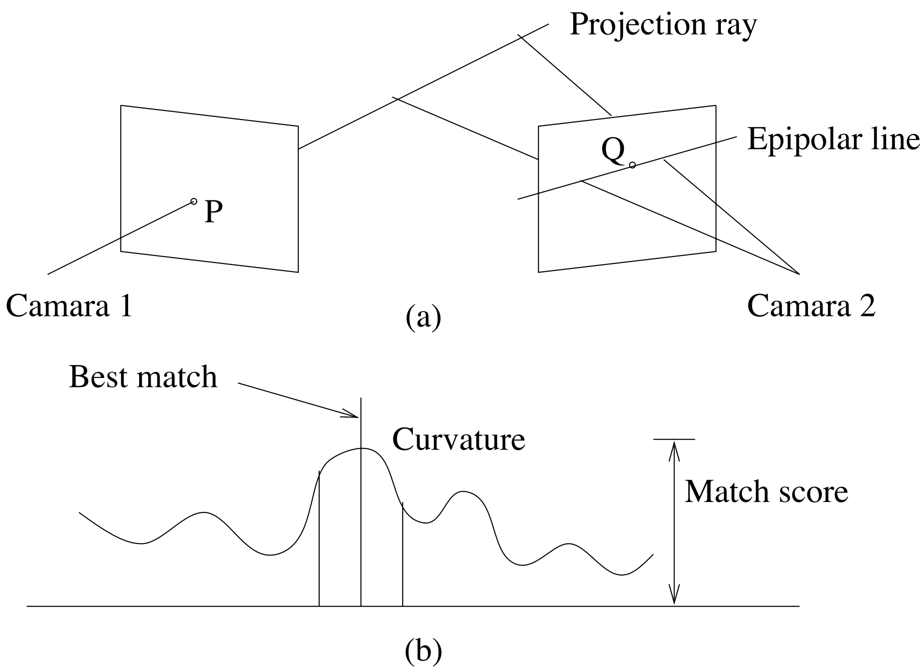


Figure 2: Concepts of epipolar geometry and similarity function

(a) epipolar geometry of an image pair

(b) similarity function of a 2-D point in the left image

two cameras, the images I^F and I^G are first resampled into epipolar coordinates I_E^F and I_E^G, which satisfy the relationship

$$I_E^F(i,j) \cong I_E^G(i + D(i,j), j), \tag{1}$$

where D is the *disparity matrix*. Under the assumption of a nearly Lambertian surface, a patch on the surface has the same apparent brightness pattern when seen from different views. The similarity function $\rho(i,j;d)$ is computed between a window centered at (i,j) in I_E^F and a series of windows centered at $(i+d,j)$ in I_E^G, where $d \in (d_{\min}, d_{\max})$ is incremented in sub-pixel steps. Suppose the size of the windows is $(2m+1) \times (2n+1)$. Using the weighted central moments

$$\mu(i,j) = \frac{1}{N} \sum_{\xi=-m}^{m} \sum_{\eta=-n}^{n} I(i+\xi, j+\eta) A(\xi,\eta), \tag{2}$$

and

$$\sigma(i,j) = \sqrt{\frac{1}{N}\sum_{\xi=-m}^{m}\sum_{\eta=-n}^{n}[I(i+\xi,j+\eta)-\mu(i,j)]^2A^2(\xi,\eta)}, \tag{3}$$

the similarity function is defined as

$$\rho(i,j;d) = \frac{1}{N}\sum_{\xi=-m}^{m}\sum_{\eta=-n}^{n}\frac{I_E^F(i+\xi,j+\eta)I_E^G(i+d+\xi,j+\eta)A^2(\xi,\eta)-\mu_E^F(i,j)\mu_E^G(i+d,j)}{\sigma_E^F(i,j)\sigma_E^G(i+d,j)}, \tag{4}$$

where $N=(2m+1)(2n+1)$ is the number of pixels in the window, and A is the weight matrix. For the experiments presented, a center-weighted Gaussian function was used for A, so that the 3-D features can be determined more reliably under perspective distortion:

$$A(\xi,\eta) = \frac{C}{2\pi\sigma_A^2}\exp\left(-\frac{\xi^2+\eta^2}{2\sigma_A^2}\right), \tag{5}$$

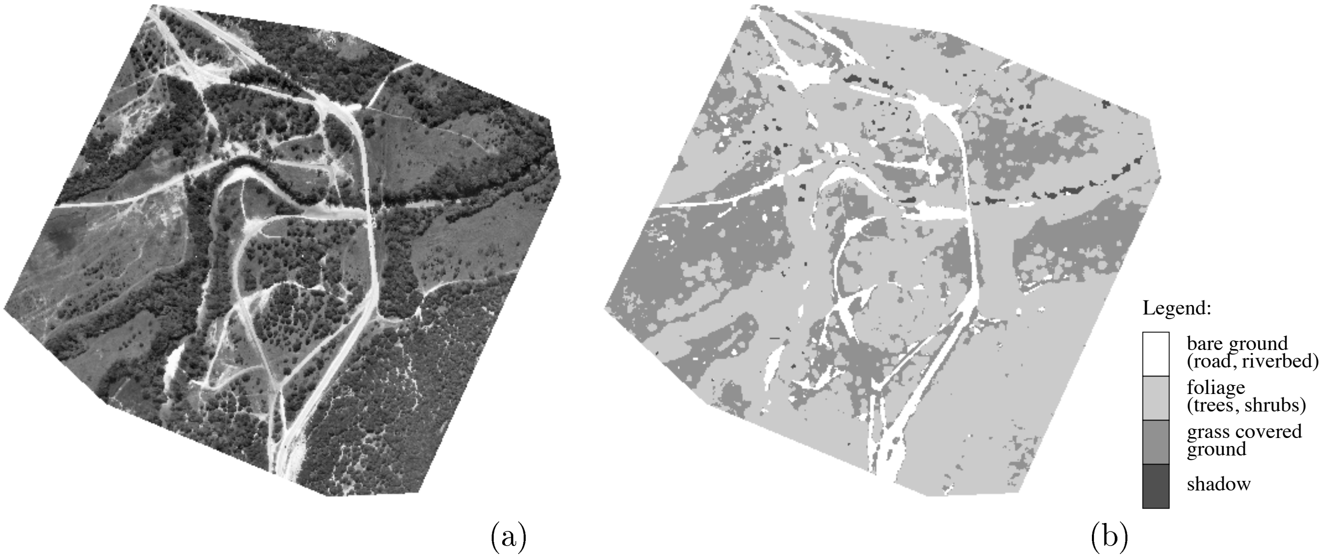
in which C is adjusted such that the average weight in the window is 1.

At a particular pixel (i,j) in Image I_E^F, the similarity function $\rho(i,j;d)$ defined in (4) is a 1-D function as depicted in Fig. 2(b), and the search range is a 1-D domain $(d_{\min}, d_{\max})$ on the epipolar line. However, when camera parameters are not accurate, the epipolar line cannot be precisely determined. In that case, the similarity function must be extended to a 2-D function $\rho(i,j;d_x,d_y)$, and the search region in Image I_E^G can be extended to a 2-D rectangular area with upper-left and bottom-right corners at $(d_{x\min}, d_{y\min})$ and $(d_{x\max}, d_{y\max})$, respectively. In the rest of this paper, we always assume that accurate camera parameters are given and the similarity function is one-dimensional.

A smooth function $\tilde{\rho}$ is fit to the discrete similarity function $\rho(i,j;d)$. At position $\hat{d}$, where

$$\tilde{\rho}(i,j;\hat{d}) = \max\{\tilde{\rho}(i,j;d)\}, \tag{6}$$

the window centered at $I_E^F(i,j)$ finds its best match in I_E^G. Thus, $\hat{d}$ is taken as the estimate for the disparity between $I_E^F(i,j)$ and the corresponding pixel in I_E^G. All disparity estimates $\hat{d}$ are stored in the disparity matrix D. The algorithm utilizes an iterative low-to-high resolution method to improve stability. The matrix D is used as an estimate at the beginning of each iteration. It is refined by warping I_E^G according to the estimate in D and computing ρ again. D is initialized with an estimate of the elevation map, if available, or a precomputed mean value. Using the final disparity matrix D, an orthographic version of the intensity image, as shown in Fig. 3(a), can be generated.



(a)

(b)

Figure 3: Orthographic intensity image and its classification

(a) the 2k×2k ortho-image from Ft. Hood Image Set

(b) classification using a combination of co-occurrence, 3-D, and intensity features (classifier trained by the four small image chips in Fig. 6)

2.2 Generating 3-D features

The 3-D features are generated from the smoothed similarity function. Intuitively, the shape of the similarity function at the best match provides us with information of the terrain texture. Fig. 4 shows some shapes of similarity functions indicating different terrain types.

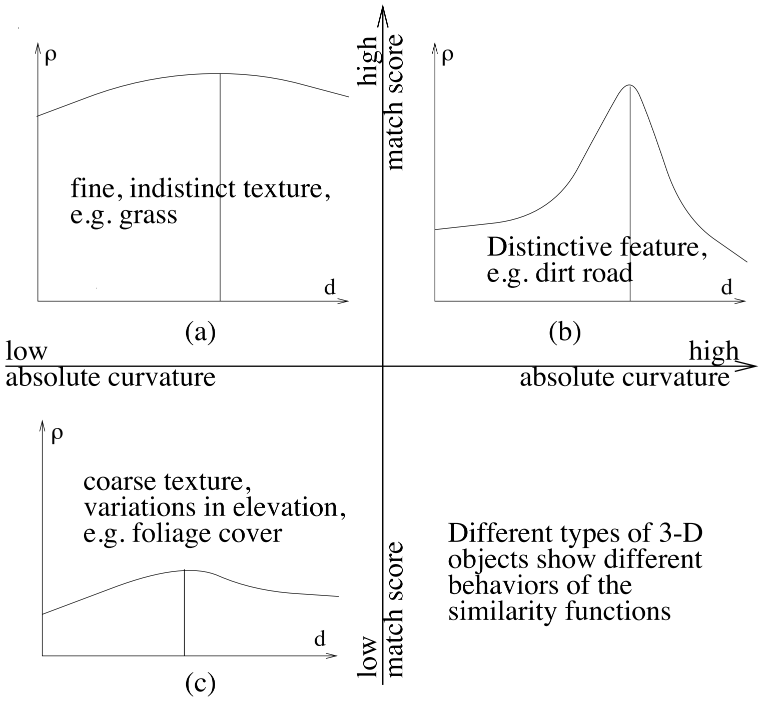


Figure 4: Qualitative types of similarity functions

The maximum value of similarity function at the best match is set to be the *match score* (MS), the first proposed 3-D feature. That is,

$$\mathrm{MS}(i,j) = \tilde{\rho}(i,j;\hat{d}). \tag{7}$$

MS tends to be high when the image patches being correlated are very similar. Taking into account perspective distortion, the highest degree of similarity can be achieved when the image patches are smooth and there is no occlusion due to change of viewpoint between the two images. In case of occlusion, the general shape of the object affected will be recovered at a lower resolution, especially when the texture in the occluded area is similar to texture elsewhere on the object. However, at a higher resolution no good match will be possible and MS will be low in the occluded area. In the case of flat ground, sufficient texture usually exists and changes in viewpoint are unlikely to cause occlusion as the 3-D surface structures are very small. Hence, MS naturally tends to be high (Fig. 4(a)(b)). In the case of foliage areas, changes in viewpoint will often cause occlusion. In these areas MS tends to be low (Fig. 4(c)).

Once the MS feature is obtained for each pixel in Image I_E^F, the second feature, called the *neighborhood variance of match score* (NVMS), can be computed. It is a feature that measures the local variance of match scores, and is computed as the second central moment of MS in a local window with size $(2m'+1) \times (2n'+1)$ in the MS map:

$$\mathrm{NVMS}(i,j) = \sqrt{\frac{1}{N'} \sum_{\xi=-m'}^{m'} \sum_{\eta=-n'}^{n'} [\mathrm{MS}(i+\xi, j+\eta) - E(\mathrm{MS}(i,j))]^2}, \tag{8}$$

in which $N' = (2m'+1)(2n'+1)$ is the number of pixels in the window and $E(\mathrm{MS}(i,j))$ is the average value of MS in the window. Clearly, NVMS is affected by the 3-D structure of the area under the window. When occlusion occurs, the value of MS is decreased; however, the decrease of MS is not homogeneous in a region when occlusion occurs in a random style, such as that in foliage. NVMS measures how MS varies in the area. For flat surfaces NVMS tends to be low; complicated 3-D structures usually have a relatively high NVMS.

The third 3-D feature, *curvature of similarity function* (CSF), describes the distinctiveness of the match between the window patches from the two views. It is computed by fitting a parabola of the form $ad^2 + bd + c$ to the smoothed similarity function $\tilde{\rho}(i,j;d)$ over the range $[D(i,j) - o, D(i,j) + o]$, where o is a parameter that determines a range of statistical significance. In the experiments o was set to a value of 1.5 pixels. The curvature of the parabola, defined to be

$$\mathrm{CSF}(i,j) = 2a, \tag{9}$$

provides an estimate of the curvature at the peak of the similarity function. Typically CSF is a negative value. The absolute value of CSF tends to be high when distinctive features or structures exist in the texture, i.e. there is a unique match (Fig. 4(b)). Texture with complicated 3-D structures such as forest, or texture with little distinctive features such as plain ground, usually have lower values of |CSF| (Fig. 4(a)(c)).

Recall that $\rho(x,y;d)$ is defined over a search range $(d_{\min}, d_{\max})$. Ideally $\tilde{\rho}(x,y;d)$ has a well-defined local maximum in the search range. If that is not the case then MS and CSF are not well-defined. To arrive at a value for MS, it is set to the value of its global maximum over the range. The undefined CSF's are fixed by applying a median filter to the CSF map. A missing local maximum generally indicates a bad match between $I_E^F(i,j)$ and its counterpart in I_E^G. We define the *neighborhood density of well-defined curvature* (NDC) as the fourth 3-D feature. It is computed as the ratio between the number of well-defined CSF in a local window of size $(2m''+1) \times (2n''+1)$ in the CSF map and the total number of pixels in the window, that is,

$$\mathrm{NDC}(i,j) = \frac{1}{N''} \sum_{\xi=-m''}^{m''} \sum_{\eta=-n''}^{n''} \delta[CSF(i+\xi, j+\eta)], \tag{10}$$

where $N'' = (2m''+1)(2n''+1)$ and

$$\delta[CSF(i,j)] = \begin{cases} 1, & \text{if CSF is well-defined at } (i,j) \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

NDC contains information about the reliability of the match for an object. If NDC is low in an area, then generally the area contained less distinguishable texture, or significant occlusion is present.

We illustrate the four types of newly defined 3-D features in Fig. 5, using a portion of the ortho-image shown in Fig. 3(a) (its position shown in Fig. 6). It contains various types of terrain textures. Features NVMS and NDC are computed in a local window of size 17×17. It is visually apparent that the various types of texture are distinguished by the features.

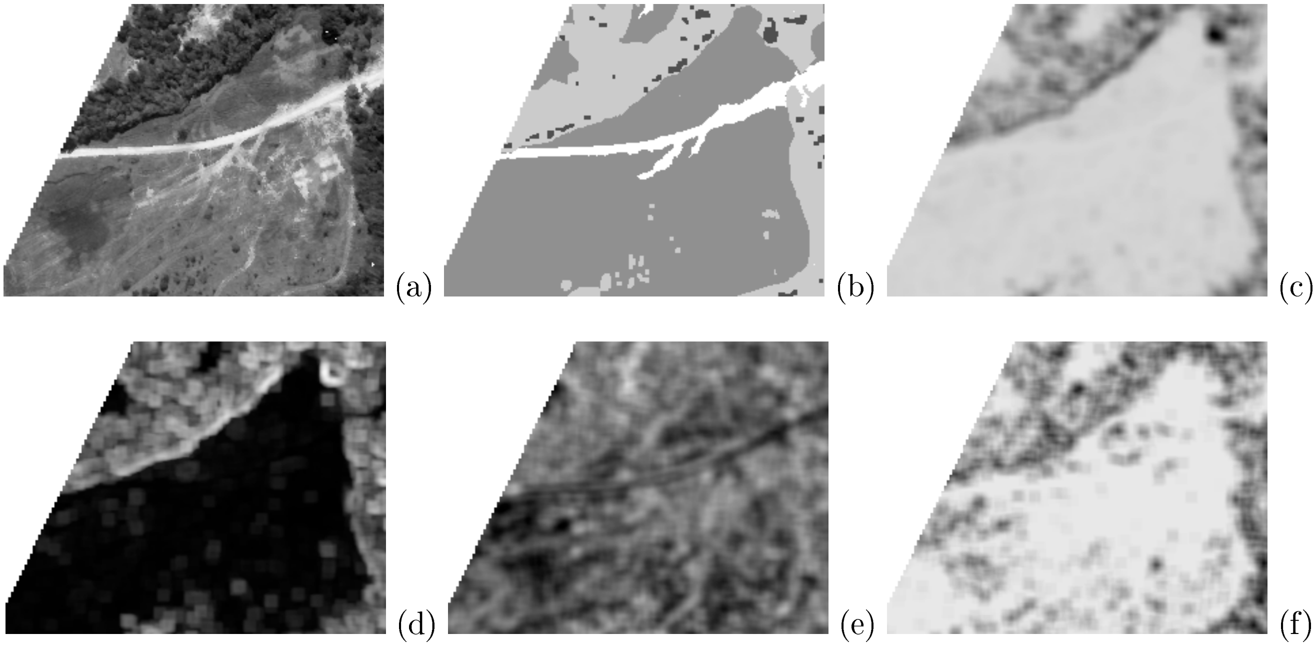


Figure 5: A portion of the 2k×2k ortho-image with a set of 3-D feature maps

(a) the intensity ortho-image

(b) ground truth hand classification

(c) match score (MS)

(d) neighborhood variation of match score (NVMS)

(e) correlation curvature (CSF)

(f) neighborhood density of well-defined curvature (NDC)

3 Experimental Set-up

3.1 Classification algorithm

As an initial attempt to test the capability of newly introduced 3-D features, we employ a linear discriminant method (the Foley-Sammon transform (FST) [12]) for classification in our system. FST is considered a very effective algorithm in terms of linear discriminant ability and has attracted significant attention in the field of pattern recognition. Tian [13] applied FST to image classification problems. Liu et al. [14] extended the FST to a wider application domain and showed experimental results on human face classification. In the field of texture classification, Weszka [16] used this linear discriminant approach for comparison of texture features.

FST provides an efficient way of reducing the dimensionality of feature vectors. As mentioned in Section 1, in terrain classification tasks, a huge number of (2-D) textural features has been studied. Thus, reducing the dimensionality of the feature space is an important consideration in pattern recognition. Using FST, the original features are projected into a lower dimensional *algebraic feature* space, which returns optimal discriminability for the classes.

Suppose there are K classes of objects: $\mathbf{C}_i, i = 1, 2, ..., K$. Let $X_j^{(i)}$ denote the jth training feature vector that belongs to $\mathbf{C}_i$, where $i = 1, 2, ..., K$ and $j = 1, 2, ..., M_i$, M_i the number of training feature vectors in $\mathbf{C}_i$. The centroid of $\mathbf{C}_i$, $\bar{X}^{(i)}$, and the centroid of the K classes, $\bar{X}$, are determined by

$$\bar{X}^{(i)} = \frac{1}{M_i} \sum_{j=1}^{M_i} X_j^{(i)}, \tag{12}$$

$$\bar{X} = \sum_{i=1}^{K} P_i \bar{X}^{(i)}, \tag{13}$$

where $P_i(i = 1, 2, ..., K)$ is a priori probability of $\mathbf{C}_i$. We define the *within-class scatter matrix*, S_w, and the *between-class scatter matrix*, S_b, by

$$S_w = \sum_{i=1}^{K} P_i \frac{1}{M_i} \sum_{k=1}^{M_i} (X_k^{(i)} - \bar{X}^{(i)})(X_k^{(i)} - \bar{X}^{(i)})^T, \tag{14}$$

$$S_b = \sum_{i=1}^{K} P_i (\bar{X}^{(i)} - \bar{X})(\bar{X}^{(i)} - \bar{X})^T. \tag{15}$$

FST uses Fisher criterion to minimize

$$J_f(x) = \frac{x^T S_b x}{x^T S_w x}. \tag{16}$$

FST is optimal in that it obtains the minimum within-class scatter and the maximum between-class scatter in the dimensionality-reduced algebraic feature space. Foley-Sammon optimal set of discriminant vectors is the set of orthonormal vectors that minimize (16). The algorithm of solving for such a set can be found in [14]. Let

$$U = [u_1, u_2, ..., u_r], \tag{17}$$

in which $u_1, u_2, ..., u_r$ are the optimal discriminant vectors. Then, for a vector X in the physical feature space, its algebraic feature vector $T(X)$ is defined by the following transformation:

$$T(X) = U^T X. \tag{18}$$

Using FST, the training vectors can be projected to the algebraic feature space. The dimension of the algebraic feature space is r. After some preliminary tests, we used a fixed number $r = 3$ as the dimensionality of the algebraic feature space in all the experiments described in Section 4 and 5.

We denote the classes in the projected space as $T(\mathbf{C}_i), i = 1, 2, ..., K$. The classifier in the current system uses a minimum distance criterion in the algebraic feature space. That is,

$$X \in \mathbf{C}_i, \text{ if } D_i(T(X)) = \min_{1 \leq k \leq K} \{D_k(T(X))\}, \tag{19}$$

in which $D_i(Y)(i = 1, 2, ..., K)$ is the distance of algebraic feature Y to the centroid of $T(\mathbf{C}_i)$.

We have further modified the FST to account for the fact that some classes have a significantly greater variance in the feature space than others. Instead of using the Euclidean distance as done by other researchers [14]'s, we used Mahalanobis distance in the classifier.

$$D_i(Y) = (Y - \bar{Y}^{(i)})^T R_i^{-1} (Y - \bar{Y}^{(i)}), \tag{20}$$

where $\bar{Y}^{(i)}$ and R_i are the centroid and scatter matrix of $T(\mathbf{C}_i)$, respectively, and can be calculated from the projected algebraic training vectors. The advantage of Mahalanobis distance in classification is that it takes into account the within-class scatter to deal with classes that have different variance.

3.2 2-D features in comparison

The experiments are designed for comparing the proposed 3-D features with traditional 2-D features in their performance in terrain classification tasks. Based on previous studies by many researchers [15, 16, 17, 18, 19], we chose *co-occurrence features* as the comparison 2-D feature set.

Co-occurrence features were introduced by Haralick et al. [15]. Weszka et al. [16] experimentally compared features on terrain images and showed that co-occurrence features were better than Fourier power spectrum features. Conners and Harlow [17] reported in their theoretical study of 2-D textural features that co-occurrence features were the most effective in comparison with the gray-level run-length, the gray-level difference, and the power spectrum features. du Buf et al. [18] compared seven types of features for image segmentation and found that co-occurrence features were the most significant in applications to aerial and satellite images. Ohanian and Dubes [19] studied four types of features, including Markov Random Field parameters, multi-channel filtering features, fractal based features, and co-occurrence features, and concluded that co-occurrence features perform best.

Co-occurrence features describe the texture within a local image window based on gray-tone spatial dependencies. A *co-occurrence matrix* is defined to represent the gray-tone spatial dependence frequencies in the window. The (i, j) matrix entry $P_{ij\theta}$ represents the relative frequency for which two pixels with gray-tone i and j are separated by distance d at angle θ in the window. Using the co-occurrence matrix, a large number of features can be defined. Haralick et al. [15] in their original work provided 14 types of features. Previous studies [16, 17, 19] suggest the use of ASM (the angular second-moment), CON (the contrast), and ENT (the entropy), which are defined as follows.

$$f_\theta^{\mathrm{ASM}} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P_{ij\theta}^2, \tag{21}$$

$$f_\theta^{\mathrm{CON}} = \sum_{n=0}^{G-1} n^2 \sum_{|i-j|=n} P_{ij\theta}, \tag{22}$$

$$f_\theta^{\mathrm{ENT}} = - \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P_{ij\theta} \ln P_{ij\theta}. \tag{23}$$

The ASM feature is a measure of homogeneity of the texture intensity. The CON feature measures the local variation present in the texture. The ENT feature measures the complexity. In the pre-processing stage the image is performed with histogram equalization to reduce the number of gray levels to G. Hence, the co-occurrence matrix has size $G \times G$. As suggested by other researchers [19], and by our preliminary experimental results, we applied the following settings in our experiments: image window size 17×17, distance $d = 3$, four angles $\theta = 0°, 45°, 90°, 135°$, and the number of gray levels $G = 8$. Under these settings, a total of twelve features were employed (i.e. three types of features, each with four angles).

3.3 Training data

The experiments have been carried out on a set of aerial images of Ft. Hood Image Set. Fig. 3(a) shows a 2k×2k image from an orthographic projection. We use a pixel classification scheme to give each pixel that belongs to different object class a distinct label. For this image set, four object classes are considered: foliage (trees, shrubs), grass covered ground, bare ground (road, riverbed), and shadow. The 2k×2k ortho-image has been independently hand-labeled into the four classes by an image analyst to act as ground truth data in the experiments. A portion of the ground truth segmentation is shown in Fig. 5(b) (legend shown in Fig. 3).

The selection of training data is an important issue for classification systems. Small training data sets are often preferred since they involve less man-machine interaction. However, more training data are typically needed to improve reliability. Reliable classifiers are those that are stable to various training data. Hence, all other issues aside, when a reliable classifier is used, it only needs a small amount of training data. In the first part of our experiments (Section 4), we attempt to use a very small training data set to design classifiers and observe their performance. The training data set includes four small image chips, each containing a texture of a particular class, randomly sampled from the ortho-image. The sizes of the four image chips are: 99×99 (foliage), 75×75 (grass covered ground), 37×37 (bare ground), and 11×11 (shadow). They account for only about 1% of the pixels in the entire ortho-image. Fig. 6 depicts the sizes and positions of image chips in the ortho-image. In the second part of the experiments (Section 5), we increase the number of image chips to test the reliability of the classifiers under different training data sets.

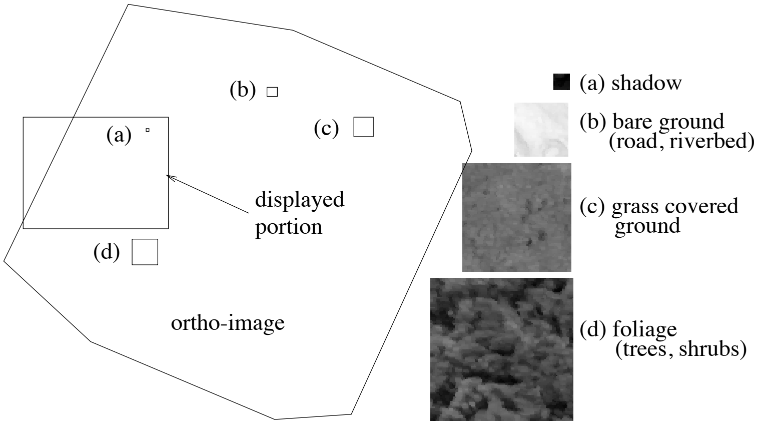


Figure 6: Positions of the training image chips and the displayed portion in Fig. 5 and 7

3.4 Feature sets for comparison

The features being used in the experiments consist of the twelve 2-D co-occurrence features, four 3-D features (MS, CSF, NVMS, NDC) generated in Section 2, and the original intensity as an additional feature. For the purpose of comparison, four classifiers are designed, each using a different set of features, denoted as *Feature Set A, B, C,* and *D.* Feature Set A contains the twelve co-occurrence features only. Feature Set B adds the intensity feature into Set A. Feature Set C consists of the four 3-D features plus the intensity feature. Feature Set D includes all the features – the twelve co-occurrence features, the four 3-D features and the intensity. All four classifiers employ the same classification algorithm based on Foley-Sammon transform and minimum Mahalanobis distance criterion. The performance of the classifiers using the four feature sets is discussed next.

4 Fundamental Experiments

In the fundamental experiments, we test the performance of the classifiers in terms of classification accuracy. We compare the four classifiers using Feature Set A, B, C, and D. As mentioned in Section 3.3, the same small training data set shown in Fig. 6 was used in all the classifiers. The entire ortho-image was then classified based on the training of the four image chips.

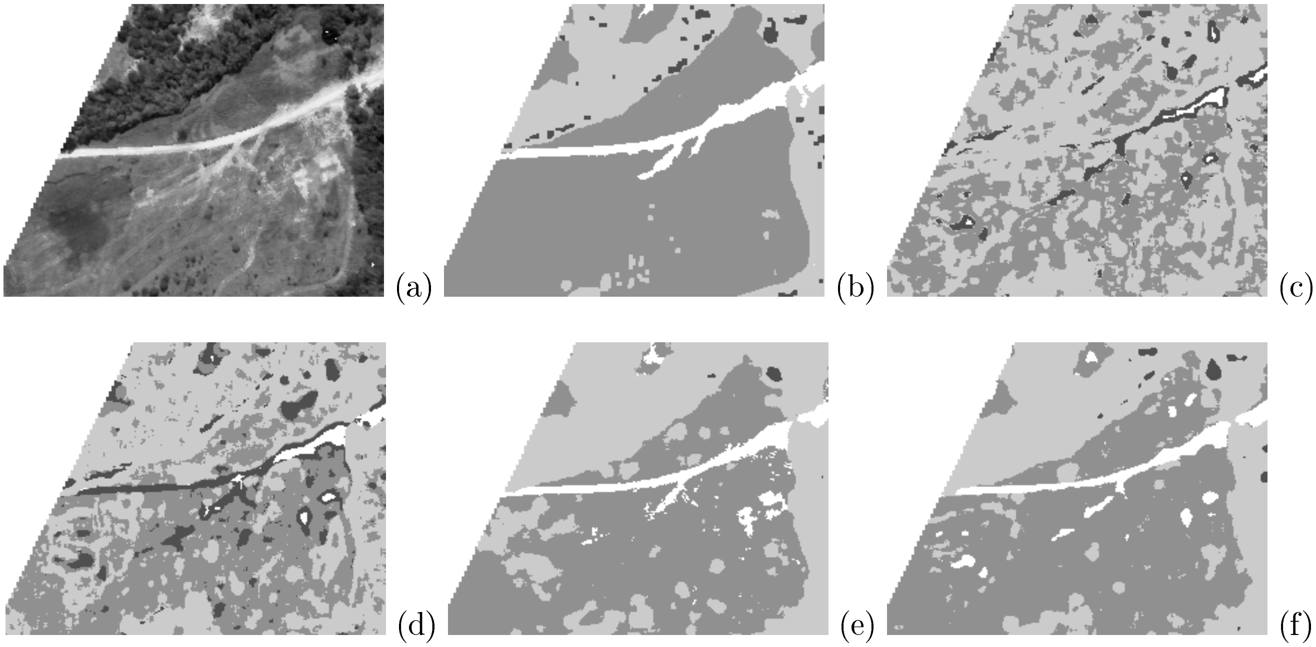


Figure 7: Classification results using various feature sets (see Fig. 3 for legend)

(a) the intensity ortho-image

(b) ground truth hand classification

(c) Feature Set A: twelve co-occurrence features

(d) Feature Set B: twelve co-occurrence features and one intensity feature

(e) Feature Set C: four 3-D features and one intensity feature

(f) Feature Set D: twelve co-occurrence features, four 3-D features, and one intensity feature

4.1 Qualitative analysis

Fig. 7 shows the classification results on the image portion of Fig. 5(a). The first observation for these results is that the classifier using only the 2-D co-occurrence features (Fig. 7(c)) produces the worst classification result. In particular, the ground and foliage labels are misclassified for each other in many places. An explanation of this phenomenon is that the ground category is affected by the existence of grass and vehicle tracks intermingled with ground patches so that it has a mottled 2-D textural appearance. Thus, the 2-D co-occurrence features find the ground regions in some cases with rough texture producing a variation of the 2-D intensity to be similar to foliage.

In contrast, since the presence of grass and vehicle tracks does not cause the 3-D smoothness of ground surface to change noticeably, they have little impact on the values of the 3-D features. For example, two image patches with the same vehicle tracks will still have a high match score because the variations are distinctive (i.e. somewhat unique) as opposed to patches of foliage texture which are far less distinct. Therefore, classifiers using 3-D features have a greater ability to discriminate the smooth ground surface from the rough forest surface. This is clearly shown in Fig. 7(e)(f).

Intensity, as a feature, provides some discriminability among the classes, especially the road and shadow classes which have extreme intensities. Thus, there is some improvement when the intensity feature is added to the co-occurrence features (Fig. 7(d)). However, because ground cover may have variations in brightness due to some variations in surface slope and soil/grass type, the intensity feature is not fully reliable. For example, in the left part of Fig. 7(a), the ground regions below the road turn darker, and classification deteriorates in these areas when the intensity feature is added into the co-occurrence feature set (Fig. 7(d)). However, once again these regions receive satisfactory classification when 3-D features are added (Fig. 7(f)), since the darker ground remains smooth (i.e. has little change in 3-D variation), allowing 3-D features to perform effectively.

The classification result on the entire 2k×2k ortho-image using Feature Set D is shown in Fig. 3(b).

4.2 Quantitative analysis

A quantitative analysis of the classification accuracy is shown in Table 1. The statistics of all the four experiments using different feature sets are given in the table. For each experiment, the entry (i, j) of the table shows the number of pixels that belong to Class i while being classified into Class j. The elements on the diagonals are the numbers of pixels that are correctly classified. The overall accuracy of each classification experiment is also given in the table (obtained by summing the diagonal elements and dividing by the total number of testing pixels).

It can be seen from Table 1 that Feature Set A (twelve co-occurrence features only) provides the lowest classification accuracy. The additional intensity feature in Feature Set B produces a minor improvement. The use of the four 3-D features and intensity of Feature Set C improves the classification dramatically over Set A – about 20 percentage points. Feature Set D of all features only shows marginal improvement over Set C. In other words, if the four 3-D features and the intensity are used, then the participation of the 2-D co-occurrence features will hardly improve the classification accuracy.

A further observation from the experimental results using Feature Set C and D in Table 1 is that major false classifications fall into two cases: (1) ground truth shadow pixels being classified into foliage, or (2) ground truth grass pixels being classified into foliage. Case (1) is easy to understand. Shadow is usually an inherent component of foliage texture. Therefore, shadow in small pieces tends to be labeled as foliage by the classifier. In Fig. 3 we can see that large and continuously shadowed areas (e.g. in the upper right part) have been correctly classified. The mis-classification in case (2) is related to the 3-D structures of trees. In Fig. 1(a) we have seen that the crown of a tree is stretched out if seen from an oblique view. If a tree is observed from two different viewpoints, the projected crown will stretch out in two different directions, and will occlude different sections of ground. Suppose that a piece of ground shows up in one source image, but is occluded by the stretch-out of tree crowns in the other. The 3-D features of this piece of ground is most likely to be close to that of foliage (e.g. MS is low). The overall effect is that the tree seems to cover more ground area than it really does, and ground pixels near a foliage area tend to be mis-classified. This problem became even worse in Ft. Hood images used in the experiments, due to the existence of a lot of sparse trees. In Fig. 7(e)(f) we can see the dilations of the sparse tree areas. There are several ways to solve this problem. One of them is to use morphological algorithms to erode the classified foliage areas, or areas of complicated 3-D structural textures. If some a priori knowledge, such as height of trees, can be obtained (e.g. via context-sensitive analysis of nearby tree regions), this problem could be solved in a knowledge-based manner. These are topics subject to future studies.

5 Additional Experiments

With the motivation of testing the reliability of the 3-D features to various training data, additional experiments were carried out. All the experimental settings were the same as in the fundamental experiments, except that different training data sets were used. For each texture class, we included two more image chips (for a total of three) as training data. They were randomly sampled from the ortho-image (Fig. 3(a)), with sizes similar to those in Fig. 6.

Twelve combinations (subsets of training chips) were tested in the experiments, and the results are shown in Table 2. Each Training Sets 1, 2, and 3 contained all the training chips for the bare ground, grass covered ground, and shadow classes, but only one (different) chip for the class of foliage. Each Training Sets 4, 5, and 6 contained one chip for the bare ground class and all the chips for other classes. Similarly, Training Sets 7, 8, and 9 each contained one chip for grass covered ground and all the chips for the others, and Training Set 10, 11, and 12 each had one for shadow and all for the others.

With each training data set, four experiments were performed to test the classification accuracy for the four feature sets, A, B, C, and D. From Table 2 we can see that, under every training data combination, the 3-D features plus the intensity consistently performed better than Feature Set A and B, where no 3-D feature was involved. The best result was always given by Feature Set D, which uses all the features. On average, with the intensity and 3-D features only (Feature Set C), the classifier outperformed Feature Set B (2-D co-occurrence features plus the intensity) by nearly 10 percentage points. Adding the four 3-D features into Feature Set B (to form Feature Set D) improves the classification accuracy by 10-15 percentage points.

From the standard deviation we can see that the set with co-occurrence features plus the intensity was most sensitive to different training sets. Hence the quality of its classification was most unreliable. The sets with 3-D features had more reliable results. This phenomenon indicates that the 3-D features generated using the algorithms in this paper really reflect some consistent physical characteristics of textures with 3-D structures.

6 Conclusions and Discussions

In this paper we have provided a new perspective for understanding the nature of texture. Based on utilizing features for 2-D image texture and 3-D world texture, we extend our analysis to 3-D structural patterns on object surfaces in the real world. Methodologically, we have proposed a set of 3-D features that takes multiple views of 3-D objects into account for terrain texture classification in aerial images. Experimental results have shown that the set of new 3-D textural features significantly outperform the set of co-occurrence features, which is one of the best traditional 2-D feature sets. It is also shown that the 3-D features are reliable to various training data, suggesting that they can be used in the circumstance that only a small amount of training data is available.

This new approach moves texture analysis beyond purely low-level image processing. Traditional 2-D texture analysis methodologies could benefit from incorporating 3-D features as suggested here. However, we believe that some traditional tools have to be revisited. For example, image mosaicking and texture synthesis [6, 5, 8, 17, 18] are not readily applicable to the new domain.

The motivation of this paper is to reveal the performance of the proposed 3-D feature set in comparison with the traditional 2-D features. To focus, we only pay attention to the raw results of pixel classification, and leave many issues that a real system must face as future topics, such as efficient use of the features and improvement of classification accuracy. These issues include designing sophisticated classifiers to deal with non-linear separability of feature space. Draper et al. [11]'s, and Jain and Karu [8]'s recent work on learning schemes are good candidates in this regard. Other image segmentation techniques will be applied to cope with small classification fragments (such as using a filter as a post-processing module to "smooth" out small fragments [5, 8]) and biased segmentation boundaries mentioned in Section 4.

We are also investigating expansion of the number of object classes and the effect of this on classification accuracy. More terrain types (such as rocky ground and water surface) will be classified by using the 3-D features or a combination of 3-D and 2-D features. As an example, the correlation curvature feature (CSF) has shown sensitivity to small objects or edges of regular large-scale structures, suggesting that a new class of this kind can be formed to support other computer vision goals.

References

[1] H. Tamura, S. Mori, and T. Yamawaki, "Textural Features Corresponding to Visual Perception," *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 8, pp. 460-473, 1978.

[2] R. M. Haralick, "Statistical and Structural Approaches to Texture," *Pro. IEEE*, Vol. 67, No. 5, pp. 786-804, May 1979.

[3] L. Van Gool, P. Dewaele, and A. Oosterlinck, "Texture Analysis Anno 1983," *Computer Vision, Graphics, and Image Processing*, Vol. 29, pp. 336-357, 1985.

[4] M. Tuceryan and A. K. Jain, "Texture Analysis," In *The Handbook of Pattern Recognition and Computer Vision*, C. H. Chen, L. F. Pau, and P. S. P. Wang, eds. World Scientific Publishing Co., pp. 235-276, 1993.

[5] J. Strand and T. Taxt, "Local Frequency Features for Texture Classification," *Pattern Recognition*, Vol. 27, No. 10, pp. 1397-1406, 1994.

[6] D. Dunn, W. E. Higgins, and J. Wakeley, "Texture Segmentation Using 2-D Gabor Elementary Functions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 2, pp. 130-149, 1994.

[7] B. B. Chaudhuri and N. Sarkar, "Texture Segmentation Using Fractal Dimension," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 17, No. 1, pp. 72-77, 1995.

[8] A. K. Jain and K. Karu, "Learning Texture Discrimination Masks," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 18, pp. 195-205, 1996.

[9] X. Wang, F. Stolle, H. Schultz, E. Riseman, and A. Hanson, Using Three-Dimensional Features to Improve Terrain Classification, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* Puerto Rico, pp. 915-920, June 1997.

[10] H. Schultz, "Terrain reconstruction from widely separated images," *Integrating Photogrammetric Techniques with Scene Analysis and Machine Vision II*, SPIE Proceedings Vol. 2486, pp. 113-123, Orlando, FL, April 1995.

[11] B. A. Draper, C. E. Brodley, and P. E. Utgoff, "Gold-Directed Classification using Linear Machine Decision Trees," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 9, pp. 888-893, 1994.

[12] D. H. Foley and J. W. Sammon, Jr., "An Optimal Set of Discriminant Vectors," *IEEE Trans. on Computers*, Vol. 24, No. 3, pp. 281-289, 1975.

[13] Q. Tian, M. Barbero, Z.-H. Gu, and S. H. Lee, "Image Classification by the Foley-Sammon Transform," *Optical Engineering*, Vol. 25, No. 7, pp. 834-840, 1986.

[14] K. Liu, Y.-Q. Cheng, and J.-Y. Yang, "Algebraic Feature Extraction for Image Recognition Based on an Optimal Discriminant Criterion," *Pattern Recognition*, Vol. 26, No. 6, pp. 903-911, 1993.

[15] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 3, No. 6, pp. 610-621, 1973.

[16] J. S. Weszka, C. R. Dyer, and A. Rosenfeld, "A Comparative Study of Texture Measures for Terrain Classification," *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 6, No. 4, pp. 269-285, 1976.

[17] R. W. Conners and C. A. Harlow, "A Theoretical Comparison of Texture Algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 2, No. 3, pp. 204-222, 1980.

[18] J. M. H. du Buf, M. Kardan, and M. Spann, "Texture Feature Performance for Image Segmentation," *Pattern Recognition*, Vol. 23, No. 3/4, pp. 291-309, 1990.

[19] P. P. Ohanian and R. C. Dubes, "Performance Evaluation for Four Classes of Textural Features," *Pattern Recognition*, Vol. 25, No. 8, pp. 819-833, 1992.

Table 1: Contingency analysis of classification results in the fundamental experiments (unit: 1000 pixels)

		classification using Feature Set A			
			grs. cvd.		bare
ground truth	(total)	shadow	ground	foliage	ground
shadow	(41.8)	21.2	7.1	11.9	1.7
grass covered ground	(683.0)	17.7	259.9	405.2	0.3
foliage	(1018.6)	10.8	121.2	886.6	0.0
bare ground	(193.4)	77.5	19.3	50.5	46.1
total	(1936.8)	127.2	407.5	1354.1	48.1
		correctly classified pixel total: 1213.8 (62.67%)			

		classification using Feature Set B			
			grs. cvd.		bare
ground truth	(total)	shadow	ground	foliage	ground
shadow	(41.8)	23.7	7.0	11.1	0.0
grass covered ground	(683.0)	40.8	332.6	308.1	1.6
foliage	(1018.6)	11.7	55.0	950.7	1.2
bare ground	(193.4)	52.2	12.9	31.2	97.2
total	(1936.8)	128.2	407.5	1301.1	100.0
		correctly classified pixel total: 1404.1 (72.50%)			

		classification using Feature Set C			
			grs. cvd.		bare
ground truth	(total)	shadow	ground	foliage	ground
shadow	(41.8)	3.0	0.0	38.8	0.0
grs. cvd. ground	(683.0)	0.0	439.5	231.4	12.2
foliage	(1018.6)	0.4	20.0	995.3	2.9
bare ground	(193.4)	0.0	17.3	25.1	150.9
total	(1936.8)	3.3	476.8	1290.6	166.0
		correctly classified pixel total: 1588.7 (82.03%)			

		classification using Feature Set D			
			grs. cvd.		bare
ground truth	(total)	shadow	ground	foliage	ground
shadow	(41.8)	13.8	0.0	27.8	0.2
grs. cvd. ground	(683.0)	0.0	468.6	202.7	11.8
foliage	(1018.6)	2.0	18.0	995.7	2.8
bare ground	(193.4)	0.0	33.9	21.4	138.1
total	(1936.8)	15.8	520.6	1247.5	152.9
		correctly classified pixel total: 1616.1 (83.44%)			

Table 2: Classification accuracy in percentage on the 2k×2k ortho-image using 12 different training data sets

	Feature Set A	Feature Set B	Feature Set C	Feature Set D
Training Set 1	63.3	65.6	77.6	83.0
Training Set 2	63.2	66.3	75.8	80.8
Training Set 3	63.6	64.7	78.8	79.5
Training Set 4	63.6	65.4	77.5	80.6
Training Set 5	63.8	65.3	74.3	79.0
Training Set 6	63.4	66.0	75.8	82.4
Training Set 7	63.6	65.5	70.9	79.8
Training Set 8	63.9	74.6	78.0	80.1
Training Set 9	63.1	66.6	73.9	78.6
Training Set 10	63.5	65.7	80.1	83.3
Training Set 11	64.7	76.5	81.1	83.0
Training Set 12	64.2	68.7	81.7	83.1
mean	63.67	67.59	77.12	81.09
standard deviation	0.19	13.81	9.25	2.82

Feature Set A: twelve co-occurrence features

Feature Set B: twelve co-occurrence features and one intensity feature

Feature Set C: four 3-D features and one intensity feature

Feature Set D: twelve co-occurrence features, four 3-D features, and one intensity feature

Training Set 1-12: see Section 5