New Nonlinear Features for Inspection, Robotics, and Face Recognition

David Casasent and Ashit Talukder

Department of Electrical and Computer Engineering, Laboratory for Optical Data Processing
Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

Classification of real-time X-ray images of randomly oriented touching pistachio nuts is discussed. The ultimate objective is the development of a system for automated non-invasive detection of defective product items on a conveyor belt. We discuss the extraction of new features that allow better discrimination between damaged and clean items (pistachio nuts). This feature extraction and classification stage is the new aspect of this paper; our new maximum representation and discriminating feature (MRDF) extraction method computes nonlinear features that are used as inputs to a new modified k nearest neighbor classifier. In this work, the MRDF is applied to standard features (rather than iconic data). The MRDF is robust to various probability distributions of the input class and is shown to provide good classification and new ROC (receiver operating characteristic) data. Other applications of these new feature spaces in robotics and face recognition are also noted.

Key Words: Classification, discrimination, detection, feature extraction, k nearest neighbor classifier (modified), nonlinear features, product inspection, X-ray sensors

1 INTRODUCTION

We emphasize a product inspection application. The internal product detail that X-ray images provide (Fig. 1) allows the presence of worm damage and other defects to be determined by non-destructive(non-invasive) methods in various agricultural products[1, 2, 3]. Current standard inspection techniques cannot determine many defects that X-ray images can, such as worms, insect damage, frass, etc. that are typically not visible through visual inspection. Prior work[2] has confirmed that X-ray images can be useful for identification of infested versus good pistachio nuts. Ready-for-sale nuts such as those used here, contain 3% of insect or feeding damage, while USDA standards require 1-3% of such damage[4]. A 50% reduction in damage, while restricting rejection of good nuts to a commercially acceptable 1% or less, would be desirable for quality and aflatoxin reduction. An automated technique is necessary to identify clean and infested pistachio nuts. In this paper, a new feature extraction method called the maximum representation and discriminating feature (MRDF)[5] extraction technique is used to extract nonlinear features useful for discrimination between bad and good items. The MRDF uses higher-order correlation information that linear features do not consider; we show that such higher-order correlation information is useful for distinguishing between good and bad products in our database. We earlier compared the performance of our feature extraction procedure with the orthonormal discriminant vector (ODV) procedure[6, 7] (and the Fisher linear discriminant [8] which is a subset of the ODV).

The classification step requires each nut to be isolated from others. It is possible to have several pistachio nuts that are touching each other in the input image. Therefore, prior to classification, preprocessing must be carried out to isolate touching pistachio nuts. Prior work on pistachio nut classification[2] did not address the problem of segmentation of touching nuts. The image preprocessing we use to segment touching pistachio nuts is discussed elsewhere [9]. The MRDF method to extract features that are used as inputs to the classifier is discussed in Sect. 2. Classification results are discussed in Sect. 3. We note (Sect. 2) the drawbacks associated with the Fisher linear discriminant and the ODV, which is an extension of the Fisher linear discriminant technique to multiple discriminant vectors.

The database we consider consisted of scanned 2-D X-ray film images of large pistachio nuts (18-20 nuts per ounce). The nuts used were obtained from a processor after sorting and sizing processing and hand-inspection. The X-ray film images used were obtained with a $(0.5mm)^2$/pixel resolution [4]. Fig. 1 shows typical X-ray images of clean and infested pistachio nuts. As seen, such X-ray images contain useful internal information about the quality of the nut. Infested pistachio nut images (Figs. 1e-h) tend to have large gray scale variations (due to the presence of worm tunnels, infestations, etc.) along with large airgaps (outer dark region between nutmeat and shell), whereas clean nuts typically have smooth nutmeat regions with smaller airgaps (Figs. 1a-d).

2

Part of the SPIE Conference on Algorithms, Devices, and Systems for Optical
Information Processing III ● Denver, Colorado ● July 1999
SPIE Vol. 3804 ● 0277-786X/99/$10.00

2 FEATURES FOR DISCRIMINATION: MRDF

The MRDF[5] was originally devised to extract nonlinear features for joint representation and discrimination. It has been proven to be very useful in active vision applications where classification and pose estimation of unknown objects (machine parts) are necessary, and has been shown to outperform standard linear feature extraction techniques such as the Fisher discriminant, Fukunaga-Koontz (FK) features, and Karhunen-Loeve (KL) features[10]. The Fisher linear discriminant and standard discriminant analysis[11] yields fewer discriminant feature vectors than the number of classes (e.g., only 1 feature vector when two classes are present). In addition, the discriminant vectors are not guaranteed to be orthogonal when the number of classes are more than two (they are orthogonal *only* if the inverses of the within-class covariance matrix $\underline{W}^{-1}$ and the between class covariance matrix $\underline{B}$ have the same eigenvectors). Orthonormal discriminant vector (ODV) techniques have been suggested that use the Gram-Schmidt orthogonalization procedure to iteratively yield orthonormal vectors that maximize the Fisher criterion at each iteration[6, 7]. We show (Sect. ??), that ODV features are not as robust as the ones used in the MRDF. A nonlinear MRDF method has also been developed[5] that provides *nonlinear transformations that are obtained in closed form* and hence offer definite advantages compared to other iterative nonlinear techniques such as the nonlinear principal component analysis (PCA)[12] and neural network techniques[13] that have convergence problems, and that require ad-hoc parameter selection. It has been shown[5] that our nonlinear MRDF transformation provide quadratic transformations of full rank in contrast with other nonlinear iterative approaches. We first discuss the linear MRDF for discrimination (Sect. 2.1), and then detail the nonlinear MRDF for discrimination (Sect. 2.2). We show that the higher-order correlation information in the data that is used by the nonlinear MRDF gives improved discrimination (Sect. 3.3.3) on our present database. *For product inspection, we apply our methods to standard feature inputs rather than iconic inputs.* In the present application, *only MRDF features for discrimination are considered.* In other applications (Sects. 4 and 5), we apply our methods to iconic data.

We use the following notation. Vectors and matrices are represented by lower and uppercase letters with underlines ($\underline{x}$ and $\underline{X}$). Classes are represented in terms of sample vectors with the mean and covariance matrices of each class corresponding to the sample mean and sample covariance respectively. For example, when considering a set of S sample data vectors $\{\underline{x}_s\}$, we describe them by the sample data matrix $\underline{X} = [\underline{x}_1 \ \underline{x}_2 \ \cdots \ \underline{x}_S]$.

2.1 Linear MRDF for Discrimination

We wish to determine the $\underline{\Phi}_M$ for an M-dimensional *linear* transform, $\underline{y} = \underline{\Phi}_M^T \underline{x}$, where $\underline{\Phi}_M$ is a matrix containing M orthogonal basis function vectors $\underline{\phi}_m$, such that the input data $\underline{x}$ is transformed into lower-dimensional (M-dimensional) feature space vectors $\underline{y}$. The orthogonality constraint on the transformation vectors $\underline{\phi}_m$ forces the output features in $\underline{y}$ to be uncorrelated. To simplify explanation, we consider two classes, where $\underline{x}_{1s}$ and $\underline{x}_{2q}$ are representative samples from classes 1 and 2. Class 1 has S samples $\{\underline{x}_{1s}\}, s = 1, \ldots S$ and a covariance matrix $\underline{C}_1 = (1/S)(\underline{X}_1 \underline{X}_1^T) - \underline{\mu}_1 \underline{\mu}_1^T$, where the S samples are the columns of the data matrix $\underline{X}_1 = [\underline{x}_{11} \ \underline{x}_{12} \ \cdots \ \underline{x}_{1S}]$ and the sample mean is $\underline{\mu}_1 = (1/S)\sum_{s=1}^{S} \underline{x}_{1s}$. The sample mean and covariance matrix for class 2, $\underline{\mu}_2$ and $\underline{C}_2$, are computed in a similar manner.

We denote the projection of a class 1 vector on the basis vector $\underline{\phi}_m$ by $y_{m1s} = \underline{\phi}_m^T \underline{x}_{1s}$; similarly, the projection of a class 2 vector is $y_{m2q} = \underline{\phi}_m^T \underline{x}_{2q}$. For the projections of the two classes to be best separated, we desire that $[(y_{m1s} - y_{m2q})^2]$ be large (or maximized) for all samples s and q and for all basis vectors $\underline{\phi}_m, 1 \leq m \leq M$, i.e. we maximize the mean squared separation between all class 1 and 2 projections on each basis vector $\underline{\phi}_m$. We also minimize the spread of the transformed samples of each class, i.e. we minimize $\underline{\phi}_m^T(\underline{C}_1 + \underline{C}_2)\underline{\phi}_m$ for all $\underline{\phi}_m$. Therefore, to determine the best set of transformation coefficients $\underline{\phi}_m$ that separate (discriminate) classes 1 and 2, we maximize

$$E_D = \sum_{m=1}^{M} E_{Dm} = \sum_{m=1}^{M} \frac{\underline{\phi}_m^T \underline{R}_{12} \underline{\phi}_m}{\underline{\phi}_m^T (\underline{C}_1 + \underline{C}_2)\underline{\phi}_m}, \tag{1}$$

where $\underline{R}_{12} = (1/SQ) \sum_{s=1}^{S} \sum_{q=1}^{Q} (\underline{x}_{1s} - \underline{x}_{2q})(\underline{x}_{1s} - \underline{x}_{2q})^T$. For the multi-class case (L classes), the separation measure to be maximized is

$$E_D = \sum_{m=1}^{M} [\underline{\phi}_m^T (\sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \underline{R}_{ij})\underline{\phi}_m] / [\underline{\phi}_m^T (\sum_{k=1}^{L} \underline{C}_k)\underline{\phi}_m]. \tag{2}$$

3

To obtain the best linear transformation coefficients $\underline{\phi}_m$ that best separate the two classes, we set the derivative of Eq. (1) with respect to the basis function $\underline{\phi}_m$ to zero and obtain

$$(\underline{C}_1 + \underline{C}_2)^{-1}(\underline{R}_{12})\underline{\phi}_m = \lambda_m \underline{\phi}_m \text{ for all } m = 1, 2..., M.$$ (3)

The $\underline{\phi}_m$ solutions to Eq. (3) are the solutions to an eigenvalue-eigenvector equation. The MRDF basis functions $\underline{\phi}_m$ that best separate the two classes are the M eigenvectors corresponding to the M largest eigenvalues λ_m of $(\underline{C}_1 + \underline{C}_2)^{-1}(\underline{R}_{12})$.

Thus, while the Fisher linear discriminant [8] and its variations (discriminant analysis[11] and the ODV[7]) considers the squared separation between the *mean* of each class and the most discriminating feature (MDF)[14] technique considers the squared separation between each class *mean* and the global *mean*, the MRDF looks at the average squared separation between *all* samples in the classes. *In the case of classes with multiple clusters, the MRDF is expected to be a more robust discrimination feature* since it considers the separation between each sample pair in two classes (not just the means). *This technique is very new since all classes are considered simultaneously (pairwise),* as compared to the two-class (macro-class) approaches (MDF, FK and Fisher linear discriminant). We have shown[5] that the MRDF is the Fisher linear discriminant (when the classes are Gaussian distributed).

2.2 Nonlinear Features for Discrimination: Nonlinear MRDF

The linear MRDF (Sect. 2.1) can only provide linear transformations. We now develop a nonlinear MRDF that provides *nonlinear transformations for discrimination* and gives a *closed-form solution.* These features use higher-order correlation information in the input data and thus is useful for handling asymmetrically distributed data. All prior nonlinear PCA etc. methods are iterative and thus do not have closed-form solutions. Our nonlinear feature extraction procedure is shown to provide advantages compared to prior nonlinear PCA methods. Nonlinear transformations are shown to be better at separating classes with a large amount of overlap (Sect. 3.3.3).

For simplicity of notation, we will refer to a sample in a class as a vector $\underline{x}$ without subscripts. Our objective is to find a nonlinear transformation (or a set of nonlinear transforms) on $\underline{x}$ that optimally discriminates between classes in a reduced dimensionality space. A general nonlinear transform on a vector $\underline{x}$ is $y = f(\underline{x})$. We consider nonlinear transforms that are polynomial mappings of the input. In this initial work, we only consider a second-order polynomial mapping (quadratic transform). The quadratic transform can be written as $y = \sum_n a_n x_n + \sum_n \sum_m a_{mn} x_m x_n$ or in matrix form as $y = \underline{x}^T \underline{A}\underline{x} + \underline{b}^T \underline{x}$ where $\underline{A}$ is a matrix and $\underline{b}$ is a vector. Alternatively, we can write a quadratic mapping as a linear transform $y = \underline{\phi}^T \underline{x}_H$ on the higher-order and higher-dimensional vector

$$\underline{x}_H = [x_1 \ x_2 \ ... \ x_N \ x_1 x_1 \ x_1 x_2 \ .... \ x_N x_N]^T, \text{ where } H = N + N(N+1)/2 \text{ is the dimension of } \underline{x}_H.$$ (4)

$\underline{x}_H$ contains higher-order terms in the original input data $\underline{x}$. This higher-dimensional vector has N linear terms, and $N(N+1)/2$ unique non-linear cross-product terms. This formulation of a quadratic mapping *allows for a closed form solution in the $\underline{x}_H$ space that is quadratic in the original $\underline{x}$ space.* Note that this can be generalized easily to yield polynomial mappings of any arbitrary order.

To construct a second-order nonlinear MRDF, we consider two classes of data described by the sample vectors $\{\underline{x}_{1s}\}$ and $\{\underline{x}_{2q}\}$. We create the augmented data vectors, $\{\underline{x}_{1sH}\}$ and $\{\underline{x}_{2qH}\}$ for each class; these contain first and second-order terms as in Eq. (4). We desire to determine the M best transforms $\underline{\Phi}_M = [\underline{\phi}_1 \underline{\phi}_2 ... \underline{\phi}_M]$ such that the transformed data for each sample from the two classes $\underline{y}_{1s} = \underline{\Phi}_M^T \underline{x}_{1sH}$ and $\underline{y}_{2q} = \underline{\Phi}_M^T \underline{x}_{2qH}$ are separated. The $\underline{\phi}_m$ are constrained to be orthogonal, implying that the output features y_m in the output feature vector $\underline{y} = [y_1 \ ...y_m \ ...y_M]^T$ are uncorrelated, i.e. $E(y_m y_n) = 0$ for $m \neq n$. The separation measure to be maximized is

$$E_D = \sum_{m=1}^{M} E_{Dm} = \sum_{m=1}^{M} \frac{\underline{\phi}_m^T \underline{R}_{12H} \underline{\phi}_m}{\underline{\phi}_m^T (\underline{C}_{1H} + \underline{C}_{2H})\underline{\phi}_m}.$$ (5)

This is the associated extension of the linear MRDF in Eq. (1), where $\underline{R}_{12H} = \sum_{q=1}^{Q} \sum_{s=1}^{S} (\underline{x}_{1sH} - \underline{x}_{2qH})(\underline{x}_{1sH} - \underline{x}_{2qH})^T]$, and $\underline{C}_{1H}$ and $\underline{C}_{2H}$ are computed from the augmented input data vectors.

To obtain the set of best nonlinear transformations that separate the two classes, we differentiate Eq. (5) with respect to $\underline{\Phi}_m$. The $\underline{\phi}_m$ solution must satisfy

$$(\underline{C}_{1H} + \underline{C}_{2H})^{-1}(\underline{R}_{12H})\underline{\Phi}_M = \underline{\Phi}_M \underline{\Lambda}.$$ (6)

This corresponds to an eigenvalue-eigenvector equation as before. Thus, the M best nonlinear MRDF basis functions $\underline{\phi}_m$ are the M dominant eigenvectors corresponding to the M largest eigenvalues of $(\underline{C}_{1H} + \underline{C}_{2H})^{-1}(\underline{R}_{12H})$. Note that *the nonlinear transformation coefficients are obtained in closed-form, in contrast to other nonlinear methods.*

3   CLASSIFICATION (PRODUCT INSPECTION)

The pistachio nuts can lie at any orientation. Hence, we first compute rotation and scale-invariant histogram features (Sect. 3.1) from each individual pistachio nut. These histogram features are not all equally useful for discriminating between clean and infested pistachio nuts. Hence, it is necessary to order these features by their importance for discrimination and select a subset of these features as inputs to the classifier (classifiers themselves *cannot* find the best combination of features to use). However, selection of an optimal subset of features is an NP-complete problem[15]. We compute our MRDF features from *all* original input histogram features; standard classifiers do not perform well using all input histogram features. *The MRDF discrimination technique allows use of* (and finds the useful discrimination information present in) *all original histogram features.* Therefore, an advantage of our MRDF features is that they *automatically* compute a linear (or nonlinear) combination of the input histogram features such that the clean and infested pistachios are separated well in MRDF feature space, and are ordered *automatically* in terms of their importance for discrimination. Therefore, a classifier operating on such features performs better. In addition, the nonlinear MRDFs extract higher-order correlation from the input histogram features; such higher-order correlation information is ignored by some classifers, such as the quadratic Gaussian classifier used in the statistical analysis software (SAS)[16], linear discriminants such as the Fukunaga Koontz and Fisher linear discriminant, and linear neural networks.

We first discuss the original histogram features extracted from each image (Sect. 3.1). The new modified k nearest neighbor (k-NN) classifier that we use was described earlier [17]. We compare (Sect. 3.3) classification results using a subset of the input histogram features (selected using forward-selection in SAS) and using our linear and nonlinear MRDF features. A block diagram of the system using quadratic MRDFs is shown in Fig. 2. We show (Table 3) that MRDFs computed from *all* input histogram features provide better classification than use of the best subset of histogram features; this shows that *all histogram features contain useful discriminatory information and that the MRDF can effectively select such information, while other approaches cannot.* Our MRDF features have also been shown elsewhere [17] to provide superior discrimination compared with the ODV.

3.1   Rotation and Scale Invariant Features (Original Histogram Features)

This section describes the original features we use. We apply our MRDF algorithm to this feature space to produce improved linear and nonlinear features from it. The size of pistachio nuts vary considerably and the nuts can be at any orientation. Therefore, features that are *both rotation and scale-invariant need to be extracted from each pistachio nut.* The features considered were the mean, variance, skewness, and kurtosis (histogram features) of four different images of each nut (raw, edge, curvature[18] of raw and curvature of edge images)[19]. This gave a total of 16 possible features. The histogram for each image of each nut was divided by the total number of pixels in each nut image (this provides scale-invariant features). The raw images for infested pistachio nuts tend to be darker with more gray-scale variations, whereas clean pistachios tend to have smoother gray values. Hence, we expect the variance of the raw images to be different for infested pistachio nuts. Edge (high-frequency) information is also expected to be useful since infested pistachios tend to have rougher texture compared to clean nuts. The curvature images provide information about the rate of change of gray-values (combination of first and second order differentials in gray-scale) over the entire gray-scale pistachio image. This information is very useful, especially in regions near the airgaps. For the curvature (of raw and edge-enhanced) and edge-enhanced images, we erode the outputs with a 3×3 structuring element prior to using them; this removes the outer boundary between the shell and the background which tend to have very high values due to the gray-scale change in such regions. For only the curvature images, the output images were clipped at $\pm T_C$ (all values $\geq T_C$ are set to T_C and all values $\leq -T_C$ were set to $-T_C$; T_C=1.5 was used) and each clipped curvature image was separately normalized to 0-255. This separate normalization reduces the usefulness of the mean of a nut in curvature data but enhances the use of variance, skewness and kurtosis, since good nuts should not have many concave regions (worm tunnel etc.) while bad nuts should have both concave and convex regions. For each of the resultant four sets of images of the segmented nuts, 4 sets of histogram features were calculated. Skewness is a measure of the symmetry of the distribution; kurtosis is a measure of how sharply peaked the distribution is.

3.2 Database used

All of the nuts were dissected and classified into good and insect damaged (bad) by visual inspection after dissection[4]. Nuts in other categories (immature kernel, large kernel spots, etc) were not included. A total of 1884 nuts were used for classification; these were divided into a training set of 942 nuts (600 good and 342 bad) and a test set of 942 nuts (600 good and 342 bad); *only the training set was used to select the classifier parameters (such as selecting the features to use, the parameters to normalize each feature before classification, and the number of features to use as inputs to the classifier)*. To obtain classification accuracy using only the training set data, we used a leave-one out testing classification scheme on the training data.

3.3 Classification Results

In *all* cases, the *modified k-NN classifier* with k = 17 was used.

3.3.1 Classification Using Original Histogram Features

The 16 original histogram features were analyzed using SAS; forward-selection was used to order the histogram features and determine the *best histogram features in order to use for classification*. Each feature was normalized to a zero to one range (using only training set data). Test set data was normalized using the training set parameters. The 7 best histogram features selected by SAS were (in order of importance): Variance of curvature of raw images, Kurtosis of raw images, Variance of curvature of edge images, Variance of raw images, Variance of edge images, Skewness of curvature of raw images, and the Mean of curvature of raw images. Table 1 shows P_C (percentage of correct classification) obtained with different numbers of the best histogram features (ordered by SAS). To *determine the number of features to use*, we calculated P_C for the training set for an increasing number of features (with the order determined by SAS). As the number of histogram features used was increased, the percentage of nuts correctly classified P_C generally increased and rapidly became fairly constant (Table 1). From training set data, we chose 7 histogram features (Table 1). P_C results for the test set follow training set data quite well in Table 1. Thus, generalization is good. *We found this technique to select the number of features to use from training set data to be quite useful.*

Classifier	No. of Features	1	2	3	4	5	6	7	16
Modified k-NN (k=17)	P_C (Train)%	82.3	87.2	87.2	87.7	87.8	87.9	**88.0**	87.6
	P_C (Test)%	80.1	86.9	86.6	87.7	87.7	88.4	**88.7**	87.4

Table 1: P_C for different numbers of histogram features (modified k-NN).

3.3.2 Classification using Linear MRDF Features

We next used linear MRDF features that were computed from subsets of the input histogram features. The advantage of using the MRDFs is that the features are ordered automatically by their "usefulness" for classification; hence no feature subset selection is required for the MRDF features. One issue is the number of useful subset of histogram features to pick as inputs for the MRDF. We therefore used the 5 and 7 best histogram features, and all 16 histogram features as inputs to the MRDF. The classification performance using MRDFs computed from the 5 and 7 best histogram features (Table 2 did not yield a large improvement in P_C compared with classification using the best histogram features directly. Note that the *maximum number of MRDF features that can be computed is equal to the number of input histogram features*. If all the MRDF are used, the P_C obtained is the same as the P_C using all the input histogram features; this occurs because the MRDF transformation is orthonormal and using all MRDF features merely results in a rotation of the input histogram feature data. The best overall P_C obtained from training data is shown as bold text (Table 2). The P_C performance improved when MRDFs computed from *all* 16 histogram features were used for classification. This shows that *all 16 input histogram features contain useful discriminatory information, and that our MRDFs can select such information from the input data*. The number of MRDFs to use was determined from the training data; P_C performance for the test set using the same number of MRDFs was similar (bold entries in (Table 2). Thus, the linear MRDFs improved classification performance by $\simeq 1.3\%$ in the training set and 0.4% in the test set compared to use of the histogram features (Table 1). This improvement in P_C is significant at this 90% classification level, since much of the input data is difficult to classify (around 10% of the data was misclassified by human experts).

6

3.3.3 Classification Results Using Quadratic MRDF Features

In Section 3.3.2, we used MRDF linear combinations of scale and rotation invariant histogram features. Such histogram features *do not necessarily ensure the best discriminatory information that is essential for classification* of overlapping data (as shown by the P_C (train) values obtained), since they only use second-order correlation information to discriminate between classes. We[5, 10] and others [20] have shown that higher-order correlations do exist in real data, and that such higher-order correlation information is useful for classification.

To investigate this, we computed quadratic MRDF features from the 16 histogram features using the training set and applied the same quadratic MRDF transformations to the histogram features for the 942 test samples. The modified k-NN classifier with k=17 was then used to classify the test samples using the quadratic MRDFs as input features. The classification results obtained using the k-NN on the quadratic MRDFs are shown in Table 3. As before, we noticed that best P_C for the training data was obtained when *all 16 original histogram features were used as inputs to the quadratic MRDF*. It is seen that for around 14 quadratic MRDF features, the classification accuracy for the training data samples (using a leave-one-out technique) is the highest ($P_C = 91.3\%$). The corresponding P_C for test data is also excellent and comparable ($P_C = 90.7\%$). These classification results are better than those obtained using any number of original histogram features in Table 1 (P_C=88% for training and P_C=88.7% for test data). This $\simeq 3.5\%$ improvement in classification performance for the training data, and 2% improvement in the test set performance is quite significant for this database, owing to the difficulty in classifying $\simeq 10\%$ of the data by human experts. The results obtained using the quadratic MRDFs in Table 3 compared to those in Table 1 show that *higher-order correlation information is indeed present* and is useful for discrimination. Standard classifiers intrinsically do not consider such higher-order correlation information for classification. Table 1 also shows that *standard classifiers tend to be confused by "un-important features"* (worse results are obtained when all 16 histogram features are used compared to when only the 7 most useful histogram features are used (Table 1)). However, we found the best P_C when all 16 original histogram features were used with our nonlinear MRDFs. Thus, *such features actually contain useful information when nonlinear features are used; including them noticeably improves classification performance* (Table 3).

3.4 Classifiers with best P_C are not Always Best

Obtaining the best overall P_C is not best in this application, since even P_C=90% can reject 10% of the good crop. Rather, we want to locate most/all of the clean nuts, while rejecting as many infested nuts as possible. To achieve preferable performance, we thus consider only detecting bad nuts. We only classify a nut as bad if the average distance of each test sample to the k-nearest *infested pistachio nut prototypes is less than the average distance to the k-nearest* clean nut prototypes by T. Using this technique, only bad nuts with high likelihood of correct classification (large T) are rejected and very few nuts are expected to have a large T and be misclassified as bad and be rejected. Note that varying T is equivalent to varying the ratio of the prior probabilities for each pistachio nut class. The results obtained by varying T with quadratic MRDFs as inputs to our modified k-NN classifier are shown in Table 4. Note that for training data, around 65% of the infested nuts are detected when P_C=99% for clean nuts. This is noticeably better than the classification accuracy of only 52% for the infested nuts (*a 13% improvement*) when the 7 best standard histogram features were used. The performance using nonlinear MRDFs is also around 5% better than when linear MRDFs were used. As seen in Table 4, *the percentage of infested nuts reduces to 1.05% of the total crop while only 1% of the good nuts were lost* when nonlinear MRDF features are used. Therefore, **the MRDF features yield improved classification performance**. The performance of the nonlinear MRDFs on the test set is similar.

Fig. 3a shows the ROC curve obtained when the value of T is varied over a wide range. The ROC curve was obtained by plotting P_C for clean nuts versus P_C for infested nuts as T was varied. The classification rates using nonlinear MRDF features are shown as a solid line and results using the 7 best histogram features are shown as a broken line in Fig. 3a. The modified k-NN classifier with k=17 was used. Better classification is characterized by a curve that is always above and to the left of other curves for all values of T. As seen, the nonlinear MRDF features are superior to the 7 best histogram features. Fig. 3b shows the percentage of correctly classified clean and infested nuts as T is varied for our nonlinear MRDF feature inputs to our modified k-NN classifier.

3.4.1 Classification (Product inspection) Summary

We have obtained excellent classification results on segmented agricultural products using nonlinear features (MRDFs) (computed from scale and rotation-invariant histogram features). Use of our new MRDF features was shown to provide better classification compared to histogram features. We also show that the classification results obtained using our
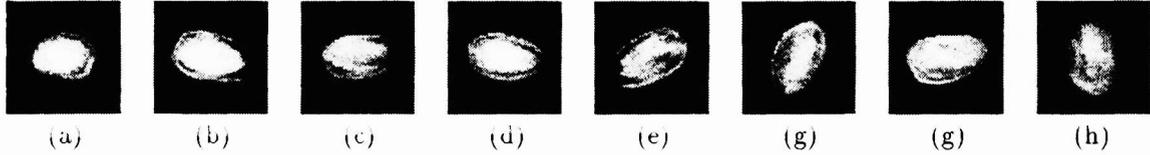
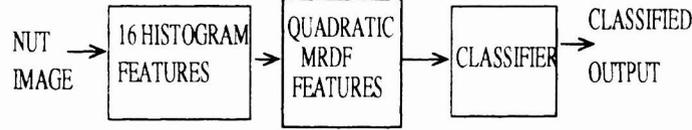Figure 1: X-ray images of clean (a-d) and infested (e-h) pistachio nuts.

Figure 2: Block diagram of the quadratic MRDF feature system used for pistachio nut classification.

Original Histogram Ftrs used	MRDF Ftrs. used	1	2	3	4	5	12
		P_C (%)					
5 Best	Train	88.1	88.7	**89.0**	-	-	-
Original Histogram Ftrs	Test	88.6	88.1	**88.2**	-	-	-
7 Best	Train	88.6	**88.9**	88.6	88.8	-	-
Original Histogram Ftrs	Test	88.5	**88.9**	88.2	88.0	-	-
All 16	Train	88.9	**89.3**	88.7	88.9	88.5	88.1
Original Histogram Ftrs.	Test	88.4	**89.1**	89.3	88.9	89.0	88.3

Table 2: P_C for linear MRDF features computed from different subsets of statistical histogram features (modified k-NN, k=17).

No. of quadratic MRDF Features (Computed from 16 Histogram Ftrs.)	3	4	6	8	10	12	14	16
P_C (Train) Modified k-NN (k=17)%	91	90.8	90.9	91.1	90.0	90.8	**91.3**	90.8
P_C (Test) Modified k-NN (k=17)%	89.2	89.2	90.02	90.02	90.33	90.5	**90.7**	90.2

Table 3: P_C (modified k-NN) with quadratic MRDF features computed from all 16 original histogram features.

T	0.091	0.0425	0.0325	0.0150	0.0045	-0.002
P_C(Good) Train	100%	99%	98%	97%	96.0%	95%
P_C(Bad) Train	42.7%	64.9%	69.6%	77.5%	81.9%	83.3%
P_C(Good) Test	99.84%	99%	98.5%	96.7%	96.5%	95.8%
P_C(Bad) Test	43.3%	60.6%	66.1%	77.2%	79.5%	81.9%

Table 4: Modified k-NN (k=17) classification accuracy with varying T using 14 quadratic MRDF features (computed from all 16 original histogram features).
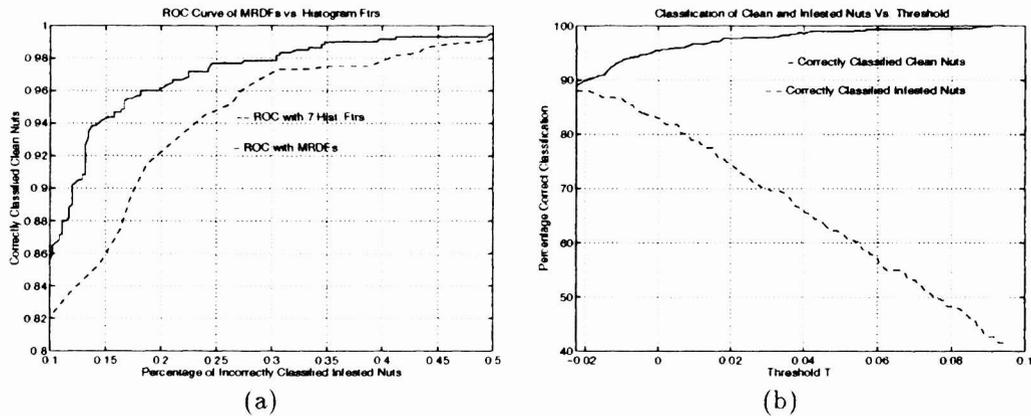

(a)

(b)

Figure 3: (a) ROC Curves using nonlinear MRDF features and 7 best histogram features, (b) Classification of clean and infested pistachio nuts by changing T.

new nonlinear MRDF features is preferable. We developed a new procedure to allow better classification performance for one class by varying the confidence thresholds on a new modified k-NN classifier.

Our nonlinear MRDF data results proves that higher-order correlations exist in such real-world data. The MRDF provides an improvement of around 2-3% in the current example, and is shown to be superior to other features by yielding far better ROC curves. *This improvement is commendable since many of the misclassified infested (clean) nuts do not look bad (good) (from X-ray images); e.g. nuts with splits can be misinterpreted as having worm tunnels, airgaps could be classified as infested regions, etc.* The truthed data provided was obtained by dissecting each nut, and visually classifying each dissected nut. Much of this information visible after dissection does not show up in the X-ray imagery. Hence, this is a formidable pattern recognition problem, and even 2-3% improvements in classification at the 90% classification level is notable. Despite these problems, the classification achieved here, 60% of bad nuts recognized reducing infested nuts to 1% of the crop with 1% of good nuts falsely rejected, appears to be very useful.

4 FEATURES FOR REPRESENTATION AND DISCRIMINATION

4.1 Feature Space

When features for representation of various versions of an object are desired, we maximize

$$E_R = \sum_n \sum_m \underline{\phi}_m^T \underline{C}_n \underline{\phi}_m \tag{7}$$

where, for generality, n classes are considered. This is standard KL eigenvector analysis. When both representation and discrimination are desired, we choose the $\underline{\phi}_m$ that maximize a weighted combination of E_D in (5) and E_R in (7), $kE_R + (k-1)E_D$. The $\underline{\phi}_m$ solutions (for the 2-class case) are the dominant eigenvectors of

$$[k\underline{I} + (1-k)(\underline{C}_1 + \underline{C}_2)]^{-1}[k(\underline{C}_1 + \underline{C}_2) + (1-k)\underline{R}_{12}]. \tag{8}$$

When only representation is desired, we set $k = 1$; when only discrimnation is desired, we set $k = 0$; when both are desired, intermediate values such as $k = 0.5$ are used. With extended $\underline{C}_H$ and $\underline{R}_H$ matrices, nonlinear solutions result *in closed form*, as before.

4.2 Feature Space Trajectory (FST) Processor

We consider robotics applications where parts must be classified and their pose estimated(for grasping, assembly, inspection, etc. applications). In these cases, we use the new features in Section 4.1 as detailed elsewhere [21]. We use these as the feature inputs to an FST neural net processor [22]. Fig. 4 shows the FST concept. Consider one machine part. At any orientation, it has some features and this pose is described by a point in feature space. We connect the points associated with adjacent vertices by straight lines to produce a trajectory (an FST). Fig. 4 shows two FSTs (FST-1 and FST-2) for 2 similar machine tool parts. To classify an unknown input object (a point in feature space), we calculate the FST to which it is closest. To estimate its pose, we determine the line segment and can thus achieve pose estimates more accurate than the angular spacing between vertices.

4.3 Results [21]

Consider the 4 similar parts in Figure 5. We used a training set of 36 views of each (at 10° differences in aspect) and a test set of 36 views of each (at intermediate aspect views). We considered the case of only discrimination (recognizing the class of an input test object); all 4 objects are quite similar. Tests (with $k = 0$) showed that fewer of our nonlinear features (than of our linear features) were needed to achieve $P_C = 100\%$ (only 3 nonlinear or 4 linear features were needed). With as many as 12 KL features or FK features, we could not achieve $P_C > 95.1\%$. Thus, *our features are preferable to standard ones and our nonlinear features perform* better. We also considered the case of both classification *and* pose estimation ($k = 0.5$). We found $P_C = 100\%$ and an average pose estimate accuracy of $7.2.^\circ$ We found that our nonlinear features were preferable and that all of our features were referable to standard ones. On other machine part problems, we found that linear features are preferable to nonlinear ones (our algorithm automatically determines this).
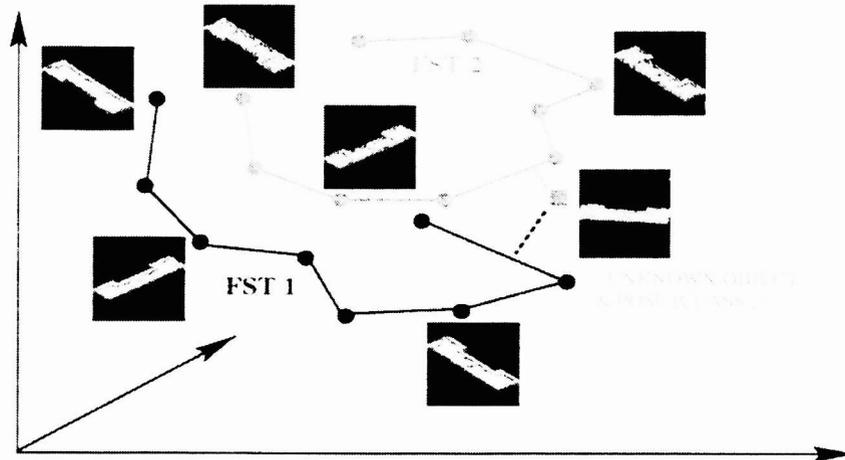
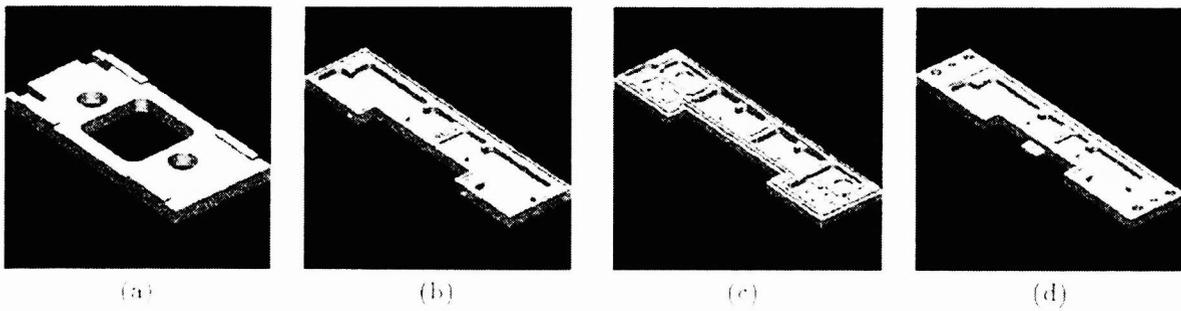Figure 4: FST for classification and pose estimation in feature space.

Figure 5: 0° aspect views of four similar flat CAD NIST machine parts from Allied Signal.

4.4 Iconic Inputs

In these machine part applications, the input data to our algorithm were image pixels (iconic). For such large input vectors, we use a new and very efficient 2-step algorithm to compute the ϕ_m transformation basis functions. It is detailed elsewhere [21]. It involves first transforming the data to a reduced orthogonal space (using a simplified performance measure) and then optimizing E_D (or a combination of E_D and E_R); this involves inverting the covariance matrix (the inversion is trivial in this new space).

5 FACE TRANSFORMATION

We have also discussed [23] the use of these new features in face recognition. One of the major problems in face recognition is that the input face can be at various poses; it is not reasonable to expect every pose view of every face in a database to be available.

We use our nonlinear discrimination features to estimate the pose of an input test face. We then use our representation features and transform the input face to any desired reference pose (we chose a frontal view). The method performs superbly. Attractive properties are that *the input face need not be present in the database and the input or reference pose of the face need not be in the database either.*

ACKNOWLEDGMENTS

References

[1] P. M. Keagy and T. Schatzki. Machine recognition of weevil damage in wheat radiographs. *Cereal Chemistry*, 70(6):696–700, 1993.

[2] P. M. Keagy, B. Parvin, and T.F. Schatzki. Machine recognition of navel worm damage in x-ray images of pastachio nuts. *Lebensm-Wiss U Technol*, 29:140–145, 1996.

[3] T.F. Schatzki, R.P. Haff, R. Young, I. Can, L.C. Lee, and N. Toyofuku. Defect detection in apples by means of x-ray imaging. *Proceedings of the Society of Photo-Optical Instrumentation Engineers*, 2907: paper 18, 1996.

[4] P. M. Keagy, B. Parvin, T. Schatzki, L.C. Le, D. Casasent, and D. Weber. Expanded image database of pistachio x-ray images and classification by conventional methods. In *Proc. SPIE*, volume 2907, pages 196–204, Nov. 1996.

[5] A. Talukder and D. Casasent. General methodology for simultaneous representation and discrimination of multiple object classes. *Optical Engineering, Special Issue on Advanced Recognition Techniques*, 37(3):904–913, March 1998.

[6] D. H. Foley and J. W. Sammon. An optimal set of discriminant vectors. *IEEE Trans. Comput.*, C-24:281–289, 1975.

[7] Y. Hamamoto, T. Kanaoka, and S. Tomita. On a theoretical comparison between the orthonormal discriminant vector method and discriminant analysis. *Pattern Recognition*, 26(12):1863–1867, 1993.

[8] R. A. Fisher. *Contributions to Mathematical Statistics*. John Wiley, New York, 1950.

[9] Ashit Talukder, David Casasent, Ha-Woon Lee, P. M. Keagy, and T. F. Schatzki. Modified binary watershed algorithm for segmentation of x-ray agricultural products. In *Proc. SPIE*, volume 3543, Nov. 1998.

[10] Ashit Talukder and David Casasent. Classification and pose estimation of objects using nonlinear features. In *Proc. SPIE: Applications and Science of Computational Intelligence*, volume 3390, pages 12–23, Apr. 1998.

[11] S.S. Wilks. *Mathematical Statistics*. Wiley, New York, 1962.

[12] J. Karhunen and J. Joutsensalo. Representation and separation of signals using nonlinear PCA type learning. *Neural Networks*, 7(1):113–127, 1994.

[13] Richard P. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Mag.*, 4(2):4–22, April 1987.

[14] D. L. Swets and J. J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. PAMI*, 18(8):831–836, Aug. 1996.

[15] K. S. Van Horn and T. R. Martinez. The minimum feature set problem. *Neural Networks*, 7(3):491–494, 1994.

[16] *SAS/STAT Users Guide: Release 6.03.* The SAS Institute, Cary, NC, 1988.

[17] Ashit Talukder and David Casasent. Nonlinear features for product inspection. In *Proc. SPIE*, volume 3715, Apr. 1999.

[18] I.D. Faux and M.J. Pratt. *Computational Geometry for Design and Manufacture.* Halstead Press, Horwood, NY, 1979.

[19] David Casasent, Ashit Talukder, and Ha-Woon Lee. X-Ray agricultural product inspection: Segmentation and classification. In *Proc. SPIE*, volume 3205, pages 46–55, Oct. 1997.

[20] G. Taylor and S. Coombes. Learning higher order correlations. *Neural Networks*, 6:423–427, 1993.

[21] Ashit Talukder and David Casasent. Nonlinear features for pose estimation and classification of machined parts. In *Proc. SPIE: Intelligent Robots and Computer Vision XVII: Algorithms, Techniques, and Active Vision*, volume 3522, Nov. 1998.

[22] D. Casasent, L. Neiberg, and M. Sipe. Feature space trajectory distorted object representation for classification and pose estimation. *Optical Engineering*, 37(3):914–923, Mar. 1998.

[23] Ashit Talukder and David Casasent. Pose estimation and transformation of faces from gray-scale images. In *Proc. SPIE: Intelligent Robots and Computer Vision XVII: Algorithms, Techniques, and Active Vision*, volume 3522, Nov. 1998.