Modeling Context as Statistical Dependence

Sriharsha Veeramachaneni[1], Prateek Sarkar[2], and George Nagy[3]

[1] SRA Division, ITC-IRST
Povo, TN 38057, Italy
sriharsha@itc.it
[2] Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304, USA
psarkar@parc.com
[3] ECSE Dept. Rensselaer Polytechnic Institute
110 Eighth Street, Troy, NY 12180, USA
nagy@ecse.rpi.edu

Abstract. Theories of context in logic enable reasoning and deduction in contexts represented as formal objects. Such theories are not readily applicable to systems that learn by induction from a set of examples. Probabilistic graphical models already provide the tools to exploit context represented as statistical dependences, thereby providing a unified methodology to incorporate context information in learning and inference. Drawing on a case study from optical character recognition, we present the various types of dependences that can occur in pattern classification problems and how such dependences can be exploited to increase classification accuracy. Learning under different conditions require differing amounts and kinds of samples and different trade-offs between modeling error due to overly strict independence assumptions and estimation error of models that are too elaborate for the size of the available training set. With a series of examples based on frames of two patterns we show how each kind of dependence can be represented using graphical models and present examples from other disciplines where the particular dependence frequently occurs.

1 Introduction

A recognition problem often pertains to the interpretation of a collection of observations in a scene. A simple approach to the solution is to interpret/classify each observation independently of others. In reality observations and their interpretations are often interdependent. This interdependence is manifested as high likelihood of occurrence of some combinations of observations and interpretations, and relative rarity or improbability of some others. Modeling the interdependence of observations and interpretations can improve the accuracy of recognition. Therefore it can be argued that scene understanding (whatever that means for the application) can be better performed by processing each object in the *context* of the others in the scene rather than independently [25]. However, it is not practical in most application domains to respect arbitrary interdependence of observations and interpretations in formulating a solution to the recognition problem. Choosing a good dependence model can improve the efficacy of learning (e.g., convergence, robustness) from available data, as well as the efficiency (e.g., accuracy, speed) of interpretation. Optical Character Recognition (OCR) is one application area where extensive work has been done in modeling and exploiting the relationships between the objects (in this case images of characters) to be classified and their intended interpretations (class labels) [14].

The recognition of a symbol, signal or object can be done in isolation by using a classification rule that has been learnt from labeled samples available for 'training' by operating on the observable data ('features') about each object (or pattern). Higher classification accuracy can be achieved by the recognition of groups of related objects, taking into account the relationship between them in addition to their individual characteristics. The cause of the relationship may be multiple observations of the same characteristic ('feature dependence'), temporal or spatial contiguity (i.e., co-articulation, ligatures, alignment), constraints imposed by a message ('language context'), or a common source ('style'). The additional information that is derived from the co-occurring objects is often called 'context.'

In logical approaches to AI, contexts are modeled as abstract mathematical entities whose values determine the values of other logical entities [12, 24]. In a setting where we have to induce general truths from thousands of noisy examples, each characterized by tens or hundreds of attributes, a statistical approach seems convenient. Moreover for the interpretation of a collection of objects, the context sometimes cannot - and often need not - be explicitly articulated as long as it is understood that one exists. In such scenarios the dependence among entities or more precisely, between the representations of the objects, can be modeled by dependent parameterized statistical distributions. Multivariable statistics provide a natural way to induce the effect of context on the interpretation of objects without the need for its explicit representation.

A close look reveals several different types of relationships between features, labels, and sources. From a statistical perspective, the joint distribution of all the variables specifies the problem completely. Because of the large number of variables, the joint distribution may be difficult to estimate accurately, and the underlying relationships are obscured rather than revealed by a model that is too elaborate for the application. We can overcome this problem by bringing to bear prior knowledge about the problem at hand to avoid modeling relationships that can be neglected.

Directed graphical models (of which Bayesian Networks are a special case) offer a way to represent even complex models by avoiding the specification of conditional independence relations. Recent advances in graphical models provide rules for efficient computation of both the joint distributions (learning), and of the conditional distributions required for classification (inference). We show how such models provide a systematic representation and efficient computational tools for the classification of groups of patterns under diverse contextual assumptions.

We present the equations and graphical models for a variety of contextual classification methods that have already found application in practice. We study the simple problem of classifying a pair of objects (patterns) illustrating with equations and examples how different types of context can be modeled. We also comment on the implications of context on the acceptable sample size of the training set, and give examples of past and future applications of broadly defined context.

Section 2 provides a brief introduction to probabilistic graphical models and a short summary of previous work. In Section 3, we show how differing assumptions on the statistical dependence affect underlying joint distributions and where such assumptions are justified. Section 4 discusses methods to learn (i.e., estimate the parameters for) the various models. Section 5 gives some examples of the expected dependence in diverse applications. Many of these are based on OCR (Optical Character Recognition) and ASR (Automatic Speech Recognition) because some of the more complex models were developed there. In the final section we demonstrate how quickly the number of parameters grows with larger frames (i.e., more patterns per frame), more features, and more classes, and discuss the trade-offs between model complexity and sample size.

2 Probabilistic Graphical Models

In probabilistic analysis with many random variables all the marginal and conditional probabilities on a subset of the variables can be computed from the joint distribution if it is known. However often in practice the joint distribution has to be estimated from (relatively) sparse data, and it is essential that we impose restrictions on the nature of the joint distributions to make this feasible.

On the other hand if all the variables are mutually independent, then the marginal distribution on each variable is simpler to estimate, and the joint distribution is simply the product of the marginals. Under such assumptions, no variable conveys information about any other variable. Useful probabilistic models often lie between these two extremes. Graphical models have emerged as an interesting class of models that constrain (or simplify) the joint density via conditional independence relations among variables which are represented by nodes in the model. The set of edges embody information about conditional independence among the variables. The graphical model formalism, drawing from graph theory and probability theory, provides a unifying framework for classical and new models such as mixture models, factor analysis, hidden Markov models, Kalman filters, Bayesian networks, Markov random fields, Ising models, and conditional random fields. [11, 13, 7] provide a good overview of graphical models.

Bayesian networks, pioneered by Pearl [15], are a class of directed graphical models where the joint distribution of random variables are completely specified by functions (or tables) stored in the nodes, which represent the distribution of the node variable when conditioned on its parents. Edge direction in these networks can be interpreted as 'causation' providing a basis for the design of the graph structure in any problem domain [16]. We recommend [15] for an excellent philosophical and mathematical foundation for Bayesian networks. There are many web tutorials for a more urgent introduction to the standard notations and diagram conventions.

Once the underlying graphical model and its parameters are known, most applications involve inference, i.e., the computation of a conditional or marginal distribution from the known joint distribution, such as $P(\text{character-label} = A|\text{bitmap} = \textit{observation})$. The graphical structure of the model leads to efficient inference algorithms, such as variable elimination [1], local message passing (for directed acyclic graphs) [15], or the junction tree algorithm when there are cycles in the graph. Heckerman [10] has written an excellent tutorial on exact inference. Often exact inference (exact reduction of the joint probability function) is computationally infeasible, requiring the use of approximate algorithms such as variational methods, Monte Carlo sampling methods and loopy belief propagation [11].

There is also extensive literature on learning the parameters of a graphical model. Jordan [11] provides a good review of previous work. In situations where the graphical structure is known a priori, the parameters of the distributions at each node are estimated either from fully observable data (direct maximum likelihood estimates) or from partially observed data (maximum likelihood estimates through Expectation-Maximization or gradient ascent). Learning the model structure from data is also an active research area.

3 Modeling context as statistical dependence

We consider the following pattern classification problem. The patterns to be classified arrive as ordered sets $z = (x_1, x_2)$ of two patterns each. Each pattern has a class label y which is one of two classes A or B. Consequently the ordered-pair pattern z has a class label (y_1, y_2). Each pattern is composed of two features $x_i = (u_i, v_i)$ that are used for classification. Given the observed pair (x_1, x_2) the objective is to classify it by assigning to it the appropriate pair label (y_1, y_2). We introduce the notion of a frame $F = (x_1, y_1, x_2, y_2) = (u_1, v_1, y_1, u_2, v_2, y_2)$ which is the ordered set of patterns and the corresponding class labels.

$F = (x_1, y_1, x_2, y_2)$ (each frame consists of two patterns and their labels)
$x_i = (u_i, v_i)$ (each pattern has a label and two features)

When the joint probability distribution over all possible frames $P(x_1, y_1, x_2, y_2)$ is completely known, the classifier with the highest accuracy (the maximum a posteriori classifier) can be constructed that assigns the class $(y_1^\star, y_2^\star)$ to the frame (x_1, x_2) where

$$(y_1^\star, y_2^\star) = \underset{(y_1, y_2)}{\operatorname{argmax}} P(x_1, y_1, x_2, y_2) \tag{1}$$

3.1 Complete Dependence

One way to exploit context during learning and inference is to make no assumptions of statistical independence a priori but to learn, under the most general model, the absence of such dependence relations from the training data. This method, however, often results in an inaccurate model because of the enormous amount of training data required to learn it accurately. It is therefore necessary to make the appropriate independence assumptions justified by the amount of data available as well as by prior knowledge about the problem.

3.2 No Dependence

The simplest independence model can be obtained by assuming mutual independence of the variables except for the dependence of each feature on the label of the corresponding pattern (without which classification would be impossible). Then the joint probability can be written as

$$\begin{aligned} P(x_1, y_1, x_2, y_2) &= P(u_1, v_1, y_1, u_2, v_2, y_2) \\ &= P(u_1|y_1)P(v_1|y_1)P(u_2|y_2)P(v_2|y_2)P(y_1)P(y_2) \qquad (2) \end{aligned}$$

The probabilistic graphical model depicting the assumption of no dependence is shown in Figure 1. The way to read the graph is to note that given the class label (y_i) of the pattern the pattern features (u_i, v_i) are statistically independent. The assumption of no dependence leads to the so-called *naive Bayes* classifier which often performs surprisingly well, especially in situations where there is little training data available.

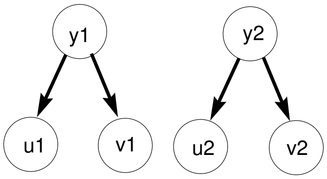

Fig. 1. Graphical model - No dependence.

3.3 Intra-pattern Class-conditional Feature Dependence

Here we drop the assumption of independence between the features u_i and v_i given the class-label y_i and therefore the complete joint distribution $P(u_i, v_i|y_i)$ has to be specified. Such dependence can be incorporated into the graphical model as shown in Figure 2. Such intra-pattern dependence is generally well understood and is discussed in most texts for binary and Gaussian variables [6, 22, 23]. The covariance matrix represents only the second-order (pairwise) dependence between the feature variables but suffices to completely specify the interdependence of Gaussian random variables regardless of dimensionality. It is possible to construct examples where taking correlation into account leads to zero error rate even when classification based only on an independence assumption is no better than a random choice.

When there are a large number of features, it may be desirable to exploit only the dependence of highly correlated features. Partial independence methods are based on dependence chains or dependence trees of features [3, 4, 6].

A different way of modeling dependence is by means of mixture distributions, which are most often used in the multidimensional case for Gaussian distributions [6, 19]. Even if the features are class-conditionally independent within each component distribution, they may be dependent in the class-conditional mixture distribution.

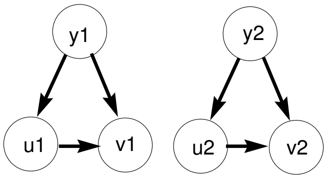

Fig. 2. Graphical model for intra-pattern class-conditional feature dependence.

3.4 Inter-pattern Class Dependence

Inter-pattern class dependence is often called *linguistic context* in OCR. When such dependence is modeled the frame distribution is given by

$$P(x_1, y_1, x_2, y_2) = P(x_1|y_1)P(x_2|y_2)P(y_1, y_2) \tag{3}$$

That is, the joint class probability is not completely described by the marginal probabilities. Figure 3 shows an example of two sequences of patterns to be recognized by a word recognizer. Although, the first three patterns are identical in both the words, knowing that the labels for the last two patterns are either ‘ks’ or ‘AD’ makes either the label ‘boo’ or ‘600’ more likely for the first three patterns. The graphical representation of inter-pattern class dependence is shown in Figure 4.

3.5 Inter-pattern Class-Feature Dependence

In some classification problems, features of a particular pattern depend on the co-occurring classes (but not necessarily on the features of co-occurring patterns).

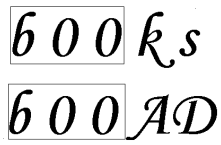

Fig. 3. Example of linguistic context in OCR.

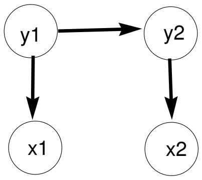

Fig. 4. Graphical model for inter-pattern class dependence.

An example of such context is shown in Figure 5. The vertical location of the apostrophe depends on whether the preceding letter was a 'T' or an 'n', but not on the particular rendering of the preceding letter. That is, the features of the apostrophe are independent of the preceding letter given its label. The graphical model that captures this kind of context is shown in Figure 6.

Such dependence can be modeled graphically as shown in Figure 6 and the joint frame distribution is given by

$$P(x_1, y_1, x_2, y_2) = P(x_1|y_1, y_2)P(x_2|y_1, y_2)P(y_1)P(y_2) \tag{4}$$

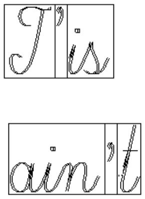

Fig. 5. Example of inter-pattern class-feature dependence.

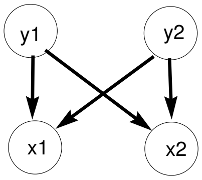

Fig. 6. Graphical model for inter-pattern class-feature dependence.

3.6 Inter-pattern Feature Dependence

(a) There are several reasons for the dependence between features of patterns in a field. In speech the articulation of a pattern, and in cursive handwriting the rendering of a pattern often determine the features of the next pattern. For example, in Figure 7 the shape of the letter 'a' depends on the shape of the preceding letter. This type of dependence can be modeled as shown in Figure 8.
(b) Another cause for such dependence is the *isogeny* (commonality of origin or source) of pattern groups. For example, handwriting or typeface consistency can lead to such a dependence in recognition problems with multiple writers or

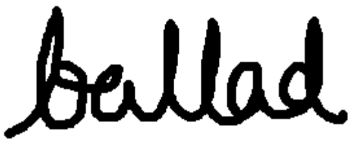

Fig. 7. Example of ligatures in OCR.

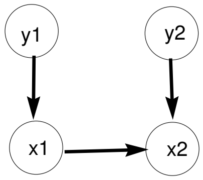

Fig. 8. Graphical model for inter-pattern feature dependence (ligatures, co-articulation etc.)

typefaces [20, 26]. Figure 9 shows the handwriting of two hypothetical writers A and B. The ‘7’ of writer A is identical to the ‘1’ of writer B. In such a scenario, no classifier operating on individual patterns can discriminate between ‘7’ from writer A and ‘1’ from writer B. However, when the test patterns appear as pairs written by the same writer, then the above confusion can be resolved using the context of the co-occurring pattern, if the other pattern in the field is a ‘1’ from writer A or a ‘7’ from writer B. Of course when the identity of the writer is known during classification such context is irrelevant, but this information is often lacking. Such a situation can be modeled as shown in Figure 10 where the variable h represents the hidden writer identity. The probabilistic graphical model for such dependence and the corresponding joint distributions is given by

$$P(x_1, y_1, x_2, y_2) = \sum_h P(x_1|y_1, h)P(x_2|y_2, h)P(y_1)P(y_2)P(h) \qquad (5)$$

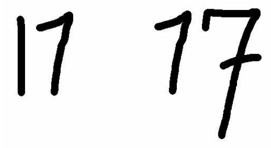

Fig. 9. Example of inter-pattern feature context due to styles.

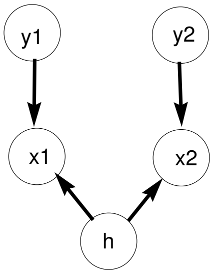

Fig. 10. Graphical model for inter-pattern feature dependence (style.)

4 Parameterization and Learning

In the previous section we showed how the various varieties of context can be visualized with probabilistic graphical models (*qualitative specification*). Now, in order to completely specify the models we need to choose the appropriate parameterization, i.e., to choose the actual families of distributions for the various dependences (*quantitative specification*). Once the model is completely specified we can estimate the parameters from the available training data (i.e., learning) in order to construct the maximum a posteriori classifier (cf. Equation 1).

In most pattern recognition applications, labeled training samples are expensive and scarce. Therefore one always faces a small-sample estimation problem in deriving the model parameters [18]. The number of parameters to be estimated increases rapidly with the number of variables that are considered dependent. Therefore most models avoid considering dependence beyond the second order (pairwise correlation). Even second-order models require estimating a number of parameters proportional to the square of the number required under independence assumptions.

Although there exist general techniques for learning the parameters for probabilistic graphical models, we discuss some specific techniques that are used for estimation of parameters for the models presented above. Since we cannot go into the intricacies of statistical parameter estimation, we indicate the major families of techniques for learning for the pattern pair recognition problem considered in the previous section and give references to more detailed sources.

4.1 Independent Variables

When the class-conditional feature distributions are modeled by a Bernoulli distribution, the necessary parameters are obtained by averaging feature values over the training samples. For Gaussian features, the sum of the squares of the feature values are also needed to obtain estimates of the class-conditional variances.

4.2 Intra-pattern Feature Dependence

The parameters of second-order models are the means and covariance matrices (or, equivalently, correlation coefficients) of the feature variables. The elements of the covariance matrix for two variables are the average values of the product of the variables minus the product of their individual averages over the training samples. The sample variances and covariances are usually scaled by $n/(n-1)$, where n is the number of samples, to obtain an unbiased estimate. The expected value of an unbiased estimator over different training sets is equal to the value of the estimated parameter [2]. When the training sample size is small the variance in the estimate of the covariance matrix can be reduced by increasing its bias by a smoothing technique called *regularization* [8].

Estimation is more difficult for mixture models if the mixture components are not explicitly indicated in the training sample. For Gaussian mixtures, algorithms and computer code based on Expectation Maximization [5, 19] are available. This is an iterative two-phase process. In the first phase, all the patterns are assigned probabilistic labels (called hidden variables) that indicate their affinity to each mixture component. In the second phase, the parameters of each mixture component are estimated from the samples weighted according to the coefficients determined in the first phase. This process converges to the population parameters of the mixture model under broad conditions.

4.3 Class-label Dependence

The transition probabilities of a first-order Markov model can be estimated directly from the sample bigram frequencies. In our example, it would suffice to count the number of occurrences of AA and BB in the training set if we assume the Markov chain to be in steady state. Some care is necessary for estimating the probabilities of very low or (rarely) very high frequency samples. *Laplace smoothing* is often performed to obtain robust estimates.

For the lexical approach, it is necessary to estimate the probabilities of all possible frame-labels. For long strings, the above caveat for low-frequency samples is even more important.

4.4 Inter-pattern Feature Dependence

For Markov models, we can follow the same route as in 4.3. If the features are class-conditionally independent, then we need to count only the frequencies of n-grams of corresponding features. If they are dependent, we must count n-grams of feature vectors.

When the styles are labeled in the training set, the feature probabilities for each style can be estimated as in 4.1 and 4.2. If not style-labeled, we can use Expectation Maximization for certain families of feature distributions. The mixture components for the class-conditional features are estimated from all the frames that include any pattern with the label for which we are estimating the features. Only the style variables are hidden.

We have not mentioned HMM (Hidden Markov Models) because they are most useful for unsegmented patterns of unknown length or duration, like cursive writing and speech, rather than discrete patterns. HMM is based on a hidden (unobservable) Markov state sequence. Only some transitions are permitted: in handwriting and speech, the network topology is usually restricted to left-to-right with self transitions. Features are generated in each state according to predetermined but unknown class-conditional probability distributions. Methods have been developed to estimate the (hidden) transition probabilities and the state-dependent feature probabilities from unsegmented sample sequences [17]. Most of these methods are specialized formulations of Expectation Maximization. The corresponding frame classification is based on Dynamic Programming (the Viterbi algorithm) to determine the most likely sequence of labels in the frame. Other algorithms such as the A^* search and Stack search are also used.

Table 1. Number of parameters to be estimated for various dependence models when the class-conditional feature distribution is assumed Gaussian.

Dependence	Number of parameters
None	CF means, CF variances $C-1$ class-probabilities
Intra-pattern feature dependence only	CF means, $CF(F+1)/2$ covariances $C-1$ class-probabilities
Class-label dependence only	C^L-1 frame-label probabilities CF means, CF variances
Inter-pattern feature dependence (style only)	$K-1$ style probabilities KCF means, KCF variances $C-1$ class probabilities

C-Number of classes; F-Number of features; K-Number of styles; L-Frame length

Table 1 shows how the number of parameters to be estimated depends on the number of patterns per frame, the number of classes, and the number of features under the various assumptions on the dependence structure. Increasing the complexity of the models increases the number of samples necessary for accurate estimation of the model parameters. Consequently better modeling of dependence (or the joint distributions) does not necessarily imply an improvement in classification accuracy. The choice of model complexity must be guided by considerations of relative gain in performance, and the sample size necessary to train the model. We need to carefully choose the appropriate dependence model given the size and type of training data available and from prior knowledge about the problem.

5 Some Applications

5.1 Intra-pattern Feature Dependence

It is difficult to find features that are truly uncorrelated. For instance, the width and height of blob patterns and alphabetic symbols, and even their geometric moments, tend to be correlated. So are the short-term frequency components of phonemes, the height and trunk-diameter of trees, and the multispectral reflection coefficients of vegetation.

An interesting example of feature dependence is induced during digitization by the random phase of the spatial sampling grid relative to a blob pattern [9]. The pixels on each edge are positively correlated (they appear or disappear together), while pixels on opposite edges are negatively correlated [21].

Sometimes principal components (also called Karhunen Loeve expansions or Hotelling transform) are extracted in an attempt to obtain orthogonal (uncorrelated) features, but this leads only to features that are overall, rather than class-conditionally, independent [6].

5.2 Inter-pattern Class Dependence

The best known examples of context are those among letters (in English 'u' is more likely given the previous letter is 'q' than if the previous letter is 'i'), among words ('sunny day' is more likely than 'sunny night'), and among phonemes in speech. However, one is also likely to find high positive correlation among the classes (species) of adjacent trees, crops, or malignant cells. These correlations are universally exploited in OCR and ASR, but only seldom in remote sensing and in biomedical image analysis.

5.3 Inter-pattern Feature Dependence

Markov type dependence occurs in ligatures in handwriting and as co-articulation in speech. Both are due to obvious physical constraints. In remote sensing, adjacent pixels are often correlated simply because the instantaneous field of view of the sensor (its point spread function) is larger than the spatial sampling interval. In forestry, one might expect that the height-dependent shadow cast by each tree, regardless of its species, affects the height of its neighbors.

Style dependence may occur because of external but unknown factors. In printed matter, it is reasonable to assume that each document has a dominant typeface. A single hand-written form, or a single telephone conversation, exhibits some homogeneity due to the writer or speaker. Trees within the same stand share the same weather and soil conditions. Cells from some organ of a given individual are more similar than cells from different organs or different individuals.

Figure 11 shows an optical illusion where the human brain is incapable of ignoring context in labeling the tiles as either 'dark' or 'light'. Pixel-features of a tile are dependent on the pixel-features of neighboring tiles due to shared illumination. In this example the dependence can be factored as: given the color-class ('dark' or 'light'), and stylistic factors (illumination, i.e., presence or absence of shadow) the features of neighboring tiles are independent.

5.4 Complete Dependence

Although the situation where everything depends on everything else may be quite prevalent it is often practical to make restrictive assumptions on the dependence structure, thereby increasing the small-sample robustness at the expense of increased modeling bias leading to higher accuracy.

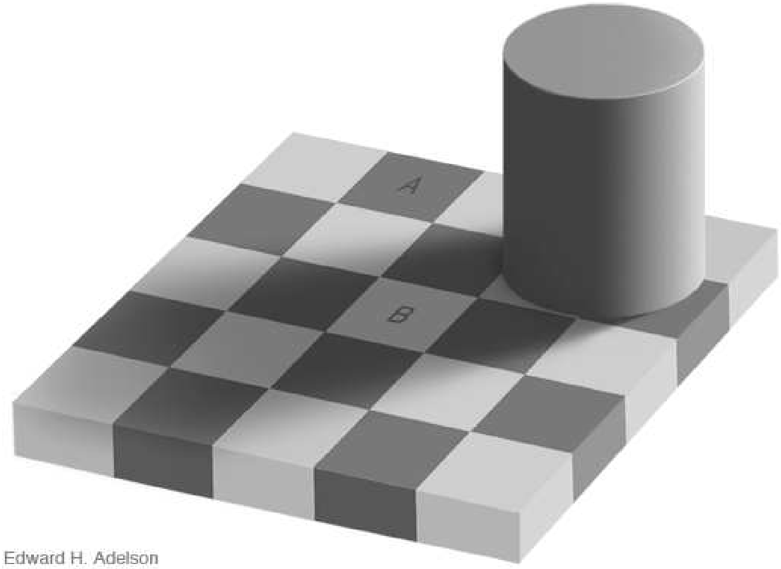

Fig. 11. The 'Checkerboard shadow illusion' created by E. H. Adelson. (http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html. Used with permission.) The tiles labeled 'A' (1, 2) and 'B' (3, 3) have exactly the same grey level.

6 Discussion and Conclusion

We have examined how different varieties of contextual information as statistical dependences can be modeled by means of directed probabilistic graphical models. The most commonly modeled types of dependence are those between features of the same pattern and linguistic label context. Dependence between nearby patterns can be modeled by Markov chains, which can be extended to Markov fields in image processing. Dependence between patterns within a frame, regardless of their position, can be modeled using styles. We have provided references to some common techniques used for learning the parameters for the various models.

We have seen that even for a simplified case of frames with only two patterns there are a number of varieties of contextual information that can be modeled and exploited. For frames with more patterns this number is larger. Although usually several types of context occur simultaneously, in most problems we expect that a few types dominate the others. The choice of a good dependence (or context) model involves art and careful empirical case analysis. We believe that rapid advances in computer speed and storage provide an opportunity to model and exploit increasingly complex manifestations of context for more accurate pattern classification.

References

1. M. Aji, S and R. J. McEliece. The generalized distributive law. *IEEE. Trans. on Information Theory*, 46(2):325–343, 2000.
2. H. D. Brunk. *An Introduction to Mathematical Statistics.* Ginn&Co., Boston, 1960.
3. C. K. Chow. A recognition method using neighbor dependence. *IRE Trans. Elec. Comp.*, EC-11:683–690, 1966.
4. C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Info. Theory*, IT-14:462–467, 1968.
5. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society*, Series B(39):1–38, 1977.
6. R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* Wiley, 1973.
7. B. J. Frey. *Graphical Models for Machine Learning and Digital Communication.* MIT Press, 1998.
8. H. Friedman. Regularized discriminant analysis. *Journal of American Statistical Association*, 84(405):166–175, 1989.
9. D. I. Havelock. The topology of locales and its effect on position uncertainty. *IEEE Trans. PAMI*, 13(4):380–386, 1991.
10. D. Heckerman. A tutorial on learning with bayesian networks. Technical report, Microsoft Research, Redmond, Washington, 1995.
11. M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.
12. J. L. McCarthy. Notes on formalizing context. In *IJCAI*, pages 555–562, 1993.
13. K. Murphy. An introduction to graphical models. Technical report, Intel Research, 2001.
14. G. Nagy. Teaching a computer to read. In *Proceedings of the Eleventh International Conference on Pattern Recognition*, volume 2, pages 225–229, 1992.
15. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.
16. J. Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2000.
17. L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proc. of the IEEE*, volume 2, pages 257–285, 1989.
18. S. J. Raudys and A. K. Jain. Small sample effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. PAMI*, 13(3):252–263, 1991.
19. R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood, and the EM algorithm. *SIAM Review*, 26(2):195–235, 1984.
20. P. Sarkar and G. Nagy. Style consistent classification of isogenous patterns. *IEEE Trans. PAMI*, 27(1):14–22, 2005.
21. P. Sarkar, G. Nagy, J. Zhou, and D. Lopresti. Spatial sampling of printed patterns. *IEEE Trans. PAMI*, 20(3):344–351, 1998.
22. R. Schalkoff. *Pattern Recognition.* Wiley, 1991.
23. J. Schurmann. *Pattern Classification.* Wiley, 1996.
24. L. Serafini and P. Bouquet. Comparing formal theories of context in AI. *Artif. Intell.*, 155(1-2):41–67, 2004.
25. G. T. Toussaint. The use of context in pattern recognition. *Pattern Recognition*, 10(3):189–204, 1978.
26. S. Veeramachaneni and G. Nagy. Style context with second-order statistics. *IEEE Trans. PAMI*, 27(1):88–98, 2005.