Image Classification with Incomplete Multiresolution Data

David Williams Lawrence Carin

Abstract

We address the problem in which imagery to be classified is available from multiple resolutions. The proposed algorithm explicitly emphasizes features extracted from fine resolution imagery over those extracted from less reliable, coarse resolution imagery. The algorithm handles missing fine resolution features by analytically integrating out the missing features via an estimated conditional density function, conditioned on the observed features. This density function exploits the statistical relationship that exists between features at different resolutions, as well as those between features from different sensors, in the multi-sensor case. Experimental results are shown on several real data sets, consistently showing the advantage of the proposed algorithm over common alternative approaches.

1. Introduction

In many classification tasks, raw data occurs in the form of imagery, from which features are subsequently extracted. This framework arises in handwritten-character or face recognition tasks, as well as in medical and remote sensing applications. High-resolution imagery can capture salient attributes that are absent in low-resolution imagery. This information can lead to better class discrimination, and in turn, improved classification performance. Unfortunately, acquiring fine resolution imagery for all data points may be prohibitively expensive. Because the representation of fine resolution imagery requires more bits — which are costly to compute, store, and transmit — it is desirable to use coarse resolution imagery when doing so can achieve an adequate level of performance.

Some image-generating instruments can function at several different resolutions. For example, when a digital camera zooms in on a particular region of an image, the resolution increases from the standard (coarse) resolution to some finer resolution. As a result, data can be characterized by features extracted from multiple image-resolution levels.

This work addresses the problem of classification for data sets in which features are extracted from imagery at different resolutions for different data points. Additionally, this work also considers the case in which multiple *sensors* — each of which may operate at multiple resolutions — are also employed. In the multiple-sensor scenario, incomplete data results when some data points are interrogated by only a subset of sensors. Incomplete data also exists in the single-sensor case, because not all data points will have features extracted from imagery at all resolution levels. In summary, the novel problem we address in this work is of multi-sensor, multiresolution, incomplete-data classification.

It is important to emphasize that this work addresses a problem that — to the best of our knowledge — has not been addressed previously. In most previous "multiresolution" image classification work (*e.g.*, [5]), the original imagery actually exists at only a single resolution; the term "multiresolution" refers simply to wavelet decompositions [6] of the original single resolution imagery. In contrast, this work utilizes multiple raw images, each at a unique resolution. The ultimate classification objective also distinguishes this work from other multiresolution image classification work. Most multiresolution classification work strives for pixel-level classification via the segmentation of large scenes (*e.g.*, [1]). In contrast, in this work, a given image belongs to a single class, as in face recognition tasks.

When features from multiple resolutions are possessed, several approaches can be used to handle the fact that some data points may be characterized by features extracted from only a subset of the resolutions (the missing data problem). One approach would build a separate classifier for data from each resolution level. Assuming the set of possible resolutions is relatively small, this approach would be reasonable. The major drawback with this method, however, is that the dependencies of the data from each resolution are not exploited. In the multi-sensor case, in which each sensor may operate at multiple resolutions, this approach becomes infeasible. In addition to ignoring the correlations between sensors, the severe fragmentation of the data set — based on the combinations of which sensors and resolutions are observed — will leave insufficient data to train each classifier.

Another method would concatenate the features from the various resolutions; incomplete data arising from missing features resolutions would be handled in some way, such as by imputation [7]. However, such an approach would treat features obtained from images at different resolutions

equally. Intuitively, we should emphasize features extracted from high resolution imagery.

In this paper, we extend the work of Williams *et al*. [9] — in which missing data is analytically integrated out — to the case of multiresolution imagery. The proposed algorithm addresses how classification should be performed when multiple resolution imagery from each sensor is available. Because the data need not be available for all sensors or at all resolutions for all data points, incomplete-data issues arise. The algorithm presented here does not suffer from any of the drawbacks that plague the aforementioned methods. The proposed algorithm *requires only a single classifier be built*, regardless of the number of sensors or the number of resolutions involved in the problem. Moreover, all data is utilized, so *correlations among sensors, as well as among features at different resolutions, can be exploited*. Additionally, features extracted from different resolutions are not treated equally; rather, fine resolution features are given more importance. Furthermore, *the missing data that exists is handled in a principled manner, avoiding explicit imputation*. Specifically, the missing data is integrated out via the use of an estimated conditional density function that relates the dependencies of features both of a single given sensor at different sampling rates, as well as of features from different sensors.

The remainder of this paper is organized as follows. In Section 2 notation is explained for the proposed classification algorithm introduced in Section 3. Section 4 describes the multiresolution data sets for which experimental results are shown in Section 5. Section 6 consists of a discussion, followed in Section 7 by concluding comments and directions for future work.

2. Notation

Consider the case in which an instrument (*e.g*., a digital camera) generates raw data in the form of an image, from which features can subsequently be extracted. Henceforth, we shall refer to the image-generating instrument as a *sensor*.

Assume we possess S such sensors, the s-th of which can operate at R_s spatial sampling rates. Each of the S sensors may or may not be of the same modality, with diversity manifested by locating the sensors at different spatial positions. The possible sampling rates of each sensor are in general unique. Define Δ_r^s to be the r-th sampling rate of the s-th sensor, for $s = 1, 2, \ldots, S$ and $r = 0, 1, 2, \ldots, R_s$. Let Δ_0^s denote the finest sampling rate of the s-th sensor. We refer to all sampling rates except the finest (*i.e*., $\Delta_r^s \ \ \forall r \neq 0$) as *coarse sampling rates*. The resolution of an image, which is a function of and directly proportional to the (spatial) sampling rate of the sensor, is written $\mathcal{R}(\cdot)$. The image that results from operating a sensor at its finest sampling rate is referred to here as a fine resolution image. Sensors operating at coarse sampling rates generate coarse resolution imagery.

Assume that for a given sensor, the set of features that are extracted from the raw image data are fixed, regardless of the sampling rate of the sensor that generated the raw data. That is, for a given sensor, the specific features extracted will be identical for all sampling rates, but the feature *values* will in general be unique for each sampling rate.

Let $\boldsymbol{x}_i^{(s)} \in \mathbb{R}^{F_s}$ be the F_s features of the s-th sensor for the i-th data point, extracted from the highest resolution image of the s-th sensor. For all coarse sampling rates, let $\boldsymbol{z}_i^{(s,r)} \in \mathbb{R}^{F_s}$ be the F_s features of the s-th sensor for the i-th data point, extracted from the image obtained with the r-th spatial sampling rate of the s-th sensor. Define $\boldsymbol{x}_i = \left[\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}, \ldots, \boldsymbol{x}_i^{(S)}\right]$ to be the concatenated feature vectors extracted from imagery at each sensor's respective finest sampling rate. Similarly, define $\boldsymbol{z}_i = \left[\boldsymbol{z}_i^{(1)}, \boldsymbol{z}_i^{(2)}, \ldots, \boldsymbol{z}_i^{(S)}\right]$ to be the concatenated feature vectors extracted from each sensor's coarse resolution imagery, where $\boldsymbol{z}_i^{(s)} = \left[\boldsymbol{z}_i^{(s,1)}, \boldsymbol{z}_i^{(s,2)}, \ldots, \boldsymbol{z}_i^{(s,R_s)}\right]$. Hereafter, we shall refer to $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ as *primary* and *auxiliary* features (or data), respectively.

The data can alternatively be partitioned in terms of its observed and missing components. Let o_i^x be the set of sensors for which the i-th data point's primary features are observed. Let m_i^x be the (complementary) set of sensors for which the primary features are missing for the i-th data point. Similarly, let o_i^z be the set of sensor and coarse-sampling-rate pairs for which the auxiliary features for the i-th data point are observed. Let m_i^z be the (complementary) set of sensor and coarse-sampling-rate pairs for which the auxiliary features for the i-th data point are missing. To simplify notation, we shall suppress the superscripts when doing so will not cause confusion (*e.g*., $\boldsymbol{x}_i^{o_i^x}$, the primary features (from all sensors) that are observed for the i-th data point, will be written as $\boldsymbol{x}_i^{o_i}$). The primary and auxiliary data of the i-th data point can thus be written as $\boldsymbol{x}_i = [\boldsymbol{x}_i^{o_i}; \boldsymbol{x}_i^{m_i}]$ and $\boldsymbol{z}_i = [\boldsymbol{z}_i^{o_i}; \boldsymbol{z}_i^{m_i}]$, respectively.

A data point is deemed to be *complete* if we possess all primary features, for all sensors, for that data point (*i.e*., $m_i^x = \emptyset$). A data point is otherwise deemed *incomplete*. It should be noted that there exist two different types of incomplete data. First, a data point would be incomplete if some subset of sensors were never deployed (at any sampling rate) to the data point. A data point could also be incomplete even when all sensors were deployed to the data point; specifically, the data point would still be considered incomplete in this case if the finest sampling rate was not used for all of the sensors on the data point.

3. Classification with Incomplete Data

Assume we have a set of labeled (incomplete) data

$$\mathcal{D}_L = \{\boldsymbol{x}_i, \boldsymbol{z}_i, y_i, o_i^x, o_i^z, m_i^x, m_i^z\}_{i=1}^{N_L} \tag{1}$$

where $y_i \in \{-1, 1\}$ is the label of the i-th data point. Let $\boldsymbol{w}_s = \left[w_s^{(1)}, w_s^{(2)}, \ldots, w_s^{(F_s)}\right]$ represent the F_s weights of a classifier on the primary features of the s-th sensor. Let $\boldsymbol{w} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_S]$ be the classifier weights on the primary features of each sensor (*i.e.*, $\boldsymbol{x}_i$). Note that the number of features from each sensor need not be identical. It must be emphasized that the weights — and hence the resulting classifier — are on the features extracted from only the fine resolution imagery. We emphasize this caveat by using different notation for primary features extracted from the fine resolution imagery ($\boldsymbol{x}_i$), and auxiliary features extracted from coarse resolution imagery ($\boldsymbol{z}_i$).

In logistic regression (with a hyperplane classifier), the probability of label y_i given feature vector $\boldsymbol{x}_i$ is

$$p\left(y_i|\boldsymbol{x}_i, \boldsymbol{w}\right) = \sigma(y_i \boldsymbol{w}^T \boldsymbol{x}_i), \tag{2}$$

where $\sigma(\eta) = (1+\exp(-\eta))^{-1}$ is the sigmoid function and $\boldsymbol{w}$ constitutes a classifier. If all data points are complete, the weights of the classifier can be learned easily. Here we consider the case in which the data points are in general *incomplete* in the sense described previously.

Recall that the classifier is to be designed for only the primary data — the features extracted from the finest resolution imagery. We partition $\boldsymbol{x}_i$ into its observed and missing parts, $\boldsymbol{x}_i = [\boldsymbol{x}_i^{o_i}; \boldsymbol{x}_i^{m_i}]$. We apply the same partition to $\boldsymbol{w}$ to obtain $\boldsymbol{w} = [\boldsymbol{w}_{o_i}; \boldsymbol{w}_{m_i}]$.

With $\eta_i = \boldsymbol{w}_{m_i}^T \boldsymbol{x}_i^{m_i}$, (2) can be written as

$$p\left(y_i|\boldsymbol{x}_i^{o_i}, \boldsymbol{w}\right) = \sigma(y_i(\boldsymbol{w}_{o_i}^T \boldsymbol{x}_i^{o_i} + \eta_i)). \tag{3}$$

If the missing data $\boldsymbol{x}_i^{m_i}$ is integrated out, the needed probability of y_i given *all* observed features can be written as

$$\begin{aligned} &p\left(y_i|\boldsymbol{x}_i^{o_i}, \boldsymbol{z}_i^{o_i}, \boldsymbol{w}\right) \\ &\quad = \int p\left(y_i|\boldsymbol{x}_i^{o_i}, \boldsymbol{w}\right) p\left(\boldsymbol{x}_i^{m_i}|\boldsymbol{x}_i^{o_i}, \boldsymbol{z}_i^{o_i}\right) d\boldsymbol{x}_i^{m_i} \qquad (4) \\ &\quad = \int \sigma(y_i(\boldsymbol{w}_{o_i}^T \boldsymbol{x}_i^{o_i} + \eta_i)) p\left(\eta_i|\boldsymbol{x}_i^{o_i}, \boldsymbol{z}_i^{o_i}\right) d\eta_i. \qquad (5) \end{aligned}$$

Although the classifier uses only the primary data, the auxiliary data *is* exploited when primary data is missing, via the conditional density function $p\left(\boldsymbol{x}_i^{m_i}|\boldsymbol{x}_i^{o_i}, \boldsymbol{z}_i^{o_i}\right)$.

The integration in (5) can be performed analytically by making two mild assumptions. First, we assume that $p\left(\boldsymbol{x}_i, \boldsymbol{z}_i\right)$ is a Gaussian mixture model (GMM). This density function describes the relationship between the same features obtained from different resolutions; it also describes the relationship between features from different *sensors*. It then follows that

$$p(\boldsymbol{x}_i, \boldsymbol{z}_i) = p(\boldsymbol{x}_i^{m_i}, \boldsymbol{x}_i^{o_i}, \boldsymbol{z}_i^{o_i}) p(\boldsymbol{z}_i^{m_i}|\boldsymbol{x}_i^{m_i}, \boldsymbol{x}_i^{o_i}, \boldsymbol{z}_i^{o_i}), \tag{6}$$

where $p(\boldsymbol{x}_i^{m_i}, \boldsymbol{x}_i^{o_i}, \boldsymbol{z}_i^{o_i})$ is also necessarily a GMM. Introducing the notation $\boldsymbol{\chi}_i^{o_i} = [\boldsymbol{x}_i^{o_i}; \boldsymbol{z}_i^{o_i}]$, this GMM is

$$\begin{aligned} p(\boldsymbol{x}_i^{m_i}, \boldsymbol{\chi}_i^{o_i}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_i^{m_i} \\ \boldsymbol{\chi}_i^{o_i} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_k^{m_i} \\ \boldsymbol{\mu}_k^{o_i} \end{bmatrix}, \right. \\ \left. \begin{bmatrix} \boldsymbol{\Sigma}_k^{m_i m_i} & \boldsymbol{\Sigma}_k^{m_i o_i} \\ \boldsymbol{\Sigma}_k^{o_i m_i} & \boldsymbol{\Sigma}_k^{o_i o_i} \end{bmatrix}\right), \end{aligned} \tag{7}$$

where π_k are the mixing proportions that sum to unity. Moreover, $p(\boldsymbol{x}_i^{m_i}|\boldsymbol{\chi}_i^{o_i})$ is a GMM as well. Because of the linear relation $\eta_i = \boldsymbol{w}_{m_i}^T \boldsymbol{x}_i^{m_i}$, $p\left(\eta_i|\boldsymbol{\chi}_i^{o_i}\right)$ is also a GMM,

$$p\left(\eta_i|\boldsymbol{\chi}_i^{o_i}\right) = \sum_{k=1}^{K} \delta_k^i \mathcal{G}\left(\frac{\eta_i - \boldsymbol{w}_{m_i}^T \boldsymbol{\xi}_k^{m_i}}{\sqrt{\boldsymbol{w}_{m_i}^T \boldsymbol{\Omega}_k^{m_i} \boldsymbol{w}_{m_i}}}\right), \tag{8}$$

with the parameters

$$\delta_k^i = \frac{\pi_k \mathcal{N}\left(\boldsymbol{\chi}_i^{o_i}; \boldsymbol{\mu}_k^{o_i}, \boldsymbol{\Sigma}_k^{o_i o_i}\right)}{\sum_{\ell=1}^{K} \pi_\ell \mathcal{N}\left(\boldsymbol{\chi}_i^{o_i}; \boldsymbol{\mu}_\ell^{o_i}, \boldsymbol{\Sigma}_\ell^{o_i o_i}\right)} \tag{9}$$

$$\boldsymbol{\xi}_k^{m_i} = \boldsymbol{\mu}_k^{m_i} + \boldsymbol{\Sigma}_k^{m_i o_i} (\boldsymbol{\Sigma}_k^{o_i o_i})^{-1} (\boldsymbol{\chi}_i^{o_i} - \boldsymbol{\mu}_k^{o_i}) \tag{10}$$

$$\boldsymbol{\Omega}_k^{m_i} = \boldsymbol{\Sigma}_k^{m_i m_i} - \boldsymbol{\Sigma}_k^{m_i o_i} (\boldsymbol{\Sigma}_k^{o_i o_i})^{-1} \boldsymbol{\Sigma}_k^{o_i m_i} \tag{11}$$

where $\mathcal{G}(\eta_i) = (2\pi)^{-1/2} \exp\left\{-\eta_i^2/2\right\}$ is the standard univariate Gaussian density function with zero mean and unit variance.

The second (very accurate) assumption is that the sigmoid function can be approximated as a probit function (*i.e.*, a Gaussian cumulative distribution function)

$$\sigma(\alpha) \approx \int_{-\infty}^{\alpha} \mathcal{G}\left(\frac{u}{\beta}\right) du \tag{12}$$

where $\beta = \frac{\pi}{\sqrt{3}}$.

Mirroring the derivation in [9], it can be shown that the probability of y_i given only the observed portions of $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ can be expressed as a mixture of *sigmoids*:

$$p\left(y_i|\boldsymbol{x}_i^{o_i}, \boldsymbol{z}_i^{o_i}, \boldsymbol{w}\right) \approx \sum_{k=1}^{K} \delta_k^i \, \sigma\left(\frac{y_i \beta (\boldsymbol{w}_{m_i}^T \boldsymbol{\xi}_k^{m_i} + \boldsymbol{w}_{o_i}^T \boldsymbol{x}_i^{o_i})}{\sqrt{\boldsymbol{w}_{m_i}^T \boldsymbol{\Omega}_k^{m_i} \boldsymbol{w}_{m_i} + \beta^2}}\right). \tag{13}$$

The log-likelihood function of the incomplete data in (1) is then

$$\ell(\boldsymbol{w}) = \log p\left(\{y_i\}_{i=1}^{N_L} | \{\boldsymbol{x}_i^{o_i}\}_{i=1}^{N_L}, \{\boldsymbol{z}_i^{o_i}\}_{i=1}^{N_L}, \boldsymbol{w}\right) \tag{14}$$

$$\approx \sum_{i=1}^{N_L} \log \left[\sum_{k=1}^{K} \delta_k^i \, \sigma\left(\frac{y_i \beta (\boldsymbol{w}_{m_i}^T \boldsymbol{\xi}_k^{m_i} + \boldsymbol{w}_{o_i}^T \boldsymbol{x}_i^{o_i})}{\sqrt{\boldsymbol{w}_{m_i}^T \boldsymbol{\Omega}_k^{m_i} \boldsymbol{w}_{m_i} + \beta^2}}\right)\right]. \tag{15}$$

Although the objective function (15) to be maximized is no longer concave, an intelligent initialization of $\boldsymbol{w}$ — such as that used in [9] — will avoid most local maxima. The

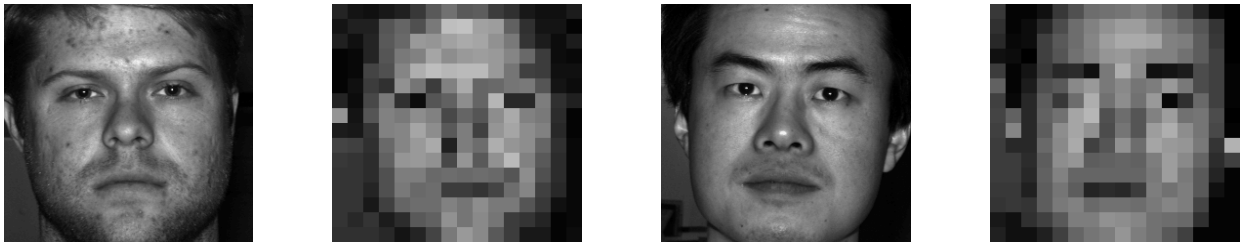

Figure 1. Example images of each of the two subjects at the two different resolutions from the YALE FACE data set. From left to right: fine resolution ($\mathcal{R}(\Delta_0)$) image of subject 1, coarse resolution ($\mathcal{R}(\Delta_0/16)$) image of subject 1, fine resolution image of subject 2, coarse resolution image of subject 2.

maximum-likelihood (ML) logistic regression classifier $\boldsymbol{w}$ can then be obtained, in spite of the missing data. Thereafter, the class predictions of an unlabeled testing data point with incomplete (missing) features is computed trivially using (13).

4. Multiresolution Data Sets

The proposed classification algorithm is intended for data sets consisting of imagery that exists at multiple resolutions. In lieu of actual multiresolution data, we simulate multiresolution imagery from single-resolution imagery in the following manner. Declare each original image to represent fine resolution imagery (at the highest spatial sampling rate, Δ_0). Coarse resolution imagery is then simulated by downsampling the imagery.

It is important to note that the particular features extracted from a given image will be identical, regardless of the image's resolution. Although the features are identical, the actual *values* of these features extracted from images at different resolutions will be unique.

The sizes of the five data sets used in this paper are summarized in Table 1. Additional data set details follow.

Table 1. Details of the data sets. For the LAND MINE data sets, class $+1$ corresponds to mines, while class -1 corresponds to clutter.

DATA SET	NUMBER OF SENSORS	NUMBER OF DATA POINTS IN CLASS $+1$	NUMBER OF DATA POINTS IN CLASS -1
YALE FACE	1	585	585
LAND MINE A	1	120	737
LAND MINE B	1	88	506
LAND MINE C	1	83	616
LAND MINE	4	91	622

4.1. Face Recognition

We first consider a face recognition task using data from the Yale Face Database B [2], which contains images of faces at various poses and under various illumination conditions. Specifically, the task we consider is to classify a given image of a face as belonging to one of two subjects under consideration. We first crop each image to a size of 256 pixels by 256 pixels, centered on the faces. Example coarse and fine resolution images of the two subjects in this set of experiments are shown in Figure 1. From each image, ten features are extracted. Each feature is the inner product between the image and one of ten "eigen-faces." These eigen-faces are the five eigen-images of each subject — which correspond to the largest eigenvalues — obtained via principal component analysis [4].

Figure 2. An example chip for one data point at the three different resolutions ($\mathcal{R}(\Delta)$, $\mathcal{R}(\Delta_0/8)$, and $\mathcal{R}(\Delta_0/16)$), from the LAND MINE B data set.

4.2. Single-Sensor Land Mine Detection

We next considered a classification task to distinguish land mines from clutter. Data in the form of an image is collected for a large area of land by a ground-penetrating radar (GPR) sensor. From this image, a simple pre-screener indicates points of interest. Each of these points of interest is subsequently considered to be a data point. For each data point, an image "chip" is extracted from the original image. These chips are 48 pixels by 48 pixels, centered at the pre-screener detection locations. Example measured image chips for one data point, at three different sampling rates, are shown in Figure 2. Each data point is subsequently characterized by 16 features that are extracted from a given chip. Three different measured land mine data sets (A, B, and C) are examined in this work; each data set is created from a unique image of a different large area of land.

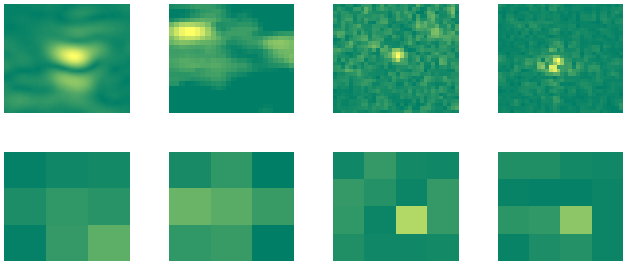

Figure 3. Example chips of the same one data point from each of the four sensors, from the multi-sensor LAND MINE data set. Fine resolution images appear in the top row; corresponding coarse resolution images appear in the bottom row.

4.3. Multi-Sensor Land Mine Detection

The objective for the multi-sensor data set used in this work is to discriminate between land mines and clutter. The data set is comprised of data collected by four airborne sensors: a GPR sensor, an electro-optic infrared (EOIR) sensor, a *Ku*-band synthetic aperature radar (SAR) sensor, and an *X*-band SAR sensor. Each sensor generates one image that covers a large area of land. The pre-processing steps taken to obtain image chips for each data point from each sensor are similar to those used in the single-sensor case. Different features are then extracted from the chips of different sensors; the number of features extracted from the chips of each sensor are 16, 5, 8, and 8, respectively. Example coarse and fine resolution image chips for one data point, from each of the four sensors, are shown in Figure 3.

5. Experimental Results

To evaluate the proposed incomplete-data classification algorithm, we applied it to the five multiresolution image data sets described above.

In each single-sensor experiment, four algorithms are applied, each of which handles the multiresolution data in a different manner. However, a logistic regression classifier is used for all methods. Table 2 summarizes the symbols that are used to denote the different methods in all figures showing results. These methods will be described in greater detail here.

The proposed approach builds a classifier for only the primary data; it handles missing primary data by integrating out the missing data, using the estimated density function relating both the primary and auxiliary data. This density function — a GMM — is accurately estimated using all available data, via the Variational Bayesian Expectation-Maximization algorithm presented in [9]. Because class labels are not used, both labeled and unlabeled data can be utilized in the estimation.

The second method builds a separate classifier for data from each resolution. That is, in the case of two resolutions, one classifier is built for features extracted from fine resolution imagery, while a second classifier is built for features extracted from coarse resolution imagery. The third method builds a classifier for the concatenated primary and auxiliary data; it handles missing primary data by integrating out the missing data, via the approach used in [9]. The difference between this method and the proposed method is that this method builds a classifier on both auxiliary and primary data, whereas the proposed method does so only on the latter. The fourth method also builds a classifier for the concatenated primary and auxiliary data; however, this method "fills in" missing primary data with the unconditional mean of the observed data.

For the multi-sensor experiments, the (second) method that builds a separate classifier for data from each resolution level is not used because insufficient training data would be available to train classifiers for each combination of observed primary and auxiliary data when multiple sensors are employed.

Each point on every curve of the figures showing results is an average over 20 trials. Each trial has a random partition of the data set into training and testing data, and randomly selected data points that are assumed to be missing the primary data. In all experiments on the LAND MINE data sets, 25% of the data points are used as labeled training data.

The area under a receiver operating characteristic curve (AUC) is given by the Wilcoxon statistic [3]

$$\text{AUC} = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \mathbf{1}_{x_m > y_n} \tag{16}$$

where $x_1, \ldots, x_M$ are the classifier outputs of data belonging to class 1, $y_1, \ldots, y_N$ are the classifier outputs of data belonging to class -1, and $\mathbf{1}$ is an indicator function. We present the results of the classification experiments in terms of the AUC. As an indicator of classification performance, the AUC is a more useful quantity than accuracy when significant class imbalance exists, as it does in the LAND MINE data sets (*c.f.* Table 1).

5.1. One Sensor with One Resolution Level

To evaluate the relative discriminative quality of the data at each resolution level, we first perform single-resolution complete-data experiments. Although no data is missing in these experiments, it is assumed that data are available from only a single resolution. The results of the complete-data case for the YALE FACE data set appear in Table 3, while those of the single-sensor LAND MINE data sets appear in Table 4. Values shown in the tables are the average AUC over 100 trials, where each trial has a unique training and testing data partition. As can be seen from the tables, the performance improves as the resolution increases.

Table 2. Explanation of symbols used in all figures showing results. A double circle, ⊙, for the proposed method indicates that it is statistically significantly the best method, according to a paired t-test, at the 95% confidence level.

SYMBOL	METHOD
○	PROPOSED METHOD
◇	SEPARATE CLASSIFIER FOR DATA FROM EACH RESOLUTION
△	SINGLE CLASSIFIER, INTEGRATE OUT MISSING DATA
▽	SINGLE CLASSIFIER, IMPUTE MEAN FOR MISSING DATA

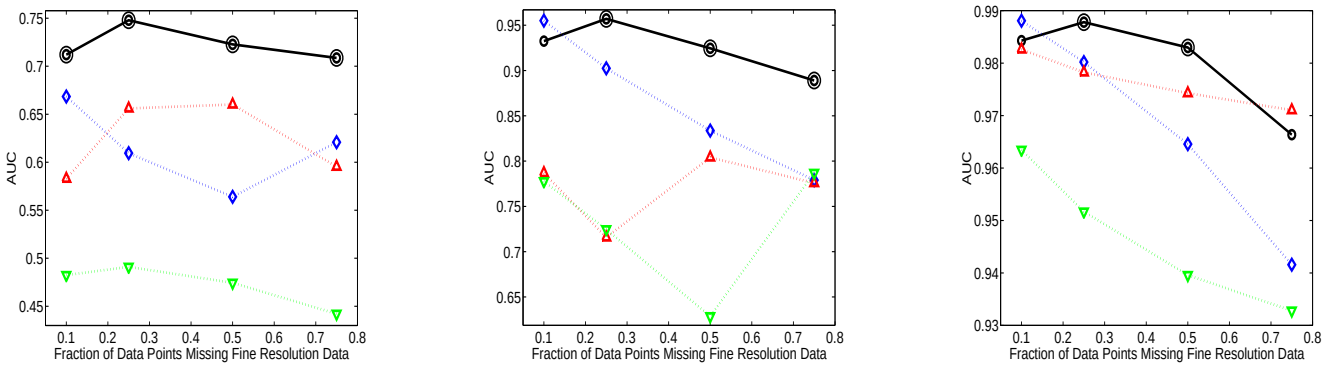

Figure 4. Experimental results for the YALE FACE data set. The results are for the cases when the (labeled) training data makes up (a) 1%, (b) 2%, and (c) 10% of the entire data set.

Table 3. The AUC for single-resolution, complete-data classification experiments from using data (features) obtained from the two resolutions, for the YALE FACE data set.

	FRACTION OF TRAINING DATA		
RESOLUTION	1%	2%	10%
$\mathcal{R}(\Delta_0)$	0.6919	0.9536	0.9947
$\mathcal{R}(\Delta_0/16)$	0.5783	0.8727	0.9420

Table 4. The AUC for single-resolution, complete-data classification experiments from using data (features) obtained from three different resolutions, for the single-sensor LAND MINE data sets.

	RESOLUTION		
DATA SET	$\mathcal{R}(\Delta_0)$	$\mathcal{R}(\Delta_0/8)$	$\mathcal{R}(\Delta_0/16)$
LAND MINE A	0.8486	0.8248	0.6628
LAND MINE B	0.8677	0.8400	0.7328
LAND MINE C	0.8293	0.8303	0.7055

5.2. One Sensor with Two Resolution Levels

We now apply our classification algorithm to the YALE FACE data set, and to three, single-sensor LAND MINE data sets. For all four data sets, we considered the case in which imagery was available at two different resolution levels. It was assumed that all data points had coarse resolution ($\mathcal{R}(\Delta_0/16)$) imagery, but certain fractions of data points were missing fine resolution ($\mathcal{R}(\Delta_0)$) imagery. The results of the face recognition experiments appear in Figure 4. The results of the land mine detection experiments appear in Figure 5. As can be seen from the figures, the proposed method consistently outperformed the alternative approaches.

5.3. One Sensor with Three Resolution Levels

We now repeat our experiments on the three, single-sensor LAND MINE data sets, with one change. Specifically, we consider the case in which imagery are available at *three* different resolution levels. It is assumed that all data points have coarse resolution ($\mathcal{R}(\Delta_0/16)$) imagery, but certain fractions of data points are missing the intermediate resolution ($\mathcal{R}(\Delta_0/8)$) and fine resolution ($\mathcal{R}(\Delta_0)$) imagery. Specifically, the fraction of data points assumed to be missing the intermediate resolution imagery is fixed to be half of the fraction of data points missing the fine resolution imagery. For example, if 50% of the data points are missing imagery with a resolution of $\mathcal{R}(\Delta_0)$, then 25% of the data points are missing imagery with a resolution of $\mathcal{R}(\Delta_0/8)$. The results of the experiments appear in Figure 6. Again, the proposed method performed favorably in comparison to the alternative approaches.

5.4. Four Sensors Each with Two Resolution Levels

The final set of experiments is on the four-sensor LAND MINE data set, where each sensor operates at two different resolutions. We consider the case in which 0%, 10%, and 25% of the data points are missing coarse resolution

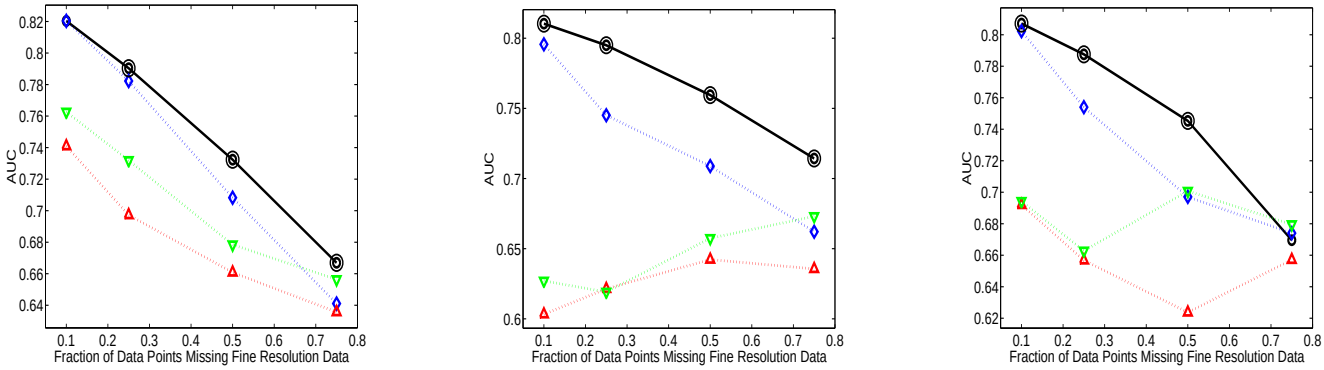

Figure 5. Experimental results for the single sensor LAND MINE data sets (a) A, (b) B, and (c) C. Two resolution levels — $\mathcal{R}(\Delta_0)$ and $\mathcal{R}(\Delta_0/16)$ — are used; no coarse resolution ($\mathcal{R}(\Delta_0/16)$) data is missing.

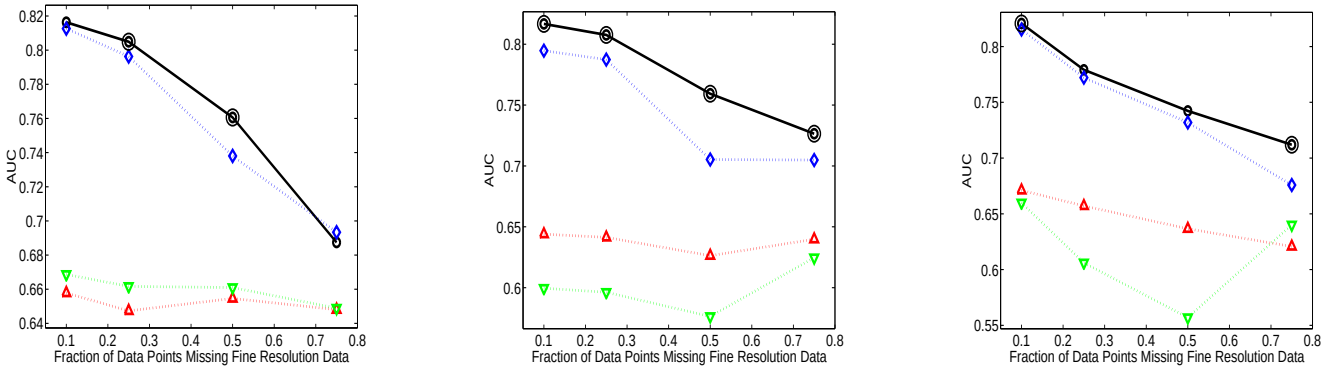

Figure 6. Experimental results for the single sensor LAND MINE data sets (a) A, (b) B, and (c) C. Three resolution levels — $\mathcal{R}(\Delta_0)$, $\mathcal{R}(\Delta_0/8)$, and $\mathcal{R}(\Delta_0/16)$ — are used; no coarse resolution ($\mathcal{R}(\Delta_0/16)$) data is missing. The intermediate resolution ($\mathcal{R}(\Delta_0/8)$) data is missing for half the number of data points that are missing the fine resolution ($\mathcal{R}(\Delta_0)$) data.

imagery from each sensor. Additionally, in all cases, certain fractions of fine resolution data are also missing. The case in which a data point is missing imagery at a particular resolution would arise if the sensor — while operating at the particular resolution — is flown over an area of land that does not contain the data point. The results of these multi-sensor, multiresolution experiments appear in Figure 7. As can be seen from the figure, the proposed approach significantly outperformed the alternative approaches.

6. Discussion

It should be emphasized that in the proposed method, the classifier weights are on the primary features, which are extracted from fine resolution imagery. However, the auxiliary features extracted from coarse resolution imagery are still utilized in the algorithm when primary data is missing. Specifically, missing primary data is analytically integrated out via the estimated density function, which models the relationship between the features from coarse resolution imagery and those same features from fine resolution imagery. The experimental results consistently show that the proposed method outperforms the alternative methods. Here we explain the reasons underlying this result.

High resolution imagery contains salient aspects that are absent in low resolution imagery. This statement is supported by the complete-data, single-resolution experiments (*c.f.* Tables 3 and 4), which showed that using the features extracted from fine resolution imagery achieves better performance than using features extracted from coarse resolution imagery. This result reinforces our belief that the finer resolution features should be trusted more than the coarse resolution features. Our proposed approach emphasizes the importance of the finer resolution data by building a classifier with only that data. The coarser resolution data is still exploited (via the estimated density function), albeit in an auxiliary role.

If a classifier is instead built on the conglomerated features extracted from different resolutions — as it was in two of the alternative methods — the information about the relative "quality" of the features is ignored. Concatenating features extracted from different resolution imagery also causes the feature-dimension to grow quickly, which can in turn lead to overfitting of the training data. The performance of the two feature-concatenation methods considered in the experiments actually degraded when features

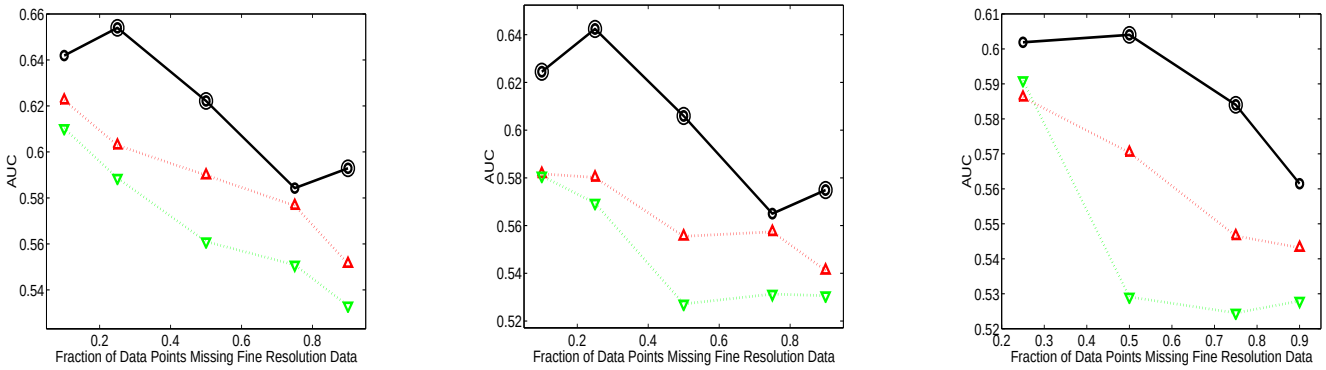

Figure 7. Experimental results for the four-sensor LAND MINE data set when (a) 0%, (b) 10%, and (c) 25% of each sensor's coarse resolution data is missing (in addition to the fine resolution data that is missing).

extracted from an additional intermediate resolution were included (*c.f*. Figures 5 and 6) because of this overfitting problem. In theory, a prior could be incorporated to combat overfitting, but additional complications arise as a result of having incomplete data. In contrast, the proposed method has no such overfitting issues.

The proposed method also consistently outperforms the method that builds a separate classifier for data from each resolution. This result is possible because the proposed method utilizes side information in the form of the estimated density function. By exploiting the statistical relationship that exists between features at different resolutions (as well as between features from different sensors, in the multi-sensor case), better performance can be achieved. This result can perhaps best be understood from the viewpoint of super-resolution techniques. Knowledge about a problem (*e.g*., that noise in an image is Gaussian) can be exploited to resolve a super-resolution image from several blurry images [8]. Similarly, in this problem, the knowledge of the statistical relationship between features at different resolutions can be exploited. Importantly, the proposed approach avoids the unnecessary intermediate step of forming an entire super-resolution *image*; instead, the ultimate goal is directly addressed, which is obtaining the equivalent of "super-resolution *features*."

7. Conclusion

Acquiring fine resolution imagery for all data points may be prohibitively expensive. This work presents a principled algorithm to classify imagery that may be available at multiple resolutions. Because some data points may possess imagery at only a subset of resolutions, the problem can be viewed as one of incomplete-data classification. The algorithm also naturally handles the case in which multiple *sensors* — each of which may operate at multiple resolutions — are used to acquire data. In summary, the novel problem we address is of multi-sensor, multiresolution, incomplete-data classification. Experimental results on several real data sets have demonstrated the advantage of the proposed algorithm. To spur interest in the multiresolution classification problem examined here, the data used in our experiments will be made freely available to interested researchers.

Future work will focus on the development of an active data acquisition algorithm that determines which data points should receive finer resolution imagery — and at which particular resolution level — in order to most improve performance. This active sensing concept is relevant for many applications, including medical imaging, remote sensing, and video tracking.

References

[1] C. Bouman and B. Liu. Multiple resolution segmentation of textured images. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 13(2):99–113, 1991. 1

[2] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001. 4

[3] J. Hanley and B. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982. 5

[4] I. Jolliffe. *Principal Component Analysis*. Springer, 1986. 4

[5] J. Li, R. Gray, and R. Olshen. Multiresolution image classification by hierarchical modeling with two dimensional hidden Markov models. *IEEE Trans. on Information Theory*, 46(5):1826–1841, 2000. 1

[6] S. Mallat. *A wavelet tour of signal processing*. Academic Press, 1998. 1

[7] D. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987. 2

[8] R. Tsai and T. Huang. Multi-frame image restoration and registration. *Advances in Computer Vision and Image Processing*, 1:317–339, 1984. 7

[9] D. Williams, X. Liao, Y. Xue, and L. Carin. Incomplete-data classification using logistic regression. In *Proc. Int'l Conf. Machine Learning (ICML)*, pages 977–984, 2005. 2, 3, 4, 5