IMM
DEPARTMENT OF MATHEMATICAL MODELLING

Technical University of Denmark
DK-2800 Lyngby – Denmark

J. No. 3D
October 7, 1997
RL

3-D CONTEXTUAL BAYESIAN CLASSIFIERS

Rasmus Larsen

TECHNICAL REPORT

IMM-REP-97-16

IMM

3-D CONTEXTUAL BAYESIAN CLASSIFIERS

Rasmus Larsen

TECHNICAL REPORT

IMM-REP-97-16

IMM

3-D Contextual Bayesian Classifiers

Rasmus Larsen, Assistant Research Professor, Ph.D.
Department of Mathematical Modelling, Building 321
Technical University of Denmark, DK-2800 Lyngby, Denmark

EMAIL: rl@imm.dtu.dk
WWW: www.imm.dtu.dk/∼rl
PHONE: +45 4588 1433
FAX: +45 4588 1397

Abstract

In this paper we will consider extensions of a series of Bayesian 2-D contextual classification pocedures proposed by Owen [1], Hjort & Mohn [2] and Welch & Salter [3] and Haslett [4] to 3 spatial dimensions. It is evident that compared to classical pixelwise classification further information can be obtained by taking into account the spatial structure of image data. The 2-D algorithms mentioned above consist of basing the classification of a pixel on the simultaneous distribution of the values of a pixel and its four nearest neighbours. This includes the specification of a Gaussian distribution for the pixel values as well as a prior distribution for the configuration of class variables within the cross that is made of a pixel and its four nearest neighbours. We will extend these algorithms to 3-D, i.e. we will specify a simultaneous Gaussian distribution for a pixel and its 6 nearest 3-D neighbours, and generalise the class variable configuration distributions within the 3-D cross given in 2-D algorithms. The new 3-D algorithms are tested on a synthetic 3-D multivariate dataset.

Keywords

Classification, Segmentation, 3-D, Contextual methods

I. Introduction

When applying classical classification schemes in image analysis the spatial structure of the datasets is neglected. This is non-satisfying, because further information obviously can be drawn from the spatial arrangement of pixels, since neighbouring pixels tend to be of the same class. We will refer to this type of information as contextual information.

Contextual information can be taken into account in a number of ways when performing classification. One important way is to include (derived) features that hold information of the neighbourhood of a given pixel, i.e. contextual features. Another way is introducing the spatial nature directly in the algorithms. Several algorithms have been proposed in the 2-D case. In [5] it is proposed simply to augment the feature vector with the average of the feature vector from the four neighbouring pixels. In order to find the maximum a posteriori estimate in a Markov random field model stochastic relaxation has been proposed in [6]. An approximation to the maximum a posteriori estimate using iterated conditional modes was proposed in [7]. In [1], [8], [4] a classification scheme for 2-D images that bases the actual classification of pixel on the feature vectors of the pixel itself and those of the 4 nearest neighbours is introduced. In [4] it is assumed that classes of the nearest neighbours of a pixel are conditionally independent given the class of the center pixel, whereas in [1], [8] it is assumed that the pixel size is small relative to the grains of the pattern under study, which leads to a vastly reduced set of possible class configurations among a pixel and its four nearest neighbours.

In this article we will extend the algorithms proposed in [1], [8], [3], [4] to 3-D images, and carry out a series of tests on a synthetic 3-D image.

II. METHODS

In this Section we will develop a 3-D contextual classification rule, specify a Gaussian distribution for the observed (and derived) features, and specify a prior distribution for the class variable.

A. Construction of a Contextual Classification Rule

Suppose that a pixel is an observation from one of the classes (populations) $\pi_1, \pi_2, \dots, \pi_k$. The classification of the observation depends on the vector of features $\boldsymbol{X} = (X_1, X_2, \dots, X_p)^T$ of that pixel. Furthermore, let us assume knowledge of the prior distribution of the classes, i.e. the prior probabilities, $P(C = \pi_i) = p_i$, $i = 0, 1, \dots, k$ where C is the class variable. This distribution determines the probability with which an arbitrary feature vector has been generated from a particular class.

We will denote the feature vector of the neighbouring pixels $\boldsymbol{X}_N$, $\boldsymbol{X}_S$, $\boldsymbol{X}_E$, $\boldsymbol{X}_W$, $\boldsymbol{X}_T$, and $\boldsymbol{X}_B$ for the north, south, east, west, top, and bottom pixel, respectively. The augmented feature vector consisting of the features vectors for the neighbours of a pixel will be denoted $\boldsymbol{D}_\Delta = (\boldsymbol{X}_N^T, \boldsymbol{X}_S^T, \boldsymbol{X}_E^T, \boldsymbol{X}_W^T, \boldsymbol{X}_T^T, \boldsymbol{X}_B^T)^T$. The augmented feature vector consisting of the feature vector of a pixel itself and those of its neighbours will be denoted $\boldsymbol{D} = (\boldsymbol{X}^T, \boldsymbol{D}_\Delta^T)^T$.

We obtain the Bayes solution for the case of equal losses by setting the discriminant score equal to the maximum a posteriori probability. The posterior distribution for the class variable becomes

$$\begin{aligned} f(\pi_\nu|\boldsymbol{d}) &= P(C{=}\pi_\nu|\boldsymbol{D}{=}\boldsymbol{d}) = \frac{P(C=\pi_\nu)P(\boldsymbol{D}=\boldsymbol{d} \mid C=\pi_\nu)}{\sum_{i=1}^{k} P(C=\pi_i)P(\boldsymbol{D}=\boldsymbol{d} \mid C=\pi_i)} \\ &= \frac{\sum_{a,b,c,d,e,f} p_\nu P(\boldsymbol{D}{=}\boldsymbol{d} \mid \boldsymbol{C}{=}(\pi_\nu,\pi_a,\pi_b,\pi_c,\pi_d,\pi_e,\pi_f)) g(\pi_a,\pi_b,\pi_c,\pi_d,\pi_e,\pi_f \mid \pi_\nu)}{h(\boldsymbol{d})} \end{aligned} \tag{1}$$

where $h(\boldsymbol{d})$ is the unconditional density of the augmented feature vector, (a, b, c, d, e, f) is one of the possible k^6 configurations of the class variables of the neighbouring pixels, $\boldsymbol{C}$ is the class configuration corresponding to the augmented feature vector $\boldsymbol{D}$, and $g(\pi_a, \pi_b, \pi_c, \pi_d, \pi_e, \pi_f \mid \pi_\nu)$ is the probability of the configuration of the class variables of the neighbouring pixels given that the center pixel has class π_ν.

If we furthermore assume that the density of the feature vector of the centerpixel is independent of the classes of the neighbouring pixels, i.e.

$$\begin{aligned} &P(\boldsymbol{D}{=}\boldsymbol{d} \mid \boldsymbol{C}{=}(\pi_\nu,\pi_a,\pi_b,\pi_c,\pi_d,\pi_e,\pi_f)) = \\ &\quad P(\boldsymbol{D}_\Delta = \boldsymbol{d}_\Delta \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{C} = (\pi_\nu,\pi_a,\pi_b,\pi_c,\pi_d,\pi_e,\pi_f)) P(\boldsymbol{X} = \boldsymbol{X} \mid C = \pi_\nu), \end{aligned} \tag{2}$$

we can rewrite Equation (1) to

$$f(\pi_\nu|\boldsymbol{d}) = \frac{p_\nu P(\boldsymbol{X} = \boldsymbol{X} \mid C = \pi_\nu)}{h(\boldsymbol{d})} R_\nu(\boldsymbol{D}). \tag{3}$$

This is the posterior probability of the non-contextual rule multiplied by $R_\nu(\boldsymbol{D})$, which is a contextual adjustment factor given by

$$R_\nu(\boldsymbol{D}) = \sum_{a,b,c,d,e,f} g(\pi_a, \pi_b, \pi_c, \pi_d, \pi_e, \pi_f \mid \pi_\nu) P(\boldsymbol{D}_\Delta = \boldsymbol{d}_\Delta \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{C} = (\pi_\nu, \pi_a, \pi_b, \pi_c, \pi_d, \pi_e, \pi_f)) \tag{4}$$

Contextual information may come into the model in two ways, first in the spatial dependence of the feature vectors (specification of the conditional distribution of the augmented feature vector), and second in the specification of prior distribution of the class configurations, g.

B. *Specification of a Gaussian distribution*

Following the 2-D algorithm as specified by Hjort et al. [8] we assume that each feature vector may be written as a sum of two terms, i.e. $\boldsymbol{X} = \boldsymbol{Y} + \boldsymbol{\epsilon}$, where the $\boldsymbol{Y}$ terms are independent given the classes and model the class dependency of the feature vectors, i.e.

$$(\boldsymbol{Y} \mid C = \pi_i) \in N(\boldsymbol{\mu}_i, (1-\theta)\boldsymbol{\Sigma}) \tag{5}$$

and $(\boldsymbol{\epsilon}^T_{\boldsymbol{s}(1)}, \ldots, \boldsymbol{\epsilon}^T_{\boldsymbol{s}(N)})$ is multinormal and model an autocorrelated noise term with

$$\begin{aligned} E\{\boldsymbol{\epsilon}_{\boldsymbol{s}(j)}\} &= 0 \\ E\{\boldsymbol{\epsilon}_{\boldsymbol{s}(j_1)}\boldsymbol{\epsilon}^T_{\boldsymbol{s}(j_2)}\} &= \rho^{\|\boldsymbol{s}(j_1)-\boldsymbol{s}(j_2)\|_2}\theta\boldsymbol{\Sigma} \end{aligned} \tag{6}$$

$\boldsymbol{s}(j)$ refers to the spatial position of pixel number j, and $\|\boldsymbol{s}(j_1) - \boldsymbol{s}(j_2)\|_2$ is the Euclidean distance (i.e. the 2-norm) between pixels j_1 and j_2, N is the total number of pixels.

Alternative models for the correlogram include using the 1-norm (i.e the city-block or Manhattan distance), or the ∞-norm. ρ is the autocorrelation between first-order neighbours, and θ is the proportion of the covariance matrix $\boldsymbol{\Sigma}$ that is due to autocorrelated noise.

Here we have chosen to use an isotropic autocorrelation function. However, the extension to an anisotropic function is straightforward. In Figure 2 realisations of autocorrelated noise patterns corresponding to using the three different norms in Equation (6) are shown.

Now it is possible to write the conditional distribution of the augmented feature vector given that the classes are π_ν, π_a, π_b, π_c, π_d, π_e, and π_f, respectively

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{X}_N \\ \boldsymbol{X}_S \\ \boldsymbol{X}_E \\ \boldsymbol{X}_W \\ \boldsymbol{X}_T \\ \boldsymbol{X}_B \end{bmatrix} \in N_{7p}\left[\begin{bmatrix} \boldsymbol{\mu}_\nu \\ \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \\ \boldsymbol{\mu}_c \\ \boldsymbol{\mu}_d \\ \boldsymbol{\mu}_e \\ \boldsymbol{\mu}_f \end{bmatrix}, \begin{bmatrix} 1 & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha \\ \alpha & 1 & \gamma & \beta & \beta & \beta & \beta \\ \alpha & \gamma & 1 & \beta & \beta & \beta & \beta \\ \alpha & \beta & \beta & 1 & \gamma & \beta & \beta \\ \alpha & \beta & \beta & \gamma & 1 & \beta & \beta \\ \alpha & \beta & \beta & \beta & \beta & 1 & \gamma \\ \alpha & \beta & \beta & \beta & \beta & \gamma & 1 \end{bmatrix} \otimes \boldsymbol{\Sigma}\right] \tag{7}$$

where $\otimes$ denotes the Kronecker (tensor) product, and the definitions of α, β, and γ are given in Table I for the different autocorrelation model mentioned above. Note, that all these models are valid corellogram models [9], thus ensuring positive definiteness of the covariance matrix in Equation (7).

From Equation (7) using a result from [10] we find that the conditional distribution of the augmented feature vector for the neighbour pixels only is multinormal with mean

$$\begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \\ \boldsymbol{\mu}_c \\ \boldsymbol{\mu}_d \\ \boldsymbol{\mu}_e \\ \boldsymbol{\mu}_f \end{bmatrix} + \begin{bmatrix} \alpha\boldsymbol{\Sigma} \\ \alpha\boldsymbol{\Sigma} \\ \alpha\boldsymbol{\Sigma} \\ \alpha\boldsymbol{\Sigma} \\ \alpha\boldsymbol{\Sigma} \\ \alpha\boldsymbol{\Sigma} \end{bmatrix} \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_\nu) = \begin{bmatrix} \boldsymbol{\mu}_a+\alpha(\boldsymbol{X}-\boldsymbol{\mu}_\nu) \\ \boldsymbol{\mu}_b+\alpha(\boldsymbol{X}-\boldsymbol{\mu}_\nu) \\ \boldsymbol{\mu}_c+\alpha(\boldsymbol{X}-\boldsymbol{\mu}_\nu) \\ \boldsymbol{\mu}_d+\alpha(\boldsymbol{X}-\boldsymbol{\mu}_\nu) \\ \boldsymbol{\mu}_e+\alpha(\boldsymbol{X}-\boldsymbol{\mu}_\nu) \\ \boldsymbol{\mu}_f+\alpha(\boldsymbol{X}-\boldsymbol{\mu}_\nu) \end{bmatrix} \tag{8}$$

and covariance matrix $\boldsymbol{S}_\Delta \otimes \boldsymbol{\Sigma}$, where

$$\boldsymbol{S}_\Delta = \begin{bmatrix} 1-\alpha^2 & \gamma-\alpha^2 & \beta-\alpha^2 & \beta-\alpha^2 & \beta-\alpha^2 & \beta-\alpha^2 \\ \gamma-\alpha^2 & 1-\alpha^2 & \beta-\alpha^2 & \beta-\alpha^2 & \beta-\alpha^2 & \beta-\alpha^2 \\ \beta-\alpha^2 & \beta-\alpha^2 & 1-\alpha^2 & \gamma-\alpha^2 & \beta-\alpha^2 & \beta-\alpha^2 \\ \beta-\alpha^2 & \beta-\alpha^2 & \gamma-\alpha^2 & 1-\alpha^2 & \beta-\alpha^2 & \beta-\alpha^2 \\ \beta-\alpha^2 & \beta-\alpha^2 & \beta-\alpha^2 & \beta-\alpha^2 & 1-\alpha^2 & \gamma-\alpha^2 \\ \beta-\alpha^2 & \beta-\alpha^2 & \beta-\alpha^2 & \beta-\alpha^2 & \gamma-\alpha^2 & 1-\alpha^2 \end{bmatrix} \tag{9}$$

C. Specification of a prior distribution for the OHM model

Following the 2-D procedure developed by Owen [1] and Hjort et al. [8] (The OHM model) we assume that pixels in a scene are assigned populations by a stochastic process, we regard a scene with pixels that have not been assigned populations. Following [1], as the first step in the process we divide the scene by planes distributed by a stochastic process. Each pixel will now be part of a region. If the size of the regions are large compared to the pixel size, it can be assumed that on the borders between regions other patterns than the Q, R, and S patterns shown in Figure 1 will occur with very small probability. Let the probability of a pixel being an interior point be p. Furthermore, let the probability of a pixel being on a border parallel to two of the coordinate axes be q, let the probability of a pixel being on a border parallel to only one of the coordinate axes be r, and let the probability of a pixel being on a plane that is not paralel to any of the coordinate axes be $s = 1 - p - q - r$. All other configurations are assumed to occur with probability 0.

In the 2-D case Owen [1] employs a dividing mechanism devised in [11] that results in the parametrisation of the probabilities of the three patterns in the 2-D model by a Poisson field intensity. However, Hjort & Mohn [2] argue that the slightly parameter richer model of estimating the pattern probalities directly results in a more model-robust classification.

As the second step we assign a population to each region independently, according to the a priori probabilities for the populations. If two neighbouring regions are assigned the same population we can delete the border between these regions.

By rotation we obtain six, twelve, and eight different CR_2, CR_3, and CR_4 patterns, respectively, i.e. we have $6(k-1)$ configurations for the CR_2 pattern corresponding to the six orientations and the $k-1$ possibilities for the neighbour region class. Note that from the assumption of the regions being larger than the pixel size we also have that the pixels within the 'cross' in the CR_2, CR_3, and CR_4 cases that are different from the center pixel, all have the same class. In all given the center pixel class we have $1+6(k-1)+12(k-1)+8(k-1) = 26k-25$ different configurations. Which should be compared with k^6 configurations if the assumption of region size vs. pixel size was not applied. These patterns are assigned positive a priori probabilities, while all other patterns are assigned the probability zero.

Under these assumptions we have the following expression for the probabilities, for each of the possible patterns.

CR_1 :
$$g(\pi_\nu, \pi_\nu, \pi_\nu, \pi_\nu\pi_\nu, \pi_\nu \mid \pi_\nu) = p + (q + r + s) \cdot p_\nu$$

CR_2 :
$$\begin{aligned}
&g(\pi_i, \pi_\nu, \pi_\nu, \pi_\nu, \pi_\nu, \pi_\nu \mid \pi_\nu) = g(\pi_\nu, \pi_i, \pi_\nu, \pi_\nu, \pi_\nu, \pi_\nu \mid \pi_\nu) = \\
&g(\pi_\nu, \pi_\nu, \pi_i, \pi_\nu, \pi_\nu, \pi_\nu \mid \pi_\nu) = g(\pi_\nu, \pi_\nu, \pi_\nu, \pi_i, \pi_\nu, \pi_\nu \mid \pi_\nu) = \\
&g(\pi_\nu, \pi_\nu, \pi_\nu, \pi_\nu, \pi_i, \pi_\nu \mid \pi_\nu) = g(\pi_\nu, \pi_\nu, \pi_\nu, \pi_\nu, \pi_\nu, \pi_i \mid \pi_\nu) = \tfrac{1}{6} q p_i
\end{aligned}$$

CR_3 :
$$\begin{aligned}
&g(\pi_i, \pi_\nu, \pi_i, \pi_\nu, \pi_\nu, \pi_\nu \mid \pi_\nu) = g(\pi_i, \pi_\nu, \pi_\nu, \pi_i, \pi_\nu, \pi_\nu \mid \pi_\nu) = \\
&g(\pi_\nu, \pi_i, \pi_i, \pi_\nu, \pi_\nu, \pi_\nu \mid \pi_\nu) = g(\pi_\nu, \pi_i, \pi_\nu, \pi_i, \pi_\nu, \pi_\nu \mid \pi_\nu) = \\
&g(\pi_i, \pi_\nu, \pi_\nu, \pi_\nu, \pi_i, \pi_\nu \mid \pi_\nu) = g(\pi_i, \pi_\nu, \pi_\nu, \pi_\nu, \pi_\nu, \pi_i \mid \pi_\nu) = \\
&g(\pi_\nu, \pi_i, \pi_\nu, \pi_\nu, \pi_i, \pi_\nu \mid \pi_\nu) = g(\pi_\nu, \pi_i, \pi_\nu, \pi_\nu, \pi_\nu, \pi_i \mid \pi_\nu) = \\
&g(\pi_\nu, \pi_\nu, \pi_i, \pi_\nu, \pi_i, \pi_\nu \mid \pi_\nu) = g(\pi_\nu, \pi_\nu, \pi_i, \pi_\nu, \pi_\nu, \pi_i \mid \pi_\nu) = \\
&g(\pi_\nu, \pi_\nu, \pi_\nu, \pi_i, \pi_i, \pi_\nu \mid \pi_\nu) = g(\pi_\nu, \pi_\nu, \pi_\nu, \pi_i, \pi_\nu, \pi_i \mid \pi_\nu) = \tfrac{1}{12} r p_i
\end{aligned}$$

CR_4 :
$$\begin{aligned}
&g(\pi_i, \pi_\nu, \pi_i, \pi_\nu, \pi_i, \pi_\nu \mid \pi_\nu) = g(\pi_i, \pi_\nu, \pi_i, \pi_\nu, \pi_\nu, \pi_i \mid \pi_\nu) = \\
&g(\pi_i, \pi_\nu, \pi_\nu, \pi_i, \pi_i, \pi_\nu \mid \pi_\nu) = g(\pi_i, \pi_\nu, \pi_\nu, \pi_i, \pi_\nu, \pi_i \mid \pi_\nu) = \\
&g(\pi_\nu, \pi_i, \pi_i, \pi_\nu, \pi_i, \pi_\nu \mid \pi_\nu) = g(\pi_\nu, \pi_i, \pi_i, \pi_\nu, \pi_\nu, \pi_i \mid \pi_\nu) = \\
&g(\pi_\nu, \pi_i, \pi_\nu, \pi_i, \pi_i, \pi_\nu \mid \pi_\nu) = g(\pi_\nu, \pi_i, \pi_\nu, \pi_i, \pi_\nu, \pi_i \mid \pi_\nu) = \tfrac{1}{8} s p_i
\end{aligned}$$

where $\nu \neq i$, and $\nu, i = 1, \ldots, k$.

In this way we have obtained a huge reduction in the number of terms in the contextual classification rule.

D. Specification of a prior distribution for the WSH model

Alternatively, following the 2-D algorithms by Welch & Salter and Haslett [3], [4] (The WSH model), we may assume independence between the class variables of the neighbours given the center pixel class, i.e.

$$\begin{aligned}
&g(\pi_a, \pi_b, \pi_c, \pi_d, \pi_e, \pi_f \mid \pi_\nu) = \\
&\quad \phi(\pi_a|\pi_\nu)\phi(\pi_b|\pi_\nu)\phi(\pi_c|\pi_\nu)\phi(\pi_d|\pi_\nu)\phi(\pi_e|\pi_\nu)\phi(\pi_f|\pi_\nu).
\end{aligned} \tag{10}$$

Here $\phi(\pi_i \mid \pi_j) = P(C_A = \pi_i \mid C_B = \pi_j)$, where A and B are immediate neighbours, i.e. a neighbour transition probability.

The model leads to a considerable simplification of the formula for the posterior distribution of the center pixel class variable (Equation (1)) in the case of conditional independence of the feature vectors given the class variables, i.e. $\theta = 0 \vee \rho = 0 \Leftrightarrow \alpha = \beta = \gamma = 0$ in Equation (7).

In the case of autocorrelated noise, however, an approximation is necessary (for computational reasons). In [8] it is suggested to approximate the matrix $\boldsymbol{S}_\Delta$ by a diagonal matrix $\boldsymbol{S}_\Delta^*$ with equal diagonal elements having the same determinant as $\boldsymbol{S}_\Delta$. Using this approximation the contextual adjustment factor from Equation (4) simplifies to

$$\sum_a \phi(\pi_a|\pi_\nu) P(\boldsymbol{X}_N = \boldsymbol{x}_N \,|\, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{C} = (\pi_\nu, \pi_a, \pi_b, \pi_c, \pi_d, \pi_e, \pi_f))$$
$$\sum_b \phi(\pi_b|\pi_\nu) P(\boldsymbol{X}_S = \boldsymbol{x}_S \,|\, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{C} = (\pi_\nu, \pi_a, \pi_b, \pi_c, \pi_d, \pi_e, \pi_f))$$
$$\sum_c \phi(\pi_c|\pi_\nu) P(\boldsymbol{X}_E = \boldsymbol{x}_E \,|\, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{C} = (\pi_\nu, \pi_a, \pi_b, \pi_c, \pi_d, \pi_e, \pi_f))$$
$$\sum_d \phi(\pi_d|\pi_\nu) P(\boldsymbol{X}_W = \boldsymbol{x}_W \,|\, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{C} = (\pi_\nu, \pi_a, \pi_b, \pi_c, \pi_d, \pi_e, \pi_f))$$
$$\sum_e \phi(\pi_e|\pi_\nu) P(\boldsymbol{X}_T = \boldsymbol{x}_T \,|\, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{C} = (\pi_\nu, \pi_a, \pi_b, \pi_c, \pi_d, \pi_e, \pi_f))$$
$$\sum_f \phi(\pi_f|\pi_\nu) P(\boldsymbol{X}_B = \boldsymbol{x}_B \,|\, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{C} = (\pi_\nu, \pi_a, \pi_b, \pi_c, \pi_d, \pi_e, \pi_f)),$$

where each of the sums are over all classes.

III. RESULTS

The procedures described above were tested on Monte Carlo simulated data; the results of the evaluations are discussed below.

A. Simulation

In order to illustrate the power of this algorithm we will apply it to a two class 3-D synthetic dataset. This dataset consists of a $64 \times 64 \times 64$ data volume with one variable at every pixel. The data volume is generated by use of a (morphological) isotropic Potts model [12]. In Figures 3(a) and 3(h) horizontal (x-y) slice 32 and vertical (y-z) slice 32 of the volume are shown, respectively.

The two classes are assigned mean values -1 and 1. We will consider two cases. First, the case of pure white noise, and second, the case of a mixture of white and autocorrelated noise. Furthermore, we will consider a moderat noise level of unit standard deviation as well as a high noise level of standard deviation two. In both cases we will compare the contextual classifiers with a classical pixelwise linear classifier (e.g. [10]). In addition to this we will make comparisons between the classifications using the 3-D algorithms with implementations where contextual information is drawn only from 2-D (corresponding to the algorithms in [2], [4]), as well as implementations where only 1-D context is used.

B. Classification Results

All classifications will be performed using the true parameters for mean values, variances, and autocorrelations. The transition probabilities of the WSH models and the prior distribution of the neighbourhood configurations of the OHM models are estimated by their relative occurences in the simulated data volume. Maximum likelihood methods of discriminant analysis are of course sensitive to the parameters of the fitted models. This is not considered a part of the subject for this article. We intend to return to this topic in a forthcoming paper.

B.1 Case 1: White noise

In this case we will degrade the data volume with independent, identically distributed Gaussian noise, with standard deviations 1 and 2, respectively. In Figures 3(b), 3(c), 3(i), and 3(j) degraded slices corresponding to Figures 3(a) and 3(h) are shown.

The misclassification rates for the classifications are shown in Table II. With respect to the classifiers, OHM and WSH refers to the Owen-Hjort-Mohn and the Welch-Salter-Haslett methods, a prefix capital A denotes use of an autocorrelated noise model (i.e. $\theta \neq 0$ in Equation (6)), whereas a missing capital A denotes the use of a white noise model only (i.e. $\theta = 0$). Finally, The postfix n-D indicates the size of the context considered in the algorithm. For the 3-D algorithms we use the north, east, south, west, top, and bottom neighbours as described in the previous Section, for the 2-D algorithms we employ the north, east, south, and west neighbours (as described in the original 2-D algorithms), and in the 1-D case we use the east and west neighbours only.

For the non-contextual classifier the classification rule should be a theshhold at 0, which for the two values of the standard deviation, σ, corresponds to $1 \cdot \sigma$ and $0.5 \cdot \sigma$. Assuming normality, this should result in misclassification rates of 15.866% and 30.854%, respectively. The obtained results agree well with this. When compared with the contextual OHM 3-D classifier, we see that the inclusion of spatial information results in a misclassification rate that for $\sigma = 1$ is a factor 15 lower and for $\sigma = 2$ is a factor 3 lower. For the WSH 3-D model the misclassication rates are also better, though not as good as for the OHM model. It is noteworthy that whereas the OHM models increase their performance as more spatial dimensions are included, the misclassification rate does not decrease for the WSH model when going from 2-D to 3-D. Also, where OHM 2-D and WSH 2-D performs equally well, the OHM model is superior in the 3-D case.

Apart from the contextual methods performing significantly better in terms of misclassifications rates the original patterns are clearly discernible when comparing with the non-contextual methods, as is shown in Figures 4 and 5. It should also be noted that the errors tend to occur on the edges, and that the errors also tend to lump together in the directions where contextual information is included (i.e. for the 1-D algorithms the errors frequently occur in east-west line segments, whereas in the other directions they seem to occur more randomly).

B.2 Case 2: Autocorrelated and white noise

In this case we will degrade the data volume with independent, identically distributed Gaussian noise and with autocorrelated Gaussian noise. The white noise and the autocorrelated noise are independent and have equal variances. We will use autocorrelated noise with an autocorrelation that decays exponentially with Euclidean distance. We will apply two cases of the autocorrelation, namely the cases of autocorrelation in lag 1 being 0.4 and 0.6. Again we will apply the algorithms to two cases with pixelwise standard deviations 1 and 2, respectively. In Figures 3(d)-3(g) and 3(k)-3(n) the degraded slices corresponding to Figures 3(a) and 3(h) are shown.

Again we see in Table II that misclassification rates for the non-contextual classifier are close to what we would expect. With respect to the contextual methods we see the same pattern as for the white noise only situation: The 2-D algorithms works equally well, whereas the extension to 3-D increase the performance only for the OHM model. The lowest misclassification rates are obtained for the (A)OHM 3-D classifiers. It should also be noted that for the WSH models the inclusion of spatial autocorrelation in the noise model does not have an effect. For the OHM models the effect of including spatial autocorrelation in the noise model is hardly discernible. Examples of the classification results on the slices shown in Figure 3 are shown in Figures 6, 7, 8, and 9.

With respect to the autocorrelated noise in the images it is clear that the relative improvement of including the context in the classification is less. The neighbours hold less extra information as is also noted in [4].

Finally is should be noted that for the high noise level and high noise autocorrelation as can be seen in Figure 9 the contextual algorithms although performing better than the non contextual method in terms of misclassification rates break down in the sense that the original patterns are hard if not impossible to discern.

IV. Conclusion

We have described extensions of 2-D contextual classification algorithms by Owen, Hjort & Mohn (OHM) and Welch, Salter & Haslett (WSH) based on the simulateneous distribution of a pixel and its nearest neighbours to the 3-D case. The algorithms include contextual information for each pixel by including the feature vector of that pixel as well as the feature vectors of the 6 nearest neighbouring pixels in the decision. A joint Gaussian distribution for these feature vectors given the classes of the pixels has been specified. It is assumed that the noise can be modelled as a sum of white noise and autocorrelated noise, where the autocorrelation function is exponentially decaying with (Euclidean) distance. Furthermore, joint prior distributions of the class variables of a pixel and its 6 nearest neighbours have been specified. In the OHM case it is assumed that the pixel size is small relative to the region sizes in the image, thus vastly decreasing the number of possible configurations to in principle four types. Whereas in the WSH case we assume independence of the class variables of the neighbours given the center pixel class.

The algorithm is tested on a synthetic two-class 3-D image. For moderate white noise levels the misclassification rate is a factor 15 lower for the OHM 3-D algorithm than the rate obtained using an ordinary linear pixelwise classifier. The relative improvement in misclassification rate decreases with increasing noise level. For the WSH algorithms the extension to a 3-D context from 2-D does not decrease the misclassification rates. In the case of a mixture of white and autocorrelated noise the improvement in misclassification rate over the pixelwise method is a factor 4 for moderate noise levels for the OHM 3-D model. In this case also the inclusion of the extra spatial dimension from 2-D to 3-D does not decrease the misclassification rate for the WSH models. Figures 4 and 6 give a good visual indication of the power of the algorithm. The non-contextual classifier gives very noisy (speckled) classification results, where the contextual methods and in particular the OHM 3-D algorithm gives well defined patterns.

V. Acknowledgements

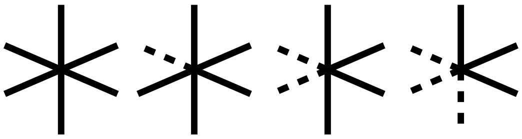
The software used to generate the data set was programmed by M. Sc. Jørgen Folm Hansen, Department of Mathematical Modelling, Technical University of Denmark.

References

[1] Art Owen, "A neighbourhood-based classifier for LANDSAT data," *The Canadian Journal of Statistics*, vol. 12, pp. 191–200, 1984.

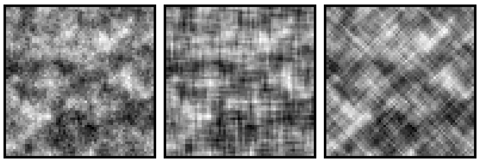
[2] Nils Lid Hjort and Erik Mohn, "A comparison of some contextual methods in remote sensing classification," in *The 18th International Symposium on Remote Sensing of Environment*, Paris, France, Oct. 1984.

[3] John R. Welch and Kenneth G. Salter, "A context algorithm for pattern recognition and image interpretation," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 1, pp. 24–30, 1971.
[4] John Haslett, "Maximum likelihood discriminant analysis on the plane using a markovian model of spatial context," *Pattern Recognition*, vol. 18, no. 3, pp. 287–296, 1985.
[5] Paul Switzer, "Extension of discriminant analysis for statistical classification of remotely sensed satellite imagery," *Journal of the International Association for Mathematical Geology*, vol. 12, pp. 367–376, 1980.
[6] Stuart Geman and Donald Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.
[7] Julian Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society, Series B*, vol. 48, no. 3, pp. 259–302, 1986.
[8] Nils Lid Hjort, Erik Mohn, and Geir Storvik, "Contextual classification of remotely sensed data, based on an auto-correlated model," in *Contextual classification of remotely sensed data: Statistical methods and development of a system*, H. V. Sæbø, K. Bråten, Nils Lid Hjort, B. Llewellyn, and Erik Mohn, Eds. Norwegian Computing Center, 1985, Technical report No. 768.
[9] Noel E. Cressie, *Statistics for Spatial Data*, John Wiley & Sons, New York, second edition, 1993, 900 pp.
[10] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, New York, second edition, 1984, 675 pp.
[11] Paul Switzer, "A random set process in the plane with a Markovian property," *Annals of Mathematical Statistics*, vol. 36, pp. 1859–1863, 1965.
[12] Jens Michael Carstensen, "Morphological Markov random fields," *Statistics and probability letters*, vol. 20, no. 4, pp. 321–326, 1994.



(a) CR_1 (b) CR_2 (c) CR_3 (d) CR_4

Fig. 1. Patterns in the model. Within the 'cross', that represents the neighbourhood of a pixel, i.e. the six nearest neighbours, it it is assumed that at most two classes are present, and that the only possible configurations are these four types of 'crosses'.



(a) 2-norm (b) 1-norm (c) ∞-norm

Fig. 2. 2-D Noise patterns corresponding to autocorrelation functions using (a) the 2-norm (Euclidean), (b) the 1-norm (Manhattan), and (c) the ∞-norm. All three realization have an autocorrelation of 0.8 in for first-order neighbours.

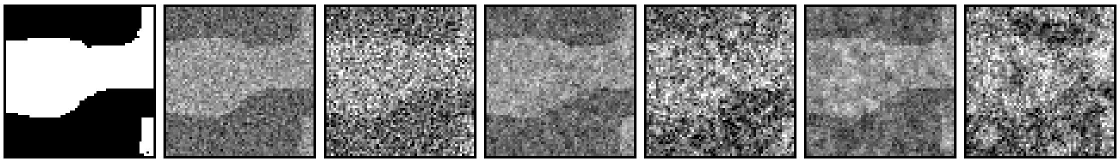
TABLE I

AUTOCORRELATIONS BETWEEN FIRST-ORDER (α), SECOND-ORDER (β), AND THIRD-ORDER (γ) NEIGHBOUR CORRESPONDING TO THREE DIFFERENT NORMS USED IN THE DEFINITION OF THE AUTOCORRELATION MODEL.

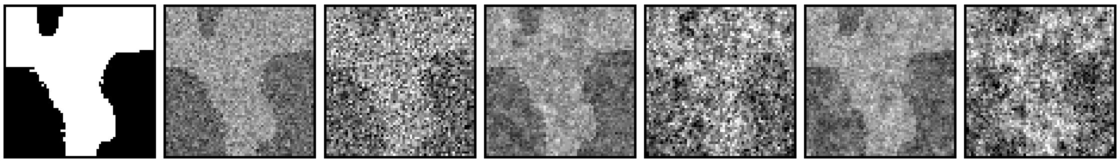
	α	β	γ
∞-norm	$\rho\theta$	$\rho\theta$	$\rho^2\theta$
2-norm	$\rho\theta$	$\rho^{\sqrt{2}}\theta$	$\rho^2\theta$
1-norm	$\rho\theta$	$\rho^2\theta$	$\rho^2\theta$

TABLE II
MISCLASSIFICATION RATES FOR EACH OF THE COMBINATIONS BETWEEN CLASSIFIER AND NOISE LEVEL.

	White noise		Autocorrelated noise $\rho = 0.4$		Autocorrelated noise $\rho = 0.6$	
	$\sigma = 1$	$\sigma = 2$	$\sigma = 1$	$\sigma = 2$	$\sigma = 1$	$\sigma = 2$
Non-context	15.8	30.9	15.8	30.8	16.0	30.9
AOHM 3-D	-	-	3.6	17.9	5.9	21.2
OHM 3-D	1.1	10.6	3.7	18.0	6.0	21.3
AWSH 3-D	-	-	5.0	20.1	7.0	22.4
WSH 3-D	2.2	14.3	4.9	20.0	7.0	22.5
AOHM 2-D	-	-	4.9	19.9	7.0	22.5
OHM 2-D	2.2	14.4	4.9	20.0	7.0	22.6
AWSH 2-D	-	-	4.9	19.9	7.1	22.6
WSH 2-D	2.2	14.3	4.9	20.0	7.1	22.6
AOHM 1-D	-	-	7.5	23.2	9.1	24.8
OHM 1-D	5.1	20.1	7.4	23.1	9.0	24.9
AWSH 1-D	-	-	7.3	23.1	8.9	24.8
WSH 1-D	4.9	20.0	7.3	23.1	9.0	24.8



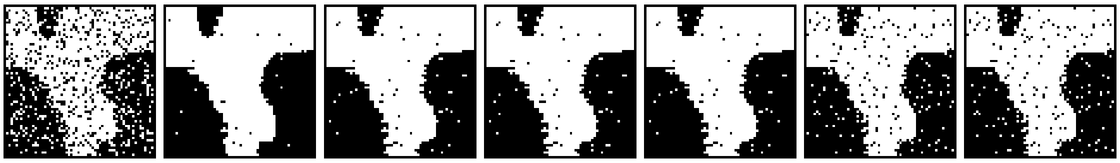
(a) Orig. (b) $\sigma = 1$ (c) $\sigma = 2$ (d) $\sigma = 1$ (e) $\sigma = 2$ (f) $\sigma = 1$ (g) $\sigma = 2$



(h) Orig. (i) $\sigma = 1$ (j) $\sigma = 2$ (k) $\sigma = 1$ (l) $\sigma = 2$ (m) $\sigma = 1$ (n) $\sigma = 2$

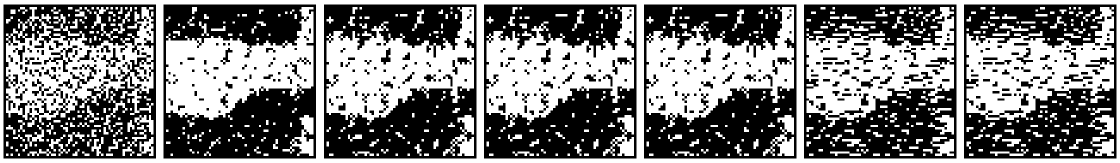
Fig. 3. Horizontal (x-y) slice 32 (top) and vertical (y-z) slice 32 (bottom) of the original data volume, and the six degraded sequences.

(a) Non-ctx	(b) 3-D OHM	(c) 3-D WSH	(d) 2-D OHM	(e) 2-D WSH	(f) 1-D OHM	(g) 1-D WSH



(h) Non-ctx	(i) 3-D OHM	(j) 3-D WSH	(k) 2-D OHM	(l) 2-D WSH	(m) 1-D OHM	(n) 1-D WSH

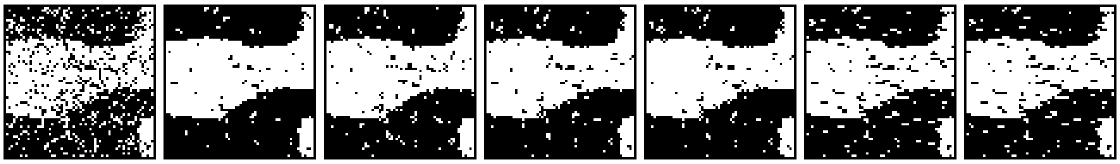
Fig. 4. Horizontal (x-y) slice 32 (top) and vertical (y-z) slice 32 (bottom) of the classifed volumes using non-contextual, 3-D OHM, 3-D WSH, 2-D OHM, 2-D WSH, 1-D OHM, and 1-D WSH in the case of $\sigma = 1$.



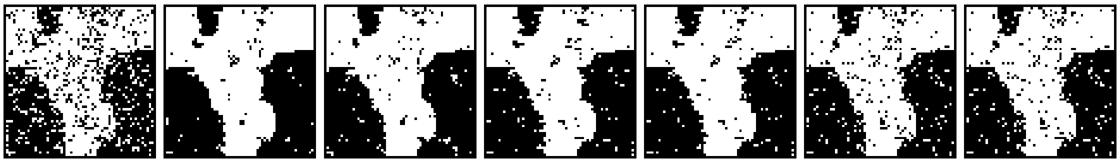
(a) Non-ctx	(b) 3-D OHM	(c) 3-D WSH	(d) 2-D OHM	(e) 2-D WSH	(f) 1-D OHM	(g) 1-D WSH

(h) Non-ctx	(i) 3-D OHM	(j) 3-D WSH	(k) 2-D OHM	(l) 2-D WSH	(m) 1-D OHM	(n) 1-D WSH

Fig. 5. Horizontal (x-y) slice 32 (top) and vertical (y-z) slice 32 (bottom) of the classifed volumes using non-contextual, 3-D OHM, 3-D WSH, 2-D OHM, 2-D WSH, 1-D OHM, and 1-D WSH in the case of $\sigma = 2$.

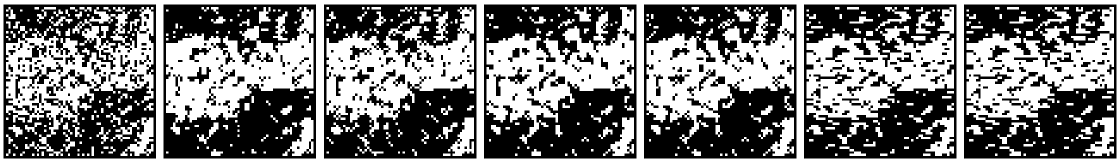


(a) Non-ctx (b) 3-D OHM (c) 3-D WSH (d) 2-D OHM (e) 2-D WSH (f) 1-D OHM (g) 1-D WSH



(h) Non-ctx (i) 3-D OHM (j) 3-D WSH (k) 2-D OHM (l) 2-D WSH (m) 1-D OHM (n) 1-D WSH

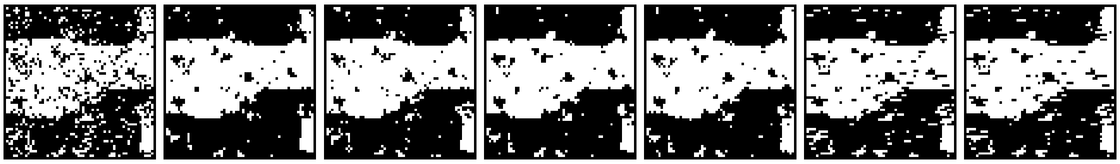
Fig. 6. Horizontal (x-y) slice 32 (top) and vertical (y-z) slice 32 (bottom) of the classifed volumes using non-contextual, 3-D OHM, 3-D WSH, 2-D OHM, 2-D WSH, 1-D OHM, and 1-D WSH in the case of $\sigma = 1$ and autocorrelated noise with $\rho = 0.4$.



(a) Non-ctx (b) 3-D OHM (c) 3-D WSH (d) 2-D OHM (e) 2-D WSH (f) 1-D OHM (g) 1-D WSH

(h) Non-ctx (i) 3-D OHM (j) 3-D WSH (k) 2-D OHM (l) 2-D WSH (m) 1-D OHM (n) 1-D WSH

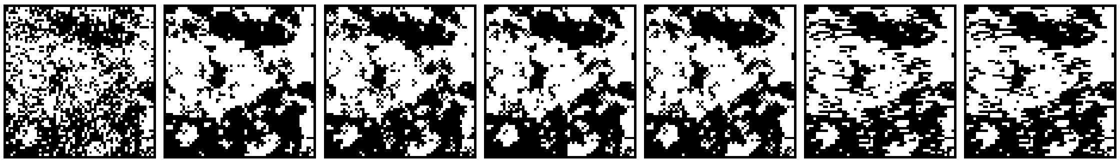
Fig. 7. Horizontal (x-y) slice 32 (top) and vertical (y-z) slice 32 (bottom) of the classifed volumes using non-contextual, 3-D OHM, 3-D WSH, 2-D OHM, 2-D WSH, 1-D OHM, and 1-D WSH in the case of $\sigma = 2$ and autocorrelated noise with $\rho = 0.4$.



(a) Non-ctx (b) 3-D OHM (c) 3-D WSH (d) 2-D OHM (e) 2-D WSH (f) 1-D OHM (g) 1-D WSH

(h) Non-ctx (i) 3-D OHM (j) 3-D WSH (k) 2-D OHM (l) 2-D WSH (m) 1-D OHM (n) 1-D WSH

Fig. 8. Horizontal (x-y) slice 32 (top) and vertical (y-z) slice 32 (bottom) of the classifed volumes using non-contextual, 3-D OHM, 3-D WSH, 2-D OHM, 2-D WSH, 1-D OHM, and 1-D WSH in the case of $\sigma = 1$ and autocorrelated noise with $\rho = 0.6$.



(a) Non-ctx (b) 3-D OHM (c) 3-D WSH (d) 2-D OHM (e) 2-D WSH (f) 1-D OHM (g) 1-D WSH

(h) Non-ctx (i) 3-D OHM (j) 3-D WSH (k) 2-D OHM (l) 2-D WSH (m) 1-D OHM (n) 1-D WSH

Fig. 9. Horizontal (x-y) slice 32 (top) and vertical (y-z) slice 32 (bottom) of the classifed volumes using non-contextual, 3-D OHM, 3-D WSH, 2-D OHM, 2-D WSH, 1-D OHM, and 1-D WSH in the case of $\sigma = 2$ and autocorrelated noise with $\rho = 0.6$.