Statistica Sinica **14**(2004), 413-430

SUPERVISED MOTION SEGMENTATION BY SPATIAL-FREQUENTIAL ANALYSIS AND DYNAMIC SLICED INVERSE REGRESSION

Han-Ming Wu and Henry Horng-Shing Lu

National Chiao-Tung University

Abstract: In this paper, we propose a new method for supervised motion segmentation based on spatial-frequential analysis and dimension reduction techniques. A sequence of images could contain non-ridge motion in the region of interest and the segmentation of these moving objects with deformation is challenging. The aim is to extract feature vectors that capture the spatial-frequential information in the training set and then to monitor the variations of the feature vectors over time in the test set. Given successive images in the training set, we consider a dynamic model that extends the sliced inverse regression in Li (1991). It is designed to capture the intrinsic dimension of feature vectors that holds over a local time scale. These projected features are then used to classify training images and predict forthcoming images in the test set into distinct categories. Theoretic properties are addressed. Simulation studies and clinical studies of a sequence of magnetic resonance images are reported, which confirm the practical feasibility of this new approach.

Key words and phrases: Dimension reduction, Gabor filter bank, motion segmentation, non-ridge motion, sliced inverse regression, spatial-frequential analysis.

1. Introduction

Motion segmentation is a preliminary step for further analysis of dynamic images, like videos, medical images, satellite images, and so forth. It is aimed to partition the images into regions that have different motion characteristics. Non-rigid motion is the motion of objects that change shape (i.e., deform) over time. Segmentation from these temporal sequences of images became difficult and challenging.

There are various approaches proposed for motion segmentation by statistical modelling (e.g., Bouthemy (1989), Nguyen, Worring and Dev (2000) and Vasconcelos and Lippman (2001)) and/or structural modelling (e.g., Duncan, Owen, Staib and Anandan (1991), Isard and Blake (1996), Mikic, Krucinski and Thomas (1998) and Deng and Manjunath (2001)). Recent development on multiple motion segmentation of images is investigated in Mansouri and Konrad (2003). They proposed a global model based on a formulation of active contour

(Blake and Isard (1998)) and solved it by level set techniques (Osher and Paragios (2003)). This method is independent of intensity boundaries and robust to the initial segmentation. Most of the above methods need to specify and estimate the parameters for their models of motion, which are either difficult to determine or computational demanding in estimation procedures. The computational overhead could be overcome by the increasing power of computers. However, as computer power increases, the volume of data and images collected also increases. We still need to find fast and effective methods to mine important features and intrinsic dimension.

Inspired by the recent studies of the early vision system of human perception in V1 cells, image synthesis and segmentation of static images based on spatial-frequential analysis that mimics human vision have been investigated (Malik and Perona (1990), Dunn, Higgins and Wakeley (1994), Tan (1995) and Zhu, Wu and Mumford (1998)). Recent progress has been made for the modelling of textures and dynamic textures by statistical models (Zhu, Liu and Wu (2000), Wu, Zhu and Liu (2000), Wu, Zhu and Guo (2002) and Doretto, Chiuso, Wu and Soatto (2003)). Currently, researchers develop methods of extraction and representation of the feature vectors that are intrinsic in the space and time domain. For example, Fablet, Bouthemy and Perez (2002) characterize the motion information by non-parametric probabilistic modelling. Time-varying texture features are further investigated in Doretto, Chiuso, Wu and Soatto (2003). It remains challenging to segment dynamic images with only a few training images. We aim to extend the studies of spatial-frequential analysis for motion segmentation.

Sliced inverse regression (SIR) was proposed by Li (1991) to find compact representations that explore the intrinsic structure of high-dimensional observations. It has been extended and used in various applications (Chen and Li (1998) and Li (2000)). We have applied this technique with spatial-frequential analysis to segment and diagnosis static images, like ultrasound images (Chen, Lu and Lin (1999), Chen, Lu and Han (2001) and Lu, Chen and Wu (2001)). This study will further investigate the possibility of extending SIR and spatial-frequential analysis for dynamic images.

We first extend the model of SIR for dynamic data and term it dynamic SIR (DSIR). Then, DSIR is combined with spatial-frequential analysis for motion segmentation. Every pixel of an image is regarded as a realization of a stochastic process over space and time. The feature vector for one pixel in one time frame is first analyzed by spatial-frequential analysis of local blocks centered at that pixel. Assume that the relationship between these feature vectors and class labels remains similar between successive frames for neighboring pixels, then we can find out the intrinsic dimension of feature vectors in the training images by DSIR. These projected feature vectors thus provide prediction rules for forthcoming images in the test set. Furthermore, we only need a small number of training images to decide the projection of feature vectors and prediction rules.

The model setup of DSIR and its properties are given in Section 2. Section 3 discusses the feature extraction methods. The algorithm of DSIR and the segmentation procedure are presented in Section 4. Section 5 reports the segmentation results both on the simulated and clinical image sequence of magnetic resonance (MR) images. Comparison studies are also performed to investigate the projection directions by principal component analysis (PCA), independent component analysis (ICA), SIR and DSIR. We then conclude in Section 6. The proof of the theoretic result is in the Appendix 1.

2. Models and Properties

Li (1991) introduced the following model

$$y = f(\beta_1'\mathbf{x}, \ldots, \beta_K'\mathbf{x}, \epsilon), \tag{1}$$

where y is a univariate variable, $\mathbf{x}$ is a random vector with dimension $p \times 1$, $p \geq K$, β's are vectors with dimension $p \times 1$, ϵ is a random variable independent of $\mathbf{x}$ and f is an arbitrary function. The β's are referred to as effective dimension reduction (*e.d.r.*) or projection directions. Sliced inverse regression (SIR) is a method for estimating the *e.d.r.* directions based on y and $\mathbf{x}$. Under regularity conditions, it is shown that the centered inverse regression curve $E[\mathbf{x}|y] - E[\mathbf{x}]$ is contained in the linear subspace spanned by $\beta_k \Sigma_{\mathrm{XX}}$ $(k = 1, \ldots, K)$, where Σ_{XX} denotes the covariance matrix of $\mathbf{x}$. Based on these facts, the estimated β's can be obtained by the procedures of standardizing $\mathbf{x}$, partitioning slices (or groups) according to the value of y, calculating the slice means of $\mathbf{x}$, and performing the principal component analysis of the slice means with weights.

The above model, (1), can be extended for dynamic data as follows:

$$y(t) = f(\beta_1'\mathbf{x}(t), \ldots, \beta_K'\mathbf{x}(t), \epsilon(t)), \tag{2}$$

where $y(t)$ and $\mathbf{x}(t)$ are response variables and p-dimensional covariates observed at time t. The projection directions, β's, are assumed to be invariant over time and ϵ is the noise process. Analogous to the steps in Li (1991, 2000), we set a condition and state a theorem. The proof is in the Appendix.

Condition 1. For any b in R^p, $E[b'\mathbf{x}(t)|\beta_1'\mathbf{x}(t), \ldots, \beta_K'\mathbf{x}(t)]$ is linear in $\beta_1'\mathbf{x}(t)$, $\ldots, \beta_K'\mathbf{x}(t)$ for any t. That is, $E[b'\mathbf{x}(t)|\beta_1'\mathbf{x}(t), \ldots, \beta_K'\mathbf{x}(t)] = c_0 + c_1\beta_1'\mathbf{x}(t) + \cdots + c_K\beta_K'\mathbf{x}(t)$ for some constants $c_0, \ldots, c_K$ and any t.

Theorem 1. *Under* (2) *and Condition* 1, $E[\mathbf{x}(t)|\, y(t)] - E[\mathbf{x}(t)]$ *falls in the linear subspace spanned by* $\mathrm{Cov}\,(\mathbf{x}(t))\beta_k$ *for any* t*,* $k = 1, \ldots, K$*.*

When $\mathbf{x}(t)$ is elliptically symmetric for any time t, the above condition is fulfilled. When this condition is violated, the biases in estimating the projection directions are not large as discussed in Li (1991, 2000). Because (2) does not introduce any structural modelling over time, we can pull together successive data to estimate the projection directions more effectively. Furthermore, all the properties of SIR can be passed along to this dynamic model without any difficulty. The algorithm for estimating projection directions in motion segmentation is specified in Section 4.

Li, Aragon and Thomos-Agan (1995) and Li (2000) investigated the analysis of multivariate outcome data. They considered the following time series model for a univariate response:

$$y(t) = \mu(t) + c_1(\mathbf{x})f_1(t) + \cdots + c_L(\mathbf{x})f_L(t) + \epsilon(t), \tag{3}$$

where $y(t)$ is a response curve, $\mu(t)$ is a baseline curve, the coefficients $c_l(\mathbf{x})$, $l = 1, \ldots, L$, are time-invariant factors that depend only on the covariate $\mathbf{x}$, $\{f_1(t), \ldots, f_L(t)\}$ is a set of functions and $\epsilon(t)$ represents the noise process. Our model (2) differs from (3) in three respects: we allow the covariate process, $\mathbf{x}(t)$ to change in time but assume the projection directions holds for any fixed time; we do not assume the noise is additive; our model allows nonlinearity in f, their model allows nonlinearity in $c_l(\mathbf{x})$ and $f_l(t)$. Therefore, our model is applicable for a sequence of data where the projection directions and the nonlinear relationship hold, which usually works for a short sequence. On the other hand, their model is useful to determine the coefficient functions of covariates and basis functions of time separately from a sequence of data. They have extended the model (3) to multivariate responses and covariates. Similarly, it is also possible to extend the model (2) to multivariate cases.

3. Feature Extraction of Images

We discuss the feature extraction of 2D images here. These methods can be extended to 3D or higher dimension easily.

Space domain: Local blocks. A rectangular lattice for a digital image in 2D of size $N \times M$ is denoted by $\mathcal{S} = \{(I, J) | 1 \leq I \leq N, 1 \leq J \leq M, I, J \in \mathcal{Z}\}$. The spatial characteristic of a pixel in a image is described by its neighboring pixels. So, a local block in one time frame with size $b \times b$ can be formed as a feature vector $\mathbf{x}^{(i)}(t_j)$ of the central pixel $i \in \mathcal{S}$, $i = 1, \ldots, n$ and $n = (M - b + 1)(N - b + 1)$, for each time point t_j, $j = 1, \ldots, m$. The dimension of the feature vector in the space domain is b^2. We do not include the pixels at different time frames into the feature vector because they may vary according to motion. For a fixed pixel i, the collection of feature vectors along the sequence of images, $\{\mathbf{x}^{(i)}(t_j), j = 1, \ldots, m\}$, represents temporal variation, or motion.

Given a sequence of m training images, we have the class labels $\{y^{(i)}(t_j), j = 1, \ldots, m\}$ and the feature vectors $\{\mathbf{x}^{(i)}(t_j), j = 1, \ldots, m\}$. We need to find the projection directions of feature vectors for classification and prediction. If the number of training images, m, is bigger than the dimension of feature vectors, b^2, then it is feasible to estimate the projection direction in the dimension of b^2. However, for a short sequence of training images, m is not necessary bigger than b^2. Therefore, we need to borrow information from neighboring pixels.

Let $\mathcal{N}_q^{(i)}$ be the set of neighboring sites for pixel i with the q-order neighborhood system. For example, in the first-order neighborhood system, every interior site has four neighbors and the size of $\mathcal{N}_1^{(i)}$ is 5. There are eight neighbors for every interior site in the second-order neighborhood system and the size of $\mathcal{N}_2^{(i)}$ is 9. Then, for pixel i, we can collect the neighboring feature vectors as the training set:

$$\mathcal{X}^{(i)} = \{\mathbf{x}_l^{(i)}(t_j), j = 1, \ldots, m, l \in \mathcal{N}_q^{(i)}\}, \tag{4}$$

where $i = 1, \ldots, n$ and $n = (M - b + 1)(N - b + 1)$. For instance, if $q = 2$, then the size of training set is $9m$. This training set will be used to estimate the projection directions of feature vectors for pixel i by DSIR in Section 4.

Frequency domain: Fourier transform of local blocks. If the feature of an image is periodic in space, then the feature of a local block in the space domain can be transformed to the frequency domain by the Fourier transform. This transform will highlight the periodical pattern (Weaver (1983)). This can be performed by the Fast Fourier Transform if the block size is of the power of 2. For a block with size $b \times b$, the two dimensional discrete Fourier transform can be expressed as

$$F(u, v) = \frac{1}{b^2} \sum_{x=0}^{b-1} \sum_{y=0}^{b-1} f(x, y) \exp\left[-\mathbf{i}2\pi\Big(\frac{ux}{b} + \frac{vy}{b}\Big)\right],$$

where $\mathbf{i} = \sqrt{-1}, u, v = 0, \ldots, b-1$. Because the image intensity is real-valued, the Fourier transform is symmetric about the center. By this symmetry, almost half of FFT calculation is redundant. Therefore, the feature in the frequency domain consists of $|F(u, v)|$ with dimension $b^2/2 + 2$ if b is power of 2.

Space-frequency domain: Gabor filter banks of local blocks. Human vision has demonstrated its superior capacity in detecting boundaries of desired objects. We use the vision model based on our previous work (Chen, Lu and Lin (1999) and Chen, Lu and Han (2001)). Other similar approaches could be applied as well. We construct a neuroimage by convolving the observed image with a bank of specific frequencies and orientation bands, such as a bank of Gabor functions. The general form of a Gabor function is given by

$$g(x,y) = \exp\{-[(x-x_0)^2a^2 + (y-y_0)^2b^2]\pi\} \\ \times \exp\{-2\pi\mathbf{i}[u_0(x-x_0) + v_0(y-y_0)]\},$$

and its Fourier transform is

$$G(u,v) = \exp\left\{-\frac{1}{\pi}\Big[\frac{(u-u_0)^2}{a^2} + \frac{(v-v_0)^2}{b^2}\Big]\right\} \\ \times \exp\{-2\pi\mathbf{i}[x_0(u-u_0) + y_0(v-v_0)]\}.$$

Each local block is convolved with a bank of Gabor filters with different orientations and frequencies. We employed the so-called *G-vector* as the feature vector at pixel (I,J), which is computed by

$$G_V(I,J) = \{g_{pk}(I,J), g_{nk}(I,J); k = 1,\ldots,r\},$$

where $g_{pk}(I,J)$ and $g_{nk}(I,J)$ are the summations of the positive and negative values for the neuroimage that is the convoluted image with the kth Gabor filter. Thus, the dimension of the feature vector is $2r$ in the space-frequency domain. For instance, we can consider a bank of $r = 3 \times 8 = 24$ Gabor filters that are designed with three scales of center frequencies, $\sqrt{2}/2, \sqrt{2}$ and $b\sqrt{2}/4 = 2\sqrt{2}$ when $b = 8$, as well as eight orientations of angles, 0, 30, 60, 90, 120, 150, 180 and 210 degrees.

4. Segmentation Procedures

For each pixel i, the training set is denoted as $\{\mathcal{Y}^{(i)}, \mathcal{X}^{(i)}\}$, where $\mathcal{X}^{(i)}$ is defined at (4) and $\mathcal{Y}^{(i)} = \{y_l^{(i)}(t_j), j = 1,\ldots,m, l \in \mathcal{N}_q^{(i)}\}$, $i = 1,\ldots,n$. The model of DSIR for this training set becomes

$$y_l^{(i)}(t_j) = f^{(i)}(\beta_1^{(i)\prime}\mathbf{x}_l^{(i)}(t_j),\ldots,\beta_K^{(i)\prime}\mathbf{x}_l^{(i)}(t_j), \epsilon_l^{(i)}(t_j)), \tag{5}$$

where $i = 1,\ldots,n$, $j = 1,\ldots,m$, and $l \in \mathcal{N}_q^{(i)}$.

We describe the algorithm of estimating the *e.d.r.* directions for each pixel i based on the training data as follows.

1. Compute the sample mean and sample covariance matrix of the $\mathbf{x}_i(t_j)$'s:

$$\bar{\mathbf{x}}^{(i)} = (m|\mathcal{N}_q^{(i)}|)^{-1} \sum_{l\in\mathcal{N}_q^{(i)}} \sum_{j=1}^{m} \mathbf{x}_l^{(i)}(t_j),$$

$$\hat{\Sigma}_{\mathrm{XX}}^{(i)} = (m|\mathcal{N}_q^{(i)}| - 1)^{-1} \sum_{l\in\mathcal{N}_q^{(i)}} \sum_{j=1}^{m} [\mathbf{x}_l^{(i)}(t_j) - \bar{\mathbf{x}}^{(i)}][\mathbf{x}_l^{(i)}(t_j) - \bar{\mathbf{x}}^{(i)}]', \tag{6}$$

where $|\mathcal{N}_q^{(i)}|$ is the number of observations in $\mathcal{N}_q^{(i)}$.

2. Sort the data by the range of $\{y_l^{(i)}(t_j), j = 1, \ldots, m,\ l \in \mathcal{N}_q^{(i)}\}$ and divide the data set into $H^{(i)}$ slices, denoted $I_1, \ldots, I_{H^{(i)}}$. Let the proportion of all observed $y_l^{(i)}(t)$'s that fall in the hth slice be $p_h^{(i)}$.

3. Within each slice, compute the sample mean of $\mathbf{x}_l^{(i)}(t_j)$'s, denoted by $\bar{\mathbf{x}}_h^{(i)}$, that is,

$$\bar{\mathbf{x}}_h^{(i)} = (m|\mathcal{N}_q^{(i)}|p_h^{(i)})^{-1} \sum_{\{j,l:y_l^{(i)}(t_j)\in I_h, l\in\mathcal{N}_q^{(i)}, j=1,\ldots,m\}} \mathbf{x}_l^{(i)}(t_j), \quad h = 1, \ldots, H^{(i)}. \tag{7}$$

4. Principal component analysis is conducted for the data $\bar{\mathbf{x}}_h^{(i)}$, $h = 1, \ldots, H^{(i)}$, with weights as follows. The weighted covariance matrix, $\hat{\Sigma}_W^{(i)} = \sum_{h=1}^H p_h^{(i)} (\bar{\mathbf{x}}_h^{(i)} - \bar{\mathbf{x}}^{(i)})(\bar{\mathbf{x}}_h^{(i)} - \bar{\mathbf{x}}^{(i)})'$, is computed first. Then, the eigenvalues and eigenvectors for $\hat{\Sigma}_W^{(i)}$ with respect to $\hat{\Sigma}_{\mathrm{XX}}^{(i)}$ are found by solving $\hat{\Sigma}_W^{(i)}\hat{\beta}_k^{(i)} = \hat{\lambda}_k^{(i)}\hat{\Sigma}_{\mathrm{XX}}^{(i)}\hat{\beta}_k^{(i)}$, where $k = 1, \ldots, p$ and $\hat{\lambda}_1^{(i)} \geq \hat{\lambda}_2^{(i)} \geq \cdots \geq \hat{\lambda}_p^{(i)}$.

5. The first K eigenvectors $\hat{\beta}_k^{(i)}$'s are used as the projection directions for the pixel i.

We let $\{\hat{\beta}_k^{(i)}, k = 1, \ldots, K\}$ be the DSIR projection directions for the pixel i. The segmentation method described above is applied to each pixel of an image. The major computation cost is only eigensystem decomposition.

For supervised segmentation, the class labels in the training set belong to G classes. That is, $y_l^{(i)}(t_j) \in \{1, \ldots, G\}$, where $j = 1, \ldots, m$, $l \in \mathcal{N}_q^{(i)}$,$i = 1, \ldots, n$. Those $p \times 1$ feature vectors, $\mathbf{x}_l^{(i)}(t_j)$, can be obtained from the space, frequency, or space-frequency domain. The DSIR algorithm can be applied to obtain the *e.d.r.* directions, $\hat{\beta}_1^{(i)}, \ldots, \hat{\beta}_K^{(i)}$, where $K \leq \min\{G - 1, p\}$. We can then project those feature vectors onto the *e.d.r.* directions. The means of the projected feature vectors for each class in the training set are the centroids. A pixel i in a test (or training) image is classified (or predicted) into the gth class if the projected feature vector of that pixel is closest to the centroid of gth class. Other classification techniques can be applied as well, like nonparametric discrimination by kernel density estimation, classification trees, and so on. Further discussion regarding the determination of K are in Li (1991, 2000) and Ferre (1998).

5. Experimental Results

The above segmentation procedures are performed on simulated and MR image sequences. All images used have 256 grey values and there is only one moving object. Thus, there are only two classes, object and background. The first projection direction from DSIR is used in segmentation.

5.1. Simulation studies

Six successive frames of images with a resolution of 80×80 pixels are simulated as in Figure 1 (a). The object looks like a ring and deforms from southwest to northeast. There are two classes in each time frame with gray levels at 90 and 120 respectively. The added noises are randomly distributed from a Gaussian distribution with mean 0 and standard deviation 20. The block size is chosen as 4×4 in the space domain. The features in the space-frequency domain are constructed by choosing a block size with 8×8 and a bank of $r = 3 \times 8 = 24$ Gabor filters. We first use frames 1 to 3 as training images and frame 4 as the test image. After supervised segmentation, the predicted class labels for frame 4 are further added to the training set to predict frame 5 and 6. We then performed the same steps by choosing frames 1 to 4 as training images and predicted frame 5. Again, the predicted class labels for frame 5 are further added to predict frame 6. Segmentation results in the space-frequency domains are shown in Figure 1 (b). The classification and prediction error rates for six frames in the space, frequency and space-frequency domains are reported on our web site.

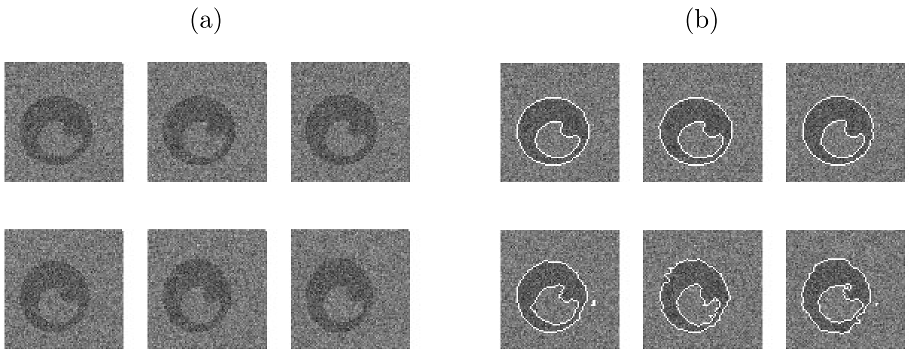

Figure 1. (a) The simulated image sequence of size 80×80 pixels. The object looks like a ring and deforms. The order of the sequence is from left to right and from top to bottom. (b) The segmentation results are displayed for the simulated images in the space-frequency domain. The top three images display the classification results and the bottom three images show the prediction results.

From these tables, the accuracy for predicting the object in frames 4 to 6 reaches over 97%. Because this object is of the same gray level with added random noises, the features in the space domain work well in the classification and prediction of the first five frames. However, the prediction error of frame 6 increases due to the motion that will mask the distinction between projected features between object and background in the space domain. The features in the frequency domain do not work as well as those in the space domain due to motion artifacts. The features in the space-frequency domain also have the least motion artifacts in Figure 1 (b). Although the classification and prediction errors in the space-frequency domain are larger, the segmented boundaries of the object for frames 5 and 6 are smoother. Thus, for simple textures in these simulation studies, the features in the space domain are sufficient for DSIR in classification and prediction.

5.2. Clinical studies of MR image sequences

We investigate the performance of our method on a MR image sequence of the epi- and endo-cardial surfaces of myocardium. Since the heart is an organ that exhibits motion, the examination of its image characteristics with 2D images sequence reveals useful information about physical condition. Figure 2 (a) shows a heart MR image sequence of 6 time frames in the region-of-interest (ROI). The goal in this experiment is to extract the boundary description of the inner and outer walls for the endocardium of the left ventricle. The training set can be obtained from the first three images by a medical expert. If we apply the SIR method as a supervised segmentation technique to segment a single image and predict the next image, the boundary of the endocardium is not segmented clearly in Figure 3. The boundary of the endocardium is connected with other tissue both in the classification result of the left image and the prediction result

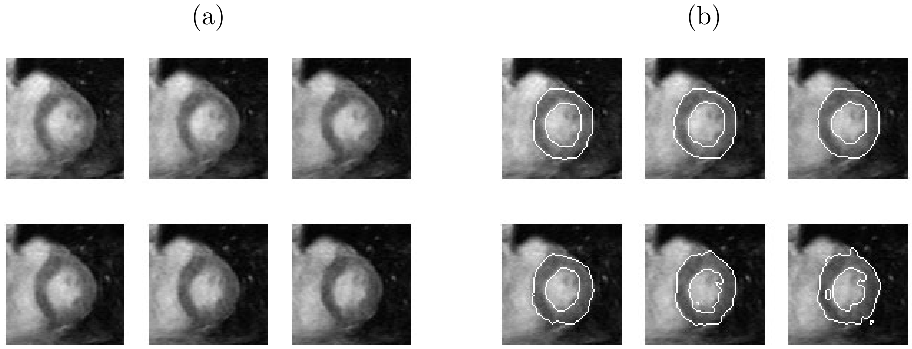

Figure 2. (a) The MR image sequence of myocardium of size 84 × 84 pixels in the ROI. The frame order is from left to right and from top to bottom. (b) The segmentation results are displayed for the MR image sequence of myocardium in space-frequency domain. The top three images display the classification results and the bottom three images show the prediction results.

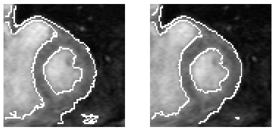

Figure 3. The segmentation result of applying SIR in the supervised segmentation of myocardium. The left image is obtained by the classification of frame 1 and the right image is obtained by the prediction of frame 2.

of the right image. Hence, it is necessary to apply the DSIR model to segment the boundary of the endocardium from a sequence of MR images. Use of our proposed procedure produces the segmentation results in the space-frequency domain as shown in Figure 2 (b). The errors in comparison to the ideal boundaries provided by the doctor are reported on our web site.

5.3. Comparison studies

To compare the performances of projection directions found by different approaches, we first apply the PCA, ICA and SIR on a single image. Then, we extend the study to motion sequences by applying DSIR.

For the image segmentation of a single image, image patches of projection directions found by PCA or ICA usually resemble edges with different lengths and widths that capture the spatial information of position, orientation, spatial frequencies and phases of objects presented in the images (Bell and Sejnowski (1997), Lewicki and Olshausen (1999) and Lee and Lewicki (2002)). Typically, one image is used and a subset of images patches are randomly selected to perform the PCA or ICA. Then the leading PCA/ICA directions represent the features of objects in the images, such as lips in face recognition.

Consider the first image from four consecutive images with natural textures in Figure 4 (a). Every image is of size 80 × 80 pixels that has an object of one texture on a background of another texture. The training set for classification is obtained by randomly selecting 1000 image patches with size 8×8 in the first image, and the second image is used as the test set for prediction. The PCA and ICA components are displayed as image patches in Figure 5 (a) and (b). These components show an increasing order of spatial frequency. However, the PCA and ICA components are different because they are derived by different criteria. We have applied the algorithm FastICA in Hyvarinen and Oja (2000) to estimate the ICA components one by one. For the PCA/ICA components with the high order of frequencies, the image patches have the checkerboard-like patterns locally. The checkerboard effect is produced when the resulting directions are redundant (Lewicki and Olshausen (1999) and Bins and Draper (2001)).

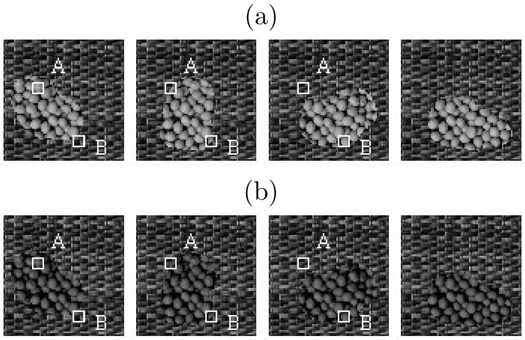

Figure 4. (a) Four consecutive frames of a moving and deforming object with one texture in a background of another texture. (b) The texture image sequence is re-scaled so that the gray levels of the local blocks in object and background have similar mean values. Two rectangles with size 8×8 (labelled as A and B) are selected to display the first DSIR projection directions.

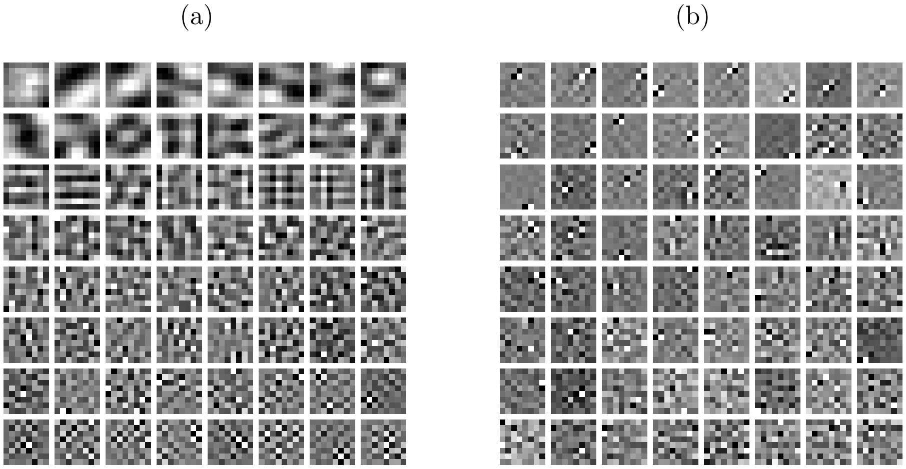

Figure 5. (a) The PCA components are displayed for the first image in Figure 4 (a) in the decreasing order of non-negative eigenvalues, from left to right and top to bottom. (b) The ICA components, using a deflation scheme, are displayed for the first image in Figure 4 (a).

Similarly to the PCA/ICA, we can also apply the SIR on the first image in Figure 4 (a). The resulting projection directions are shown in decreasing order of the corresponding non-negative eigenvalues in Figure 6. The first eigenvalue λ_1 corresponding to first projection direction accounts for $\lambda_1/\sum_i \lambda_i = 99.9\%$ of the variance. Although the first eigenvector has checkerboard patterns that appear in some local spots, the weights in the middle region have more contribution to classify class labels. The other image patches have more evident checkerboard-like patterns. Since there are two classes in this example, the first eigenvector of SIR is the only component that carries information for classification from the perspective of Fisher's linear discriminant analysis (Chen and Li (2001)).

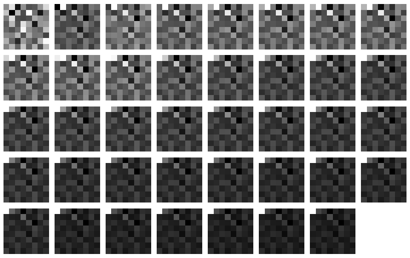

Figure 6. The projection directions of SIR in decreasing order of nonnegative corresponding eigenvalues, from left to right and top to bottom.

Figure 7 (a) shows the segmentation results using the first projection direction in the space domain. These results indicate that PCA, ICA and SIR can identify the object in the space domain with averaged error rates lower than 0.0352. Moreover, SIR performed better than PCA and ICA with classification error rate 0.0340 and prediction error rate 0.0289.

Next, we re-scale the grey levels of the objects so that the average gray levels of local blocks in the object texture and background texture are similar, as in Figure 4 (b). This makes the task of segmentation more difficult. Using the same setting as in the previous example, we apply PCA, ICA and SIR to the first image in Figure 4 (b). The patterns of the PCA and ICA components are similar to those found in the previous example. The first eigenvalue of the SIR components corresponding to first projection direction accounts for 99.9% of the variance, and the remaining image patches show more checkerboard-like patterns than the previous example.

Figure 7 (b) displays classification (at the left column) and prediction results (at the right column) in the space-frequency domain for SIR. (The segmentation results for PCA and ICA are available in a technical report.) According to the results, PCA, ICA and SIR fail to identify the object in these three feature domains. The averaged classification and prediction error rates range from 0.2429 to 0.4444, as shown in Table 1. This failure can be attributed to the fact that the training set was formed globally. The change of local structures can affect classification and prediction results. This motivated us to develop the DSIR approach on motion segmentation to explore stable features in local blocks by spatial, frequential and temporal analysis.

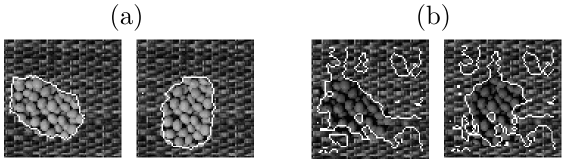

Figure 7. The segmentation results of applying SIR on a single image are displayed. The images in the left column in (a) and (b) are obtained by classification, and the images in the right column are obtained by prediction. (a) is for the first two images in Figure 4 (a) in space domain, and (b) is for the first two images in Figure 4 (b) in the space-frequency domain.

Table 1. The classification and prediction error rates of PCA, ICA and SIR for first two consecutive texture images in Figure 4 (b), where Obj represents the object, Bg denotes the background, and Total means the average error of the whole image.

		Classification			Prediction		
Methods	Feature	Obj	Bg	Total	Obj	Bg	Total
PCA	Space	0.4348	0.4480	0.4444	0.4389	0.4457	0.4438
	FFT	0.3652	0.4022	0.3920	0.3514	0.3990	0.3858
	Gabor	0.2130	0.3189	0.2890	0.2673	0.3307	0.3127
ICA	Space	0.3980	0.4371	0.4263	0.3853	0.4340	0.4205
	FFT	0.3754	0.4394	0.4218	0.3921	0.4457	0.4309
	Gabor	0.2129	0.3189	0.2889	0.2673	0.3307	0.3127
SIR	Space	0.4908	0.3483	0.3875	0.3826	0.4358	0.4211
	FFT	0.3530	0.4205	0.4020	0.3751	0.4280	0.4134
	Gabor	0.1570	0.2767	0.2429	0.2028	0.2914	0.2662

To explore the capability of DSIR in capturing the directions of features localized in space and time, we consider four consecutive frames of natural textures images given in Figure 4 (a). The first three frames are used as the training set to construct projection directions, and the last image frame is used as the test set. The feature vectors are formed by choosing a block size 8 × 8 with the 9-order neighborhood system $|\mathcal{N}_9^{(i)}|$. For each pixel i in the image, those 64 gray levels (e.g., in the space domain) are averaged (or projected) by the weights in the projection direction of β. The class label is assumed to be determined by a nonlinear function through the projected values and random noises. Assume the weights in projection directions hold for the next time frame, then the DSIR model (2) fits in this situation.

Two rectangles of size 8 × 8 (labelled as A and B) are selected to investigate the first projection directions by DSIR. These two rectangles contain background and moving textures. The resulting patterns for the rectangle A and B are shown in Figure 8, in which each image patch represents the first projection direction by DSIR and corresponds to a single pixel in the rectangle A or B. The brightness of each image pixel is proportional to its weight in the projection direction. These segmentation results (in the technical report) demonstrate that the moving boundaries of objects are identified by the proposed approach in all three feature domains with small error rates, see Table 2 and 3 (a).

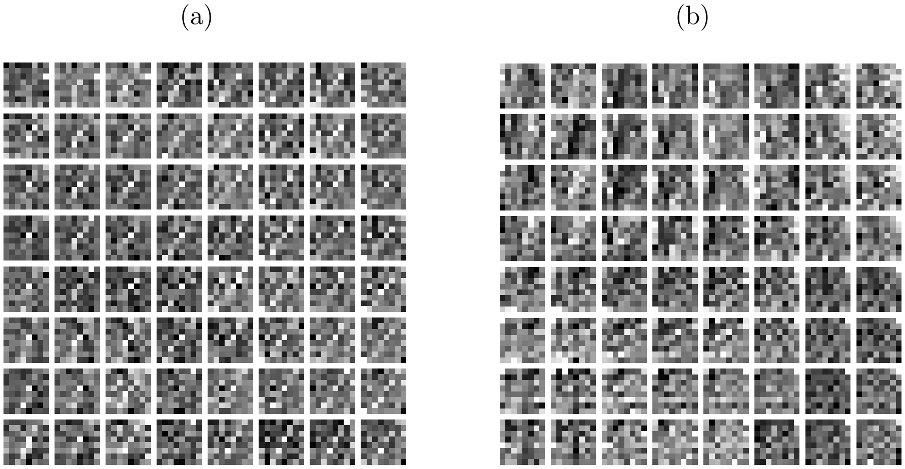

Figure 8. The first projection directions of DSIR for the selected block A and B in Figure 4 (a) are displayed. Each image patch represents the first projection direction of DSIR on a single pixel in the rectangle A or B.

Table 2. The classification error rates of DSIR for texture image sequence in Figure4 (a).

	Frame 1			Frame 2			Frame 3		
Feature	Obj	Bg	Total	Obj	Bg	Total	Obj	Bg	Total
Space	0.0034	0.0041	0.0039	0.0122	0.0018	0.0047	0.0134	0.0016	0.0049
FFT	0.0089	0.0031	0.0047	0.0047	0.0023	0.0030	0.0087	0.0044	0.0056
Gabor	0.0055	0.0027	0.0035	0.0088	0.0040	0.0054	0.0061	0.0041	0.0046

Table 3. The prediction error rates of DSIR for Frame 4 in Figure 4 (a) and (b).

	Figure 4 (a)			Figure 4 (b)		
Feature	Obj	Bg	Total	Obj	Bg	Total
Space	0.0655	0.0161	0.0298	0.2438	0.0265	0.0869
FFT	0.0858	0.0247	0.0417	0.1458	0.0161	0.0522
Gabor	0.0932	0.0232	0.0432	0.1262	0.0259	0.0545

Next, we apply DSIR to the re-scaled images sequences given in Figure 4 (b). Using the same setting as above, the resulting projection directions for the rectangle A and B exhibit more spatial variation than those in Figure 8. Figure 9 shows the segmentation results. The classification and prediction error rates are reported on our web site, they are reduced significantly comparing to those by PCA, ICA and SIR approaches. The performance in the space domain is worse than that in the frequency and space-frequency domains. This illustrates the necessity of DSIR with spatial-frequential analysis for motion segmentation.

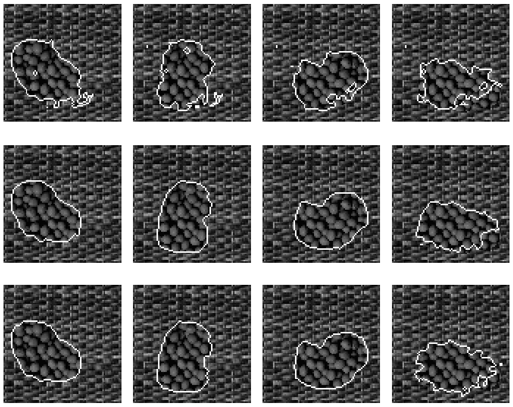

Figure 9. The segmentation results of Figure 4 (b) are displayed for the texture images sequence in the space domain (at the top row), the frequency domain (at the second row) and the space-frequency domain (at the bottom row). For each domain, the first three images show the classification results and the last one displays the prediction result.

Unlike the PCA, ICA and SIR approaches on a single image, DSIR uses local blocks from the successive images to construct the projection directions. In our experimental results, there are no apparent patterns of DSIR components like those of PCA/ICA components. The image patches of DSIR components contain the spatiotemporal information regarding the weights at pixels that could classify the class labels differently than the spatial information contained in the PCA/ICA components.

6. Conclusion and Discussion

The advantage in our recursive procedure of updating is that we can adopt the results of segmentation in previous images to predict future images. We only need a small number of sequence of images in a training set with the capability of feature extraction and dimension reduction. Prediction errors will accumulate as the segmentation process proceeds. Further tuning by image processing techniques, such as smoothing, erosion, dilation, and so on, can be applied to obtain preferred segmentation.

DSIR extracts local motion information captured by spatial-frequential properties over a short time interval. An alternative approach to spatial-temporal structures can be the space-time Gabor filters (Heeger (1998)), in which the filters are created by extending the Gabor function to include the temporal variable in both Gaussian and trigonometric functions. Once the local statistics of these features are evaluated, supervised learning methods similar to the SIR technique can be implemented.

Given a sequence of images of size m, the selection of block size (b), the order of neighborhood system (q), the parameters in the feature extraction (like the number of Gabor filters, r, in the space-frequency domain) and other parameters need further study. The application of other classification tools and dimension reduction techniques, like the principal Hessian directions algorithm in Li (1992), are also of interest. Besides the application of motion segmentation, DSIR can provide a useful tool for mining the relationship of responses and regressors over time from the perspective of dimension reduction.

Acknowledgement

Due to space limitations, the figures and tables are partially presented in the article. The complete version of figures and tables are available in our technical report at http://www.stat.nctu.edu.tw/faculty/hslu/techreport.htm. The authors thank Dr. Hui-Cheng Cheng at Veterans General Hospital, Taipei, Taiwan, for providing and explaining the clinical trial data analyzed herein. The authors also thank an associate editor and the referees for their valuable comments. This research was supported by the grants from National Science Council at Taiwan, R.O.C.

Appendix: Proof of Theorem 1

Assume that $E[\mathbf{x}(t)] = 0$ without loss of generality and let $B = (\beta_1, \ldots, \beta_K)$, a p by K matrix. Then

$$\begin{aligned} E[\mathbf{x}(t)|\beta_1'\mathbf{x}(t), \ldots, \beta_K'\mathbf{x}(t)] &= E[\mathbf{x}(t)|B'\mathbf{x}(t)] \\ &= [\mathrm{Cov}\,(B'\mathbf{x}(t))^{-1}\,\mathrm{Cov}\,(B'\mathbf{x}(t), \mathbf{x}(t))]'B'\mathbf{x}(t) \qquad (1) \\ &= [(B'\,\mathrm{Cov}\,(\mathbf{x}(t))B)^{-1}B'\,\mathrm{Cov}\,(\mathbf{x}(t))]'B'\mathbf{x}(t). \end{aligned}$$

The equation (1) is obtained by regressing each component of $\mathbf{x}(t)$ separately against $B'\mathbf{x}(t)$. Let $\mathbf{k}(y) = (B'\,\mathrm{Cov}\,(\mathbf{x}(t))B)^{-1}E[B'\mathbf{x}(t)|y(t)]$, then

$$\begin{aligned} E[\mathbf{x}(t)|y(t)] &= E(E[\mathbf{x}(t)|B'\mathbf{x}(t), \epsilon(t)]|y(t)) \\ &= E(E[\mathbf{x}(t)|B'\mathbf{x}(t)]|y(t)) \\ &= (B'\,\mathrm{Cov}\,(\mathbf{x}(t)))\mathbf{k}(y). \end{aligned}$$

References

Bell, A. J. and Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Res.* **37**, 3327-3338.

Bins, J. and Draper, B. A. (2001). Feature selection from huge feature sets. *Internat. Conference on Comput. Vision* **2**, 159-165.

Blake, A. and Isard, M. (1998). *Active Contours.* Springer-Verlag, Berlin.

Bouthemy, P. (1989). A maximum likelihood framework for determining moving edges. *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 499-511.

Chen, C. H. and Li, K. C. (1998). Can SIR be as popular as multiple linear regression? *Statist. Sinica* **8**, 289-316.

Chen, C. H. and Li, K. C. (2001). Generalization of Fisher's linear discriminant analysis via the approach of sliced inverse regression. *J. Korean Statist. Soc.* **30**, 193-217.

Chen, C. M., Lu, H. H.-S. and Lin, Y. C. (1999). An early vision based snake model for ultrasound image segmentation. *Ultrasound in Medicine and Biology* **26**, 273-285.

Chen, C. M., Lu, H. H.-S. and Han, K. C. (2001). A textural approach based on Gabor functions for texture edge detection in ultrasound images. *Ultrasound in Medicine and Biology* **27**, 513-534.

Deng, Y. and Manjunath, B. S. (2001). Unsupervised segmentation of color-texture regions in images and video. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**, 800-810.

Doretto, G., Chiuso, A., Wu, Y. and Soatto, S. (2003). Dynamic textures. *Internat. J. Comput. Vision* **51**, 91-109.

Duncan, J. S., Owen, R. L., Staib, L. H. and Anandan, P. (1991). Measurement of non-rigid motion using contour shape descriptors. *Comput. Vision and Pattern Recognition* **91**, 318-324.

Dunn, D., Higgins, W. E. and Wakeley, J. (1994). Texture segmentation using 2-D Gabor elementary functions. *IEEE Trans. Pattern Anal. Mach. Intell.* **16**, 130-149.

Fablet, R., Bouthemy, P. and Perez, P. (2002). Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval. *IEEE Trans. Image Process.* **11**, 393-407.

Ferre, L. (1998). Determining dimension in sliced inverse regression and related methods. *J. Amer. Statist. Assoc.* **93**, 132-140.

Heeger, D. J. (1988). Optical flow using spatiotemporal filters. *Internat. J. Comput. Vision* **1**, 270-302.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks* **10**, 626-634.

Hyvarinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks* **13**, 411-430.

Isard, M. and Blake, A. (1996). Contour tracking by stochastic propagation of conditional density. *European Conference on Computer Vision* **1**, 343-356.

Lee, T. W. and Lewicki, M. S. (2002). Unsupervised image classification, segmentation, and enhancement using ICA mixture models. *IEEE Trans. Image Process.* **11**, 270-279.

Lewicki, M. S. and Olshausen, B. A. (1999). A probabilistic framework for adaptation and comparison of image codes. *J. Optical Soc. Amer.* **16**, 1587-1600.

Li, K. C. (1991). Sliced inverse regression for dimensional reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316-342.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87**, 1025-1039.

Li, K. C. (2000). High dimensional data analysis via the SIR/PHD approach. Lecture Notes are available at http://www.stat.ucla.edu/~ kcli/.

Li, K. C., Aragon, Y. and Thomos-Agan, C. (1995). Analysis of multivariate outcome data: SIR and a nonlinear theory of Hotelling's most predictable variates. Technical Report, Department of Statistics, UCLA.

Lu, H. H.-S., Chen, C. M. and Wu, J. S. (2001). Statistical analysis of liver cirrhosis in ultrasound images by fractal dimension, dimension reduction and classification trees. *The 5th World Multi-Conference on Systemics, Cybernetics and Informatics*, Vol. XIII, Part II, 351-356.

Malik, J. and Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *J. Optical Soc. Amer. A* **7**, 923-932.

Mansouri, A. R. and Konrad, J. (2003). Multiple motion segmentation with level sets. *IEEE Trans. Image Process.* **12**, 201-220.

Mikic, I., Krucinski, S. and Thomas, J. D. (1998). Segmentation and tracking in echocardiographic sequences: active contours guided by optical flow estimates. *IEEE Trans. Medical Imaging* **17**, 274-284.

Nguyen, H. T., Worring, M. and Dev, A. (2000). Detection of moving objects in video using a robust motion similarity measure. *IEEE Trans. Image Process.* **9**, 137-141.

Osher, S. and Paragios, N. (2003). *Geometric Level Set Methods in Imaging, Vision and Graphics.* Springer-Verlag, New York.

Tan, T. N. (1995). Texture edge detection by modelling visual cortical channels. *Pattern Recognition* **28**, 1283-1298.

Vasconcelos, N. and Lippman, A. (2001). Empirical bayesian motion segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**, 217 -221.

Weaver, H. J. (1983). *Applications of Discrete and Continuous Fourier Analysis.* Wiley, New York.

Wu, Y., Zhu, S. C. and Guo, C. (2002). Statistical modelling of texture sketch. *European Conference of Computer Vision* **2352**, 240-254.

Wu, Y., Zhu, S. C. and Liu, X. (2000). Equivalence of Julesz texture ensembles and FRAME models. *Internat. J. Comput. Vision* **38**, 247-265.

Zhu, S. C., Liu, X. and Wu, Y. (2000). Exploring texture ensembles by efficient Markov Chain Monte Carlo - towards a 'Trichromacy' theory of texture. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 554-569.

Zhu, S. C., Wu, Y. and Mumford, D. B. (1998). Filter, random field, and maximum entropy (FRAME): towards a unified theory for texture modelling. *Internat. J. Comput. Vision* **27**, 107-126.

Institute of Statistics, National Chiao-Tung University, 1001 Ta Hsueh Road, Hsinchu 30050, Taiwan, R.O.C.

E-mail: hmwu.st86g@nctu.edu.tw

Institute of Statistics, National Chiao-Tung University, 1001 Ta Hsueh Road, Hsinchu 30050, Taiwan, R.O.C.

E-mail: hslu@stat.nctu.edu.tw

(Received August 2002; accepted September 2003)