Information Fusion 3 (2002) 277–287

INFORMATION FUSION

www.elsevier.com/locate/inffus

Robust image recognition by fusion of contextual information

Xubo B. Song [a,*], Yaser Abu-Mostafa [b], Joseph Sill [c], Harvey Kasdan [d], Misha Pavel [a]

[a] *Department of Electrical and Computer Engineering, OGI School of Science and Engineering, Oregon Health and Science University, 20000 NW Walker Road, Beaverton, OR 97006, USA*

[b] *Learning Systems Group, Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125, USA*

[c] *Ripfire, 870 Market Street, Suite 1105, San Francisco, CA 94102, USA*

[d] *International Remote Imaging Systems, Inc., 9162 Eton Avenue, Chatsworth, CA 91311, USA*

Received 19 November 2001; received in revised form 13 May 2002; accepted 12 August 2002

Abstract

This paper studies the fusion of contextual information in pattern recognition, with applications to biomedical image identification. In the real world there are cases where the identity of an object is ambiguous if the classification is based only on its own features. It is helpful to reduce the ambiguity by utilizing extra information, referred to as context, provided by accompanying objects. We investigate two techniques that incorporate context. The first approach, based on compound Bayesian theory, incorporates context by fusing the measurements of all objects under consideration. It is an optimal strategy in terms of achieving minimum set-by-set error probability. The second approach fuses the measurements of an object with explicitly extracted context. Its linear computational complexity makes it more tractable than the first approach, which requires exponential computation. These two techniques are applied to two medical applications: white blood cell image classification and microscopic urinalysis. It is demonstrated that superior classification performances are achieved by using context. In our particular applications, it reduces overall classification error, as well as false positive and false negative diagnosis rates.

Keywords: Context; Contextual information; Fusion; Pattern recognition; Image recognition; Compound Bayesian Theory

1. Introduction

For pattern classification, the primary source of information for each object of interest is the set of measurements associated with the object, commonly referred to as “features”. Often the identity of an object inferred solely from its features is ambiguous. This can be due to the noise in the measurements, less-than-optimal feature selection and extraction, or intrinsic overlap among the class-conditional feature distributions. This fundamentally limits the performance of a pattern recognition system.

In many application domains, the classification of an object can be assisted by considering more than simply the features of the object itself. One form of such extra information is “context”. In remote sensing image classification, where each pixel is part of ground cover, a pixel is more likely to be a glacier if it is in a mountainous area, than if surrounded by pixels of residential areas. In text analysis, one can expect to find certain letters occurring regularly in particular arrangement with other letters (qu, ee, est, tion, etc.). Such extra information conveyed by the accompanying entities is referred to as *contextual information*, or *context*. Context is fundamental to many, if not all, spheres of human endeavor. It occurs at many different levels including perceptual, cognitive, and statistical [30]. Context is also an essential component for sensor fusion [21]. Context incorporation is a mechanism that ensures accurate perception and appropriate interpretation of ambiguities. How humans appear to use context suggests that in automated pattern recognition we should be able to use context in conjunction with primary features in order to tackle such challenges as disambiguation and error-correction.

Context can assume various properties in different applications. It can be stationary or non-stationary (the relationship among contextual variables changes spatially or temporally). It can be local (only objects in a spatially or temporally local neighborhood are contextually relevant) or global (all objects under consideration are contextually relevant).

[*] Corresponding author.
E-mail address: xubosong@ece.ogi.edu (X.B. Song).

1566-2535/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S1566-2535(02)00092-1

Extensive efforts have been made to incorporate contextual information at one level or another [8–10,12,16,25,28,29,31,32]. One major school of context incorporation techniques falls in the Bayesian framework. A well known example is relaxation labeling [6,17,19,24]. It is based on a probabilistic iterative procedure for reducing ambiguities in the labeling of scene elements. As is typical for such Bayesian formulation, the computation complexity is exponential in the number of co-existing objects. A good review of methods that deal with such complexity is given in [18]. In one form or another, these methods take advantage of the locality of context to simplify computation. Another important school for probabilistic modeling of context is hidden Markov model (HMM) [13,23]. The definition of a HMM indicates that such context is local. Global context calls for a different approach, which is part of the aim of this paper.

This paper first studies the incorporation of contextual information formulated under compound Bayesian framework. The optimality of such formulation is investigated, both in terms of error probabilities and in terms of information gain. The benefit of such approach is exemplified by the application of white blood cell (WBC) image recognition. Compound Bayesian formulation can be thought of as context incorporation by fusing the measurements of all objects. To circumvent the exponential complexity of the compound Bayesian framework when simplifications are not feasible, we then introduce a method of context incorporation by fusing measurements of an object directly with derived context. This approach has only linear computational complexity. Its effectiveness is demonstrated by the application of automatic microscopic urinalysis.

2. Mathematical framework

2.1. Compound Bayesian theory for context

2.1.1. Formulation

Let us consider a set of N objects T_i, $i = 1, \dots N$. We associate each object T_i with a label c_i that is a member of a label set $\Omega = \{\omega_1, \dots, \omega_D\}$. Each object T_i is characterized by a feature vector $\mathbf{x}_i \in \mathbb{R}^{\mathrm{P}}$.

We make a conditional independence assumption which states $p(\mathbf{x}_i|c_i; c_1, c_2, \dots, c_J; \mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_K) = p(\mathbf{x}_i|c_i)$ for any $J = 0, 1, \dots, N$ and $J \neq i$, and $K = 0, 1, \dots, N$ (where $\mathbf{x}_0$ and c_0 are null elements). This assumption corresponds to a case where the appearance of an object is fully determined by its identity and not affected by the appearances and the identities of other accompanying objects. This is not always true in the real world. However, it is a good approximation for many domains. Under this assumption, it is easy to see that $p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N|c_1, c_2, \dots, c_N) = p(\mathbf{x}_1|c_1) \cdots p(\mathbf{x}_N|c_N)$.

Using the Compound Bayes rule, it follows that

$$p(c_1, c_2, \dots, c_N|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) = \frac{p(c_1|\mathbf{x}_1) \cdots p(c_N|\mathbf{x}_N) p(\mathbf{x}_1) \cdots p(\mathbf{x}_N) p(c_1, c_2, \dots, c_N)}{p(c_1) \cdots p(c_N) p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)} \tag{1}$$

Since $p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$ and $p(\mathbf{x}_i)$ are constant for a given set of objects, then

$$\begin{aligned} p(c_1, \dots, c_N|\mathbf{x}_1, \dots, \mathbf{x}_N) &\propto p(c_1|\mathbf{x}_1) \cdots p(c_N|\mathbf{x}_N) \frac{p(c_1, \dots, c_N)}{p(c_1) \cdots p(c_N)} \\ &= p(c_1|\mathbf{x}_1) \cdots p(c_N|\mathbf{x}_N) \rho(c_1, c_2, \dots, c_N) \end{aligned} \tag{2}$$

where

$$\rho(c_1, c_2, \dots, c_N) \triangleq \frac{p(c_1, c_2, \dots, c_N)}{p(c_1) \cdots p(c_N)} \tag{3}$$

The decision rule chooses class labels $\hat{c}_1, \hat{c}_2, \dots, \hat{c}_N$ such that

$$(\hat{c}_1, \hat{c}_2, \dots, \hat{c}_N) = \underset{(c_1, c_2, \dots, c_N)}{\operatorname{argmax}}\, p(c_1, c_2, \dots, c_N|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) \tag{4}$$

We term this decision rule the full context maximum posterior (FCMP) rule, in comparison with the partial context maximum posterior (PCMP) decision rule to be introduced in Section 2.2. Both FCMP and PCMP are context sensitive. In contrast, context-free approach views each object in isolation, assuming its identity to depend only on its own feature vector. The corresponding context-free maximum posterior (CFMP) decision rule selects the class label $\hat{c}_i$ for $i = 1, \dots, N$ such that $\hat{c}_i = \underset{c_i}{\operatorname{argmax}}\, p(c_i|\mathbf{x}_i)$.

The quantity $\rho(c_1, c_2, \dots, c_N)$, which we call the *context ratio* and through which the context plays its role, captures the dependence among the objects. In the special case where all the objects are independent, $p(c_1, c_2, \dots, c_N) = p(c_1) \cdots p(c_N)$, which implies $\rho(c_1, c_2, \dots, c_N) = 1$. In this case, there is no contextual information, and maximizing the context sensitive posterior probability in Eq. (1) is equivalent to maximizing the context-free posterior probabilities $p(c_i|\mathbf{x}_i)$ for all i.

2.1.2. Optimality of FCMP decision rule

We introduce the following notation: for a set of N elements $\{T_1, \dots, T_N\}$, let vector random variable $\underline{c} = (c_1, c_2, \dots, c_N)$ be the true labeling of the set, $\underline{\hat{c}} = (\hat{c}_1, \hat{c}_2, \dots, \hat{c}_N)$ the estimated labeling, and $\underline{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$ the feature vectors of the set of elements. For any element in the set, let scalar random variable c be its true labeling, $\hat{c}$ its estimated labeling, and $\mathbf{x}$ its feature vector.

Two types of errors occur when our inference of the class is different from the true class. The *set-by-set* error probability is defined as

$$P_e^{\mathrm{set}} = P(\underline{\hat{c}} \neq \underline{c}) = \int_{\underline{c}} \int_{\underline{x}} P(\underline{\hat{c}}(\underline{x}) \neq \underline{c} | \underline{x}, \underline{c}) p(\underline{x}, \underline{c}) \, \mathrm{d}\underline{x} \, \mathrm{d}\underline{c}$$

$$= \int_{\underline{c}} \int_{\underline{x}} [1 - \delta(\underline{\hat{c}}(\underline{x}) - \underline{c})] p(\underline{x}, \underline{c}) \, \mathrm{d}\underline{x} \, \mathrm{d}\underline{c}$$

where

$$\delta(\underline{s}) = \begin{cases} 1 & \text{if } \underline{s} = \underline{0} \\ 0 & \text{otherwise} \end{cases}$$

is the Kronecker delta.

The *element-by-element* error probability is defined as

$$P_e^{\mathrm{element}} = P(\hat{c} \neq c)$$

$$= \int_{\underline{c}} \int_{\underline{x}} \left[\frac{1}{N} \sum_{n=1}^{N} p(\hat{c}_n(\underline{x}) \neq c_n | \underline{x}, \underline{c}) \right] p(\underline{x}, \underline{c}) \, \mathrm{d}\underline{x} \, \mathrm{d}\underline{c}$$

$$= \int_{\underline{c}} \int_{\underline{x}} \frac{1}{N} \sum_{n=1}^{N} [1 - \delta(\hat{c}_n(\underline{x}) - c_n)] p(\underline{x}, \underline{c}) \, \mathrm{d}\underline{x} \, \mathrm{d}\underline{c}$$

According to the definition of set-by-set error, a set of elements is correctly classified *if and only if* every single element in the set is correctly classified. For element-by-element error, the error is counted on an element-by-element basis. The difference between a set and an element is analogous to that between a word and a letter. We are usually more concerned with set-by-set error than with element-by-element error.

It is important to point out that FCMP in Eq. (4) is the decision rule which achieves minimum set-by-set error probability. This is essentially the optimality of Bayes error rate. Conditioned on the collective feature vectors $\underline{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, no other decision rule is better in terms of achieving a smaller set-by-set error probability. The same logic implies that if conditioned only on isolated feature $\mathbf{x}_i$, the maximum posterior decision rule for $p(c_i|\mathbf{x}_i)$ given $\mathbf{x}_i$ achieves minimum element-by-element probability of error. However, it is possible that a decision rule conditioned on $\underline{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ has smaller element-by-element error probability than the one obtained conditioning only on isolated feature $\mathbf{x}_i$, since more information is being utilized.

2.1.3. Information-theoretic interpretation of FCMP decision rule

Define N_d as the number of objects in class d, and $v_d = N_d/N$ the frequency of class d. Clearly, $\sum_{d=1}^{D} N_d = N$ and $\sum_{d=1}^{D} v_d = 1$. Let P_d be the prior probability of class d, for $d = 1, \ldots, D$, and $\mathbf{P} = (P_1, P_2, \ldots, P_D)$ be the class prior probability vector. Let $v = (v_1, v_2, \ldots, v_D)$ the class frequency vector. Taking logarithms on both sides of Eq. (2) gives

$$\ln p(c_1, \ldots, c_N | \mathbf{x}_1, \ldots, \mathbf{x}_N) = \sum_{i=1}^{N} \ln p(c_i|\mathbf{x}_i) + \ln \frac{p(c_1, \ldots, c_N)}{p(c_1) \cdots p(c_N)} + \text{constant} \tag{5}$$

$$= \sum_{i=1}^{N} \ln p(c_i|\mathbf{x}_i) + \ln p(c_1, \ldots, c_N) + N\mathscr{H}(v \| \mathbf{P}) + N\mathscr{H}(v) + \text{constant} \tag{6}$$

where $\mathscr{H}(v\|\mathbf{P}) = \sum_{d=1}^{D} v_d \ln(v_d/P_d)$ is the relative entropy between v and $\mathbf{P}$, and $\mathscr{H}(v) = -\sum_{d=1}^{D} v_d \ln v_d$ is the entropy of the class frequency.

The above relation implies that maximizing $p(c_1, \ldots, c_N|\mathbf{x}_1, \ldots, \mathbf{x}_N)$ using context has the effect of trying to achieve a trade-off among several factors: the likelihood of each object given its feature (the first term), the likelihood of the set of objects appearing jointly (the second term), the distance of the class frequency profile of the set of objects from the prior distribution of the classes (the third term), and the entropy of the class frequency profile (the forth term). The maximization of the fourth term implies that the least amount of further information is assumed about the frequency profile. The first term depends on the features, and the other three depend only on the labels.

Another way to look at the benefit of context is its information gain. It follows from the chain rule for conditional entropy [5] that

$$H(c_1, c_2, \ldots, c_N | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) \leqslant H(c_1|\mathbf{x}_1) + H(c_2|\mathbf{x}_2) + \cdots + H(c_N|\mathbf{x}_N)$$

Equality is achieved if and only if $p(c_i|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N; c_1, \ldots, c_{i-1}) = p(c_i|\mathbf{x}_i)$ hold true for all $i = 1, \ldots, N$. This condition implies that c_i is fully determined by $\mathbf{x}_i$ and nothing else, which means that there is no context and therefore no information gain by considering context. When there is contextual information conveyed by other objects, this condition does not hold, in which case $H(c_1, c_2, \ldots, c_N|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ is strictly less than $H(c_1|\mathbf{x}_1) + H(c_2|\mathbf{x}_2) + \cdots + H(c_N|\mathbf{x}_N)$, and context provides information gain.

2.1.4. Complexity problem

When implementing the FCMP decision rule, we want to find a combination of $(c_1, \ldots, c_N)$ that maximizes $p(c_1, \ldots, c_N|\mathbf{x}_1, \ldots, \mathbf{x}_N)$. Suppose the D dimensional context-free probability vector $p(c_i|\mathbf{x}_i)$ is known for all $i = 1, \dots, N$. To compute $p(c_1, \dots, c_N|\mathbf{x}_1, \dots, \mathbf{x}_N)$ for all possible combinations of $(c_1, \dots, c_N)$, the total number of multiplications is $(2N+1)D^N$, and the complexity to find the maximum is D^N. For the WBC recognition problem, $D = 14$ and N is typically around 600, the computation is virtually intractable. Sometimes additional constraints can be used to reduce computation, as is the case of WBC identification (Section 3). Often such simplifications are not feasible. This limits the direct use of compound Bayesian formulation.

2.2. Fusion of measurements with contextual information

Compound Bayesian theory can be viewed as a way of incorporating context by fusing the measurements for all objects in a set together. It is set-centered in the sense that a decision is made simultaneously on all the objects in the set. To avoid the exponential computation, an alternative is to fuse the measurements of an object directly with the context. Such context can be derived/extracted from the measurements of all objects in the set. It will be object-centered—only the decision about one object is made at a time. Its formulation is given as follows.

2.2.1. Formulation

Let A be the derived context. The physical definition of A depends on the problem at hand. For example, A can be the percentage profile of all the classes, or the binary presence vector of the classes, or the presence of one or a few particular classes. A can also represent certain external information sources, such as the chemistry result in urinalysis which is a urine test different from, yet related to, microscopic urinalysis [1].

Once again, we make the conditional independence assumption $p(\mathbf{x}_i|c_i, A) = p(\mathbf{x}_i|c_i)$, then

$$\begin{aligned} p(c_i|\mathbf{x}_i, A) &= p(c_i|\mathbf{x}_i)\frac{p(c_i|A)}{p(c_i)}\frac{p(A)p(\mathbf{x}_i)}{p(\mathbf{x}_i;A)} \\ &\propto p(c_i|\mathbf{x}_i)\frac{p(c_i|A)}{p(c_i)} = p(c_i|\mathbf{x}_i)\rho(c_i, A) \qquad (7) \end{aligned}$$

The context sensitive posterior probability $p(c_i|\mathbf{x}_i, A)$ is obtained through the context-free posterior probability $p(c_i|\mathbf{x}_i)$ modified by context ratio $\rho(c_i, A) = p(c_i|A)/p(c_i)$.

The corresponding decision rule chooses class label $\hat{c}_i$ for element i such that

$$\hat{c}_i = \operatorname*{argmax}_{c_i} p(c_i|\mathbf{x}_i, A) \qquad (8)$$

which we term as the PCMP decision rule, since it uses derived context.

2.2.2. Computational complexity

This PCMP approach treats each element in a set individually, with additional information from context-bearing factor A. Again we assume the D dimensional context-free probability vector $p(c_i|\mathbf{x}_i)$ is already known for all $i = 1, \dots, N$. Once the context A is obtained, we want to maximize $p(c_i|\mathbf{x}_i, A)$ from D possible values that c_i can take on, for all i. The total number of multiplications is $2N$, and the complexity for finding the maximum is ND. Both are linear in N. Table 1 gives the comparison.

Table 1
Comparison of computational costs

Methods	Number of multiplications	Complexity for finding max
Compound Bayesian (FCMP)	$(2N+1)D^N$	D^N
Fusing measurements with context (PCMP)	$2N$	ND
Context free (CFMP)	0	ND

2.3. Test on a toy problem

The fusion of context to improve performance can be first demonstrated by a toy problem. We compare the performance of context sensitive approaches FCMP (Eq. (4)) and PCMP (Eq. (8)) to that of the CFMP decision rule for the both set-by-set and element-by-element error probability. In this toy problem, there are $N = 3$ elements in each set $\underline{c} = (c_1, c_2, c_3)$. Each element takes ternary values from $\Omega = \{0, 1, 2\}$, therefore $D = 3$. The joint distribution $p(\underline{c})$ is specified in Table 2.

Table 2
$p(\underline{c})$ for the toy problem

$\underline{c}$	$p(\underline{c})$
0 0 0	0
0 0 1	0.15
0 0 2	0
0 1 0	0.05
0 1 1	0.02
0 1 2	0
0 2 0	0
0 2 1	0
0 2 2	0
1 0 0	0.08
1 0 1	0.15
1 0 2	0
1 1 0	0.05
1 1 1	0.05
1 1 2	0.10
1 2 0	0
1 2 1	0.15
1 2 2	0.02
2 0 0	0
2 0 1	0
2 0 2	0
2 1 0	0
2 1 1	0.02
2 1 2	0.02
2 2 0	0
2 2 1	0.02
2 2 2	0.12

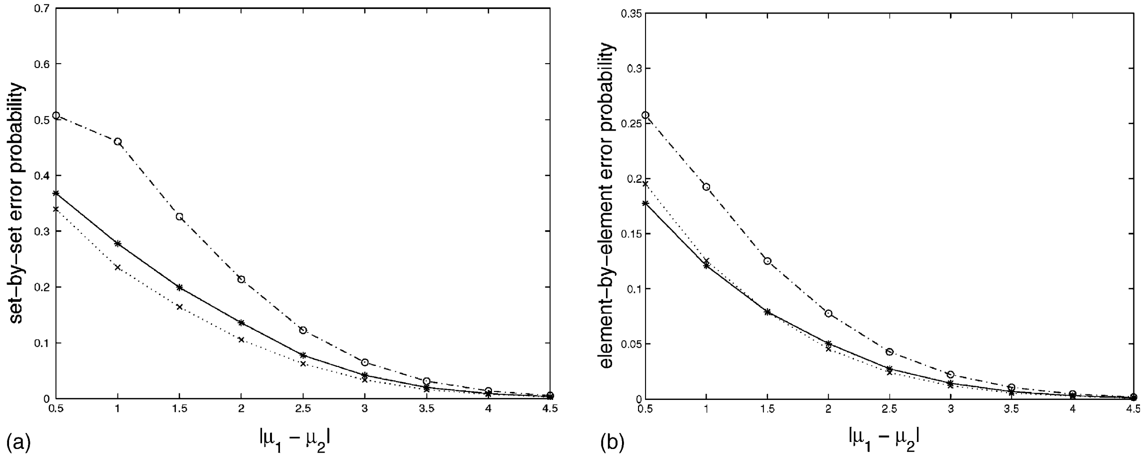

Fig. 1. Comparison of error probabilities with and without context: (a) set-by-set error, (b) element-by-element error. In both figures, the dash-dot line is for CFMP, the dotted line is with FCMP, and the solid line is with PCMP.

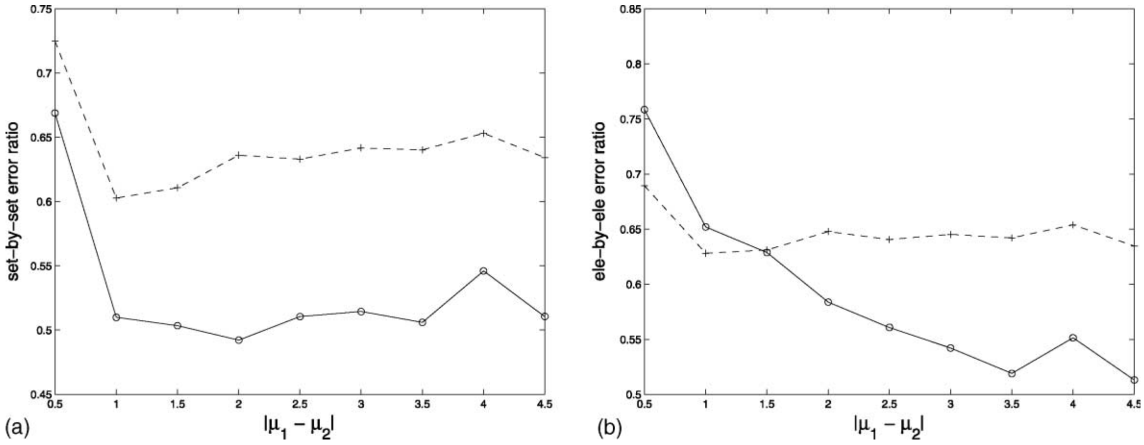

Fig. 2. The ratio of with-context error probability to without-context error probability: (a) set-by-set error probabilities, (b) element-by-element error probabilities. In both figures, the dashed line is with PCMP, and the solid line is with FCMP.

The conditional independence assumption leads to $p(\underline{x}|\underline{c}) = \prod_{i=1}^{N} p(x_i|c_i)$. The class-conditional feature distributions are Gaussians: $p(x|c=0) = N(\mu_0, \sigma_0)$, $p(x|c=1) = N(\mu_1, \sigma_1)$, and $p(x|c=2) = N(\mu_2, \sigma_2)$. In this toy problem, the derived context A is the presence or absence of class 0. We choose $\mu_0 = 0$, $\sigma_0 = 0.01$ and both μ_1, $\mu_2 \gg \mu_0$, and $\sigma_1 = 1.0$ and $\sigma_2 = 0.6$. The distance between μ_1 and μ_2 determines the separability of class 1 and 2. We compare the performances of the three approaches as we vary $|\mu_1 - \mu_2|$. Monte Carlo experiments were run and the results are illustrated in Fig. 1. As we can see, both the context sensitive approaches FCMP and PCMP consistently outperform the context-free approach, for both set-by-set and element-by-element error probability. In terms of set-by-set error, the FCMP algorithm is the best, which is expected due to its optimality. In terms of element-by-element error probability, both FCMP and PCMP algorithms are better than the CFMP algorithm, but there is no clear winner between the former two. Error probability, both set-by-set and element-by-element, decrease as $|\mu_1 - \mu_2|$ becomes larger. This is not surprising since there is less ambiguity. However, the significance of context does not diminish. The ratios of error probabilities for both FCMP and PCMP to context-free error probabilities actually decrease as $|\mu_0 - \mu_1|$ becomes large, as shown in Fig. 2, which implies that the effect of context becomes more significant in a relative sense.

3. White blood cell identification

3.1. Introduction

WBC analysis is one of the major routine laboratory examinations. In various physiological and pathological conditions the relative percentage composition of the WBC changes. An estimate of the percentage of each class present in a blood sample conveys information which is pertinent to the hematological diagnosis. Most WBC differentiation depends almost entirely on manual specimen preparation and human interpretation, and

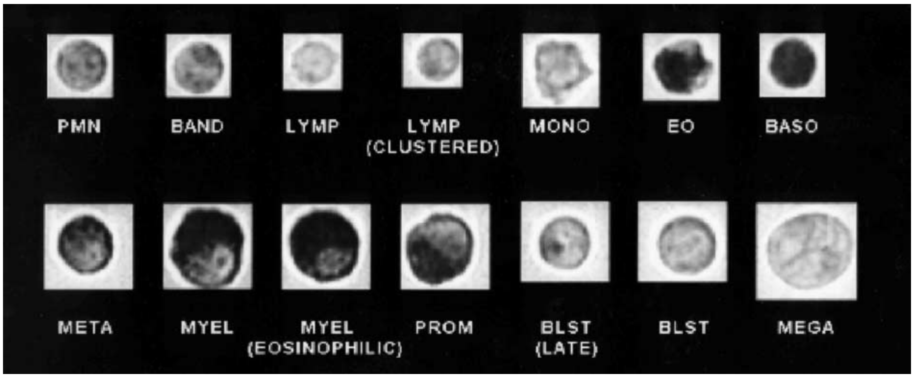

Fig. 3. Examples of some of the WBC images.

more than 90% of the direct costs are labor. The availability of Automated Intelligent Microscopy Flow Imaging technology [15] makes it possible to have automated differentiation, which will reduce labor and health care costs, and is more efficient. Typical commercial differential WBC counting systems are designed to identify five major mature cell types. But blood samples may also contain immature cells. These cells occur infrequently in a normal specimen, and most commercial systems will simply indicate the presence of these cells because they cannot be individually identified by the systems. But it is precisely these cell types that relate to the production rate and maturation of new cells and thus are important indicators of hematological disorders. Our system is designed to differentiate fourteen WBC types which includes the five major mature types: segmented neutrophils, lymphocytes, monocytes, eosinophils, and basophils; *and* the immature types: bands (unsegmented neutrophils), metamyelocytes, myelocytes, promyelocytes, blasts, and variant lymphocytes; as well as nucleated red blood cells and artifacts. Differential counts are made based on the cell classifications, which further leads to diagnosis or prognosis. There is a range of differential counts of all cell types within which a specimen is considered normal. A specimen is abnormal if the differential counts of one or more cell types fall out of their ranges.

The data is provided by International Remote Imaging Systems (IRIS), Inc. Blood specimens are collected at Harbor UCLA Medical Center from local patients, then dyed with Basic Orange 21 metachromatic dye supravital stain. The specimen is then passed through a flow microscopic imaging and image processing instrument, where the blood cell images are captured. Each image contains a single cell with full color. There are typically 600 images from each specimen. The task of the cell recognition system is to categorize the cells based on the images. Fig. 3 is an example of cell images of various types.

The size of cell images are automatically tailored according to the size of the cell in the images. Images containing larger cells have bigger sizes than those with small cells. The range varies from 20×20 to 40×40 pixels, and the average size is around 25×25. At the preprocessing stage, the images are segmented using adaptive thresholding to set the cell interior apart from the background. Features based on the interior of the cells are extracted from the images. The features include size, shape, color [1] and texture. See Table 3 for the list of features. [2] These features by design are rotation and translation invariant.

The features are fed into a non-linear feed-forward neural network with 20 inputs, 15 hidden units with sigmoid transfer functions, and 14 sigmoid output units. A cross-entropy error function is used in order to give the output a probability interpretation. Denote the input feature vector as $\mathbf{x}$, the network outputs a D dimensional vector ($D = 14$ in this case) $\mathbf{p} = \{p(d|\mathbf{x})\}, d = 1, \ldots, D$, where $p(d|\mathbf{x}) = \text{Prob}(\text{a cell belongs to class } d|\text{feature } \mathbf{x})$. The maximum posterior context-free decision rule at this stage is $d(\mathbf{x}) = \underset{d}{\text{argmax}}\, p(d|\mathbf{x})$.

3.2. Context for WBC analysis

The context-free cell-by-cell decision is only based on the features presented by a cell, without looking at any other cells. When human experts make decisions, they always look at the whole specimen, taking into consideration the identities of other cells and adjusting the cell-by-cell decision on a single cell according to the company it keeps. On top of the visual perception of the cell patterns, such as shape, color, size and texture, comparisons and associations, either mental or visual, with other cells in the same specimen are made to infer the final decision. A cell is assigned a certain identity if the company it keeps supports that identity. For instance,

[1] A color image is decomposed into three intensity images—red, green and blue respectively.

[2] The red–blue distribution is the pixel-by-pixel log(red) − log(blue) distribution for pixels in cell interior. The red distribution is the distribution of the red intensity in cell interior.

Table 3
Features extracted from cell images

Feature number	Feature description
1	Cell area
2	Number of pixels on cell edge
3	The 4th quantile of red–blue distribution
4	The 4th quantile of green–red distribution
5	The median of red–blue distribution
6	The median of green–red distribution
7	The median of blue–green distribution
8	The standard deviation of red–blue distribution
9	The standard deviation of green–red distribution
10	The standard deviation of blue–green distribution
11	The 4th quantile of red distribution
12	The 4th quantile of green distribution
13	The 4th quantile of blue distribution
14	The median of red distribution
15	The median of green distribution
16	The median of blue distribution
17	The standard deviation of red distribution
18	The standard deviation of green distribution
19	The standard deviation of blue distribution
20	The standard deviation of the distance from the edge to the mass center

the difference between lymphocyte and blast can be very subtle sometimes, especially when the cell is large. A large unusual mononuclear cell with the characteristics of both blast and lymphocyte is more likely to be a blast if accompanied by other abnormal cells or an abnormal distribution of the cells. Context incorporation is treated as the post-processing after the cell-by-cell decisions.

In this application, it is the count in each class, rather than the particular ordering or numbering of the objects, that matters, since it is the percentage profile of all classes in a specimen that convey diagnostic information [26]. Under such circumstance the contextual ratio $\rho(c_1, c_2, \ldots, c_N)$ is a function of the counts in each class. It can be shown that

$$\rho(c_1, c_2, \ldots, c_N) = \frac{p(c_1, c_2, \ldots, c_N)}{p(c_1) \cdots p(c_N)} = \frac{N_1! \cdots N_D! p(v_1, v_2, \ldots, v_D)}{N! P_1^{N_1} \cdots P_D^{N_D}} \doteq \rho(v_1, \ldots, v_D)$$

Therefore,

$$\begin{aligned} &p(c_1, \ldots, c_N | \mathbf{x}_1, \ldots, \mathbf{x}_N) \\ &\quad \propto p(c_1|\mathbf{x}_1) \cdots p(c_N|\mathbf{x}_N) \rho(c_1, c_2, \ldots, c_N) \qquad (9) \\ &\quad \propto p(c_1|\mathbf{x}_1) \cdots p(c_N|\mathbf{x}_N) \rho(v_1, v_2, \ldots, v_D) \qquad (10) \end{aligned}$$

3.3. Observations and simplifications

Direct implementation of the proposed algorithm is difficult due to the computational complexity. In the application of WBC identification, simplification is possible. We observed the following: First, we are primarily concerned with the class of blast, whose presence has clinical significance. Secondly, we only confuse blast with the class of lymphocyte. (The difference in appearances between these two classes can be very subtle.) In other words, for a potential blast, $p(\text{blast}|\mathbf{x}) \gg 0$, $p(\text{lymphocyte}|\mathbf{x}) \gg 0$, $p(\text{any other class}|\mathbf{x}) \approx 0$. Finally, we are fairly certain about the classification of all other classes, i.e., $p(\text{a certain class}|\mathbf{x}) \approx 1$, $p(\text{any other class}|\mathbf{x}) \approx 0$. Based on the above observations, we can simplify the algorithm, instead of doing an exhaustive search.

Let $p_i^d = p(c_i = d|\mathbf{x}_i)$, $i = 1, \ldots, N$. More specifically, let $p_i^{\mathrm{B}} = p(\text{blast}|\mathbf{x}_i)$, $p_i^{\mathrm{L}} = p(\text{lymphocyte}|\mathbf{x}_i)$ and $p_i^* = p(\text{class} * |\mathbf{x}_i)$ where $*$ is any class but blast. Suppose there are K potential blasts. Order the $p_1^{\mathrm{B}}, p_2^{\mathrm{B}}, \ldots, p_K^{\mathrm{B}}$'s in a descending manner over i, such that $p_1^{\mathrm{B}} \geqslant p_2^{\mathrm{B}} \geqslant \cdots \geqslant p_K^{\mathrm{B}}$. Then the probability that there are k blasts is

$$P_{\mathrm{B}}(k) = p_1^{\mathrm{B}} \cdots p_k^{\mathrm{B}} p_{k+1}^{\mathrm{L}} \cdots p_K^{\mathrm{L}} p_{K+1}^* \cdots p_N^* \rho\left(v_{\mathrm{B}} = \frac{k}{N}, v_{\mathrm{L}} = v'_{\mathrm{L}} + \frac{K-k}{N}, v_3, \ldots, v_D\right)$$

where v'_{L} is the proportion of unambiguous lymphocytes and $v_3, \ldots, v_D$ are the proportions of the other cell types.

We pick the optimal number of blasts k^* that maximizes $P_{\mathrm{B}}(k)$, $k = 1, \ldots, K$.

3.4. The algorithm and complexity

Step 1: Estimate $\rho(v_1, \ldots, v_D)$ from the database, for $d = 1, \ldots, D$.
Step 2: Compute the object-by-object "no context" posterior probability $p(c_i|\mathbf{x}_i)$, $i = 1, \ldots, N$, and $c_i \in \{1, \ldots, D\}$.
Step 3: Compute $P_{\mathrm{B}}(k)$ and find k^* for $k = 1, \ldots, K$, and relabel the cells accordingly.

We would like to point out that the number of terms to compute and compare drops from D^N to 2^N after simplification, and further to N after ordering.

3.5. Results

The algorithm has been intensively tested at IRIS, Inc. on the specimens obtained at Harbor UCLA Medical Center. We compare the performances with and without using contextual information on blood samples from 220 specimens (consisting of 13,200 cells). In about 50% of the cases, a false alarm would have occurred had context not been used. Most cells are correctly classified, but a few are incorrectly labeled as immature cells, which raises a flag for the doctors. Change of the classification of the specimen to abnormal requires expert intervention before the false alarm is eliminated, and it may cause unnecessary expenses and worry. When context is applied, the false alarms for most of the

Table 4
Comparison of with and without using contextual information

Methods	Cell classification	Normality identification	False positive	False negative
No context	88%	~50%	~50%	0%
With context	89%	~90%	~10%	0%

specimens were eliminated, and no false negative was introduced (Table 4).

4. Incorporating context into urinalysis

4.1. Introduction

Urine is one of the most complex body fluid specimens: it potentially contains about 60 meaningful types of elements. Examination of the urine sediment plays a critical role in urinalysis. It detects the presence of elements that often provide early diagnostic information concerning dysfunction, infection, or inflammation of the kidneys and urinary tract. Thus this non-invasive technique can be of great value in clinical case management. Traditional microscopic urinalysis systems rely on human operators who read the samples visually and identify them, and thus is time consuming, labor intensive and difficult to standardize. Automated microscopy of all specimens is more practical than manual microscopy because it eliminates variation among different technologists. This becomes more pronounced when the same technologist examines increasing numbers of specimens. Also, it is less labor intensive and thus less costly than manual microscopy. It also provides more consistent and accurate results.

An automated urinalysis system workstation (The *YellowIRIS*™, IRIS, Inc.) has been introduced in numerous clinical laboratories for automated microscopy. Urine samples are processed and examined at ×100 (low power field) and ×400 magnifications (high power field) with bright-field illumination. The *YellowIRIS*™ automated system collects video images of formed elements in a stream of uncentrifuged urine passing an optical assembly. These images are given to a computer algorithm for automatic identification.

Among the elements (analytes) found in microscopic urinalysis are various casts, epithelial cells, blood cells (including both white and red blood cells), crystals, as well as other elements including bacteria, yeast. Fig. 4 shows some analyte examples. Some of the analytes found in urine are pathological. There is a range of counts of these analyte types within which a specimen is considered normal. A specimen is abnormal if the counts of one or more these types fall out of their ranges.

Context is rich in urinalysis and plays a crucial role in analyte classification [27]. Some combinations of reasonable analytes are more likely than others. For

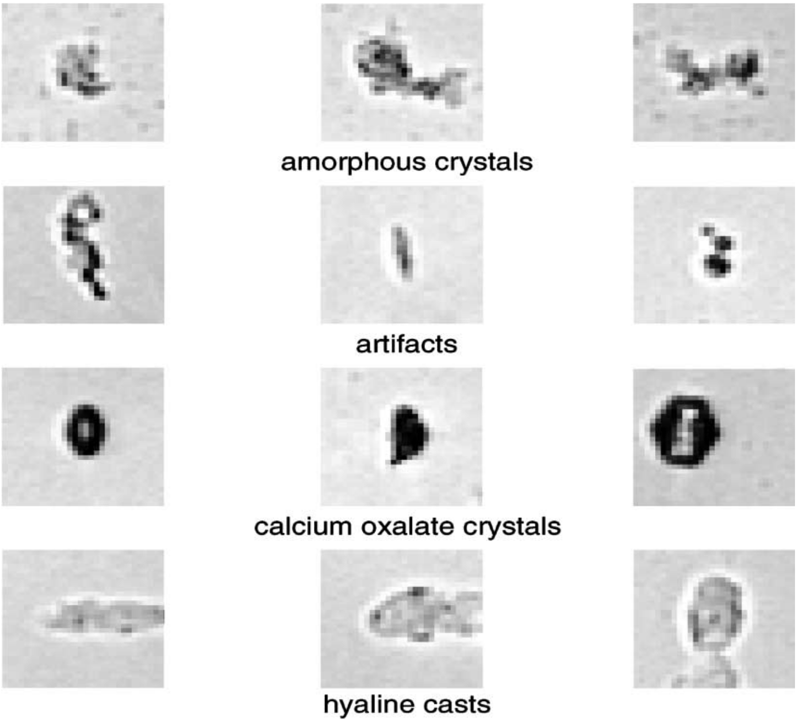

Fig. 4. Examples of some of the analyte images.

Table 5
Features extracted from urine anylates images

Feature number	Feature description
1	Area
2	Length of edge
3	Square root of area/length of edge
4	(Standard deviation/mean) of distance from center to edge
5	λ_1/λ_2
6	Sum of length of two longest straight edges/ total length of edge
7	Sum of length of four longest straight edges/ total length of edge
8	Sum of length of two longest semi-straight edges/ total length of edge
9	Sum of length of four longest semi-straight edges/ total length of edge
10	The mean of red distribution
11	The mean of blue distribution
12	The mean of green distribution
13	15th percentile of gray level histogram
14	85th percentile of gray level histogram
15	The standard deviation of gray level intensity
16	Energy of the Laplacian transformation of gray level image

instance, the presence of bacteria indicates the presence of WBCs, since bacteria tend to cause infection and thus trigger the production of more WBCs. Squamous epithelial cells can appear both flat or rolled up. If squamous epithelial cells in one form are detected, then it is likely that there are squamous epithelial cells in the other form. WBC clusters in the low power field usually indicate WBCs in high power field. Utilizing such context will hopefully improve classification accuracy.

The task of automated microscopic urinalysis is, given a urine specimen that consists of up to a few hundred images of analytes, to classify each analyte into one of the classes.

Similar to the WBC identification task discussed in the previous chapter, the automated urinalysis consists of the following steps: image processing and feature extraction, learning and pattern recognition, and context incorporation. The first two steps are very similar to that of the WBC identification, therefore these details will not be discussed. Table 5 gives a list of features extracted from analyte images.[3] The classes we are looking at are artifacts, bacteria, calcium oxalate crystals, red blood cells, WBCs, budding yeast, amorphous crystals, and uric acid crystals. All these analytes are in the high power field.

The form of context in urinalysis, especially the fact that context is contained in the presence of some types of analytes, makes it well suited for the framework discussed in Section 2.2. The context A is the presence of several relevant classes. The criteria for relevance will be discussed in Section 4.2. The maximum posterior decision rule chooses class label $\hat{c}_i$ for element i such that $\hat{c}_i = \operatorname{argmax}_{c_i} p(c_i|\mathbf{x}_i, A)$.

4.2. Identification of relevant classes

Not all classes are relevant in terms of carrying contextual information. We propose three criteria based on which we can systematically investigate the relevance of the class presence.

The first criterion is the correlation coefficient between the presence of any two classes. One type of analyte is considered relevant to another if the absolute value of their correlation coefficient is beyond a certain threshold. The graph in Fig. 5 illustrates the relevance between any two analyte types according to various thresholds. In this figure, two types are related or relevant to each other only if their nodes are connected by a line. The solid lines correspond to threshold 0.25 and the added dotted lines to 0.10. Not surprisingly, lowering the threshold leads to more relevant classes. It shows that uric acid crystals, budding yeast and calcium oxalate crystals are not relevant to any other types even by a generous threshold of 0.10.

The second criterion is the classical mutual information $I(c; A_d)$ between the presence of a class A_d and the class probability $p(c)$. The bigger the mutual information between the presence of a class and the class distribution, the more relevant this class is. Ranking the analyte types in terms of $I(c; A_d)$ in a descending manner gives rise to the following list: bacteria, amorphous crystals, artifact, red blood cells, WBCs, uric acid crystals, budding yeast and calcium oxalate crystals. The relevance level decreases in the list.

The third criterion is what we call the *expected relative entropy* $D[c||A_d]$ between the presence of a class A_d and the labeling probability $p(c)$, defined as

$$D[c||A_d] = P(A_d = 1)D[p(c)||p(c|A_d = 1)] + P(A_d = 0)D[p(c)||p(c|A_d = 0)]$$

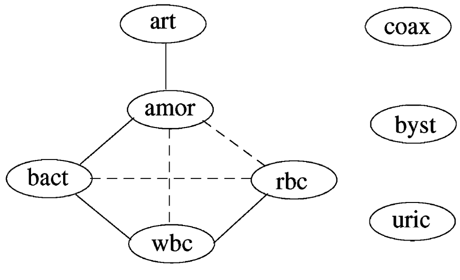

Fig. 5. Relevant classes.

[3] λ_1 and λ_2 are respectively the bigger and the smaller eigenvalues of the second moment matrix of an image.

where

$$D[p(c)||p(c|A_d=1)] = \sum_{i=1}^{D} p(c=i|A_d=1)\ln\left(\frac{p(c=i|A_d=1)}{p(c=i)}\right)$$

and

$$D[p(c)||p(c|A_d=0)] = \sum_{i=1}^{D} p(c=i|A_d=0)\ln\left(\frac{p(c=i|A_d=0)}{p(c=i)}\right)$$

Similarly, ranking the analyte types in terms of $D(c||A_d)$ in a descending manner gives rise to the following list: bacteria, artifact, red blood cells, amorphous crystals, WBCs, calcium oxalate crystals, budding yeast and uric acid crystals.

Thresholding correlation coefficient explores the pairwise relevance of classes, whereas mutual information and expected relative entropy indicate the general relevance of a class to all other classes in an expectation sense. All three criteria lead to similar conclusions regarding the relevance of all classes.

4.3. The algorithm

Once we identify the M relevant classes, we use the following algorithm to incorporate partial context.

Step 0: Estimate for the database the priors $p(c|A_d)$ and $p(c)$, for $c \in 1,2,\ldots,D$ and $d \in 1,2,\ldots,D$.

Step 1: For a given $\mathbf{x}_i$, compute $p(c_i|\mathbf{x}_i)$ for $c_i = 1,2,\ldots,D$, which are the outputs of the trained neural network.

Step 2: Let the M relevant classes be $R_1,\ldots,R_M$. According to the no context $p(c_i|\mathbf{x}_i)$ and certain criteria for detecting the presence or absence of all the relevant classes, get $A_{R_1},\ldots,A_{R_M}$.

Step 3: Let $p(c_i|\mathbf{x}_i,A_0) = p(c_i|\mathbf{x}_i)$, where A_0 is the null element. Then, for $m = 1$–M, iteratively compute$p(c_i|x_i;\underline{A_0,\ldots,A_{R_{m-1}},A_{R_m}}) = p(c_i|x_i,\underline{A_0,\ldots,A_{R_{m-1}}})\times (p(c_i|A_{R_m})p(A_{R_m}))/p(c)$.

Step 4: Label the objects according to the final context-containing $p(c_i|\mathbf{x}_i,\underline{A_{R_1},\ldots,A_{R_M}})$, i.e., $\hat{c}_i = \underset{c_i}{\operatorname{argmax}}\, p(c_i|\mathbf{x}_i,\underline{A_{R_1},\ldots,A_{R_M}})$ for $i = 1,\ldots,N$.

This algorithm is invariant with respect to the ordering of the M relevant classes in $(A_1,\ldots,A_M)$.

4.4. Results

The algorithm using partial context was tested on a database with 83 urine specimens that contains 20,276 analyte images. Four classes are considered relevant according to the criteria described in Section 4.2: bacteria, red blood cells, WBCs and amorphous crystals. We measure two types of error: analyte-by-analyte error, and specimen diagnostic error. The error means and standard deviations are derived from 50 bootstrap samples of 75 specimens out of the original set of 83 specimens. The average analyte-by-analyte error is reduced from 44.48% before using context to 36.66% after, resulting a relative error reduction of 17.6% (Table 6). The diagnosis for a specimen is either normal or abnormal. Tables 7 and 8 compare the diagnostic performance with and without using context, and Table 9 lists the relative changes. We can see using context significantly increases correct diagnosis for both normal and abnormal specimens, and reduces both false positive and false negative.

Table 6
Comparison of using and not using contextual information for analyte-by-analyte error

	Without context	With context
Average element-by-element error	44.48 ± 1.14%	36.66 ± 0.97%

Table 7
Diagnostic confusion matrix not using context

	Estimated normal	Estimated abnormal
Truly normal	40.96 ± 1.69%	7.23 ± 0.94%
Truly abnormal	19.28 ± 1.18%	32.53 ± 1.80%

Table 8
Diagnostic confusion matrix using context

	Estimated normal	Estimated abnormal
Truly normal	42.17 ± 1.89%	6.02 ± 0.80%
Truly abnormal	16.87 ± 1.15%	34.94 ± 1.85%

Table 9
Relative accuracy improvement (diagonal elements) and error reduction (off diagonal elements) in the diagnostic confusion matrix by using context

	Estimated normal	Estimated abnormal
Truly normal	+2.95 ± 1.21%	−16.73 ± 7.2%
Truly abnormal	−12.50 ± 2.65%	+7.41 ± 1.67%

5. Discussions

This paper has addressed the question of contextual information fusion. A straight-forward use of compound Bayesian theory is theoretically elegant and is optimal in terms of error probability and information gain. When applied to the problem of WBC image recognition, it significantly reduces false alarm rate and thus greatly reduces the cost due to expensive clinical tests. However, its exponential computational complexity makes it intractable for many real world problems. An effective and yet computationally linear method has been formulated. This approach explicitly derive contextual information, and fuse it with the measurements of the object of interest. It is object-centered, and naturally leads to an iterative procedure which typically converges in a few rounds. When applied to the problem of microscopic urinalysis, it significantly improves correct classification rate and reduces false alarm as well as false negative rate.

Since for the second approach we need to explicitly derive context, we are faced with the issue of identifying contextual relevant variables and defining relevancy measures. Aside from the ones proposed in this paper, another approach to identify contextual relevancy is by learning the structure of a Belief Network from data [2–4,7,11,14,20,22]. This is among the future topics for this research. Other future topics include extracting contextually relevant features, and identifying hidden context (e.g., speaker gender for speech recognition).

Acknowledgements

This work was partially supported by NSF IIS-0070842 and NASA NCC-2-1218.

References

[1] Boehringer-Mannheim-Corporation. Urinalysis Today, 1991.

[2] W. Buntine, A guide to the literature on learning probabilistic networks from data, IEEE Transactions on Knowledge and Data Engineering 8 (1996) 195–210.

[3] J. Cheng, D.A. Bell, W. Liu. Learning belief networks from data: An information theory based approach, in: Proceedings of the 6th ACM international Conference on Information and Knowledge management, 1997.

[4] G.F. Cooper, E. Herskovits, A Bayesian method for the induction of probabilistic networks from data, Machine Learning 9 (1992) 309–347.

[5] T.M. Cover, J.A. Thomas, in: Elements of Information Theory, Wiley Series in Telecommunications, Wiley, 1991.

[6] O.D. Faugeras, M. Berthod, Improving consistency and reducing ambiguity in stochastic labeling: An optimization approach, IEEE Transactions on Pattern Analysis and Machine Intelligence 3 (1981) 412–424.

[7] N. Friedman, M. Goldszmidt, Learning Bayesian networks with local structure, in: M.I. Jordan (Ed.), Learning in Graphical Models, Kluwer, Dordrecht, Netherlands, 1998.

[8] S.I. Gallant, A practical approach for representing context and for performing word sense disambiguation using neural networks, Neural Computation 3 (1991) 293–309.

[9] R.M. Haralick, H. Joo, A context classifier, IEEE Transactions on Geoscience and Remote Sensing 24 (1986) 997–1007.

[10] M. Harries, C. Sammut, K. Horn. Extracting hidden context. Technical report, School of Computer Science and Engineering, University of NSW, 1997. UNSW-CSE-TR-9708.

[11] D. Heckerman, D. Geiger, D.M. Chickering, Learning Bayesian networks: The combination of knowledge and statistical data, Machine Learning 20 (1995) 197–243.

[12] J. Illingworth, J. Kittler, Optimization algorithms in probability relaxation labeling, in: P.A. Devijver, J. Kittler (Eds.), Pattern Recognition Theory and Applications, Springer-Verlag, 1987, pp. 109–117.

[13] F. Jelinek, Statistical Methods for Speech Recognition, MIT Press, 1997.

[14] M.I. Jordan (Ed.), Learning in Graphical Models, The MIT Press, 1998.

[15] H.K. Kasdan, J.P. Pelmulder, L. Spolter, G.B. Levitt, M.R. Lincir, G.N. Coward, S.I. Haiby, J. Lives, N.C.J. Sun, F.H. Deindoerfer, The *WhiteIRIS*$^{\mathrm{TM}}$ leukocyte differential analyzer for rapid high-precision differentials based on images of cytoprobe-reacted cells, Clinical Chemistry 40 (1994) 1850–1861.

[16] A.J. Katz, M.T. Gately, D.R. Collins, Robust classifiers without robust features, Neural Computation 2 (1990) 472–479.

[17] J. Kittler, Relaxation labelling, in: P.A. Devijver, J. Kittler (Eds.), Pattern Recognition Theory and Applications, Springer-Verlag, Berlin, 1987, pp. 99–108.

[18] J. Kittler, Probabilistic relaxation: Potential, relationships and open problems, in: M. Pelillo, E.R. Hancock (Eds.), International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, Venice, Italy, 1997, pp. 393–408.

[19] J. Kittler, J. Illingworth, Relaxation labelling algorithms—a review, Image and Vision Computing 3 (1985) 206–216.

[20] W. Lam, F. Bacchus, Learning Bayesian belief networks: An approach based on mdl principle, Computational Intelligence 10 (1994) 269–293.

[21] R.R Murphy, Biological and cognitive foundations of intelligent sensor fusion, IEEE Transactions on Systems, Man and Cybernetics 26 (6) (1998) 42–51.

[22] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference, Morgan Kaufmann, 1988.

[23] L.R. Rabiner, A tutorial on hidden Markov models, Proceedings of the IEEE 73 (1989) 1349–1387.

[24] A. Rosenfeld, R. Hummel, S. Zacker, Scene labeling by relaxation operations, IEEE Transactions on Systems Man and Cybernetics 6 (1976) 420–433.

[25] J. Sherrah, S. Gong, Exploiting context in gesture recognition, in: P. Bouquet, L. Serafini, P. Brezillon, M. Benerecetti, F. Castellani (Eds.), CONTEXT'99, LNAI 1688, Springer, 1999, pp. 515–519.

[26] X.B. Song, Y. Abu-Mostafa, J. Sill, H. Kasdan, Incorporating contextual information into white blood cell recognition, in: M.I. Jordan, M.J. Kearns, S.A. Solla (Eds.), Advances of Neural Information Processing, vol. 10, MIT Press, 1997, pp. 950–956.

[27] X.B. Song, Y. Abu-Mostafa, J. Sill, H. Kasdan, Image recognition in context: Application to microscopic urinalysis, in: S.A. Solla, T.K. Leen, K.-R. Muller (Eds.), Advances of Neural Information Processing, vol. 12, MIT Press, 1999, pp. 963–969.

[28] T.M. Strat, Employing contextual information in computer vision, in: DARPA Workshop on Image Understanding, 1994, pp. 217–229.

[29] B. Tian, M.A. Shaikh, M.R. Azimi-Sadjadi, T.H. Vonder Haar, D. Reinke, A study of cloud classification with neural networks using spectral and textual features, IEEE Transactions on Neural Networks 10 (1999) 138–151.

[30] G. Toussaint, The use of context in pattern recognition, Pattern Recognition 10 (1978) 189–204.

[31] D.P. Turney, Exploiting context when learning to classify, in: Proceedings of the European Conference on Machine Learning, Springer-Verlag, Vienna, Austria, 1993, pp. 268–276.

[32] R.L. Watrous, Context-modulated vowel discrimination using connectionist networks, Computer Speech and Language 5 (1991) 341–362.