Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®

Discrete Applied Mathematics 139 (2004) 113–135

DISCRETE
APPLIED
MATHEMATICS

www.elsevier.com/locate/dam

Spatial pattern discovery by learning a probabilistic parametric model from multiple attributed relational graphs

Pengyu Hong[a], Thomas S. Huang[b]

[a]Science Center 601, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA
[b]Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Received 31 October 2001; received in revised form 12 September 2002; accepted 23 November 2002

Abstract

This paper presents the methodology and theory for automatic spatial pattern discovery from multiple attributed relational graph samples. The spatial pattern is modelled as a mixture of probabilistic parametric attributed relational graphs. A statistic learning procedure is designed to learn the parameters of the spatial pattern model from the attributed relational graph samples. The learning procedure is formulated as a combinatorial non-deterministic process, which uses the expectation–maximization (EM) algorithm to find the maximum-likelihood estimates for the parameters of the spatial pattern model. The learned model summarizes the samples and captures the statistic characteristics of the appearance and structure of the spatial pattern, which is observed under various conditions. It can be used to detect the spatial pattern in new samples. The proposed approach is applied to unsupervised visual pattern extraction from multiple images in the experiments.

Keywords: Spatial pattern discovery; Attributed relational graph; Parametric attributed relational graph; EM algorithm

1. Introduction

In many application domains (e.g., image/video retrieval, software engineering, understanding the biological activity of chemical compounds, etc.), structured information is dependently distributed among the basic primitives and the relationships between them. Extracting regular structured information from observations is an interesting and

E-mail addresses: hong@stat.harvard.edu (P. Hong), huang@ifp.uiuc.edu (T.S. Huang).

0166-218X/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2002.11.007

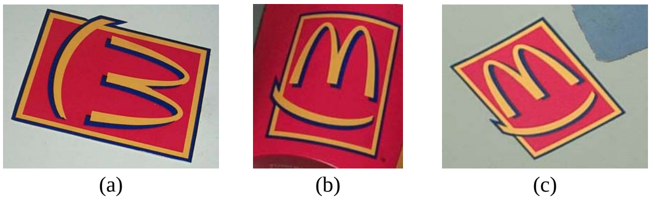


Fig. 1. The McDonald’s logo is observed under different conditions.

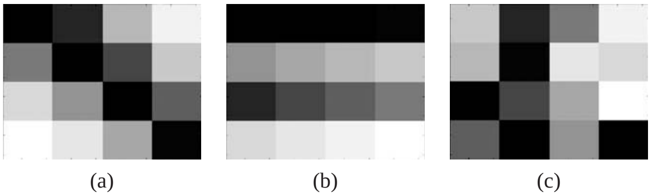


Fig. 2. Three different image patterns are formed by shuffling the same set of image blocks so that the spatial relationships between the image blocks in the patterns are different from each other.

challenging problem. The extracted information can be used to summarize old observations and predict new observations. This paper reports our work on automatic regular structured information extraction from samples. The regular structured information is represented as a spatial pattern. The samples consist of various backgrounds as well as the instances of the spatial pattern (see Fig. 1).

We chose a general graphic representation, attributed relational graphs (ARGs) [17], to represent structured information. An ARG consists of a set of nodes that are connected by a set of arcs. The nodes represent the basic primitives (e.g., image pixels, edges, image segments, atoms, molecules, gene, computer programming segments, etc.). The arcs represent the relationships between the primitives. The attributes of the nodes encode the properties of the primitives. The attributes of the relationships describe the context of the primitives. In the rest of this paper, we call the ARG representations of the samples as sample ARGs.

There are many approaches for learning spatial pattern models from multiple samples of the spatial patterns. Ratan et al. [13] used the diverse density algorithm [12] to learn “visual concepts” (i.e., spatial patterns), from multiple images. A “visual concept” is a pre-specified conjunction of several image primitives. The representation of the “visual concept” in [13] is similar to ARG. Nonetheless, the relationships between the image primitives are not modelled in [13]. Different spatial patterns may share the same set of primitives while containing considerably different relationships (see Fig. 2).

Adopting the data augmentation scheme [5,16], Frey and Jojic [7] treated transformations as latent variables and used the probabilistic graphical model to represent image patterns and their transformations. They used the EM algorithm [5] to learn the model from image samples. The transformations of an image pattern are defined as shuffles of image pixels inside the pattern. They limited the value range of the transformations to a small pre-defined discrete set because the number of the potential transformations is exponential with respect to the number of pixels of an image pattern. It requires non-trivial prior knowledge to define the discrete transformation set. In addition, the image pixels of an image pattern are just like the nodes of an ARG. However, similar to [13], they did not model the relationships between image pixels.

Zhu and Guo [20] studied the conceptualization and modelling of visual patterns from the perspective of statistical physics. They proposed a Gestalt ensemble for modelling spatial organization of attributed points (i.e., the nodes in an ARG). The Gestalt ensemble is associated with a probability model, which can be learned from samples via a minimax entropy learning scheme [21]. The learned model captures visual patterns by examining the local interactions among attributed points in a dynamic local neighborhood.

The contextual information of image pixels was utilized by Hong and Huang to automatically detect recurrent image patterns in a big image [9]. They showed how image patterns could be extracted by the local interactions of image pixels. However, the only allowable transformation of the image patterns was translation. Hong et al. [10] used the generalized EM algorithm to learn the spatial pattern model from multiple sample ARGs. Nevertheless, the theory in [10] is far from fully developed. This paper improves the work of [10], and reports the methodology and theory for unsupervised spatial pattern discovery by learning a spatial pattern model as a probabilistic parametric model from multiple sample ARGs.

We assume that the instances of the spatial pattern are governed by some underlying probabilistic distribution, which is represented by a parametric spatial pattern model. The task is to infer the parameters of the model from multiple sample ARGs. Section 2 introduces the mathematic representations of the sample ARGs and the parametric spatial pattern model. Section 3 mathematically formulates the task and uses the EM algorithm to learn the maximum-likelihood parameters for the parametric model. Section 4 addresses implementation issues and analyzes the computational complexity. Section 5 discusses how to use the learned model for pattern detection. Experimental results are shown in Section 6. Finally, the paper closes with summary and discussions in Section 7.

2. The representations

In reality, the instances of a spatial pattern will not be the same because of the noise of sensors, different observation conditions, and so on. Probabilistic modelling tools have been shown to be effective for handling noise and variation. We design a probabilistic parametric model to represent the spatial pattern.

2.1. Probabilistic modelling of the spatial pattern

Without losing generality, we assume that the instances of the spatial pattern are governed by some probability distribution function (PDF) $f(G|Z)$, where Z is the spatial pattern model of interest and G is a variable representing the instance of Z. Our goal is to infer Z given the sample ARGs.

It is in general very difficult to estimate Z without any prior knowledge about $f(G|Z)$. In practice, the PDF $f(G|Z)$ is usually assumed to have a structure, for example, a linear combination of parametric mixtures. Adopting this method, we assume that $f(G|Z)$ is a linear combination of parametric mixtures and Z consists of a set of parametric model components $\{\Phi_w\}_{w=1}^{W}$, where W is the number of model components. Each mixture of $f(G|Z)$ is represented by a model component Φ_w. Hence, we have

$$f(G|Z) = \sum_{w=1}^{W} \alpha_w \xi(G|\Phi_w), \tag{1}$$

where $\xi(G|\Phi_w)$ is a parametric mixture (or parametric distribution sub-function) of $f(G|Z)$, α_w is the weight of $\xi(G|\Phi_w)$, and $\sum_{w=1}^{W} \alpha_w = 1$. $\xi(G|\Phi_w)$ has simpler structure and is easier to estimate. The value of α_w implies the amount of information which is captured by $\xi(G|\Phi_w)$. For the purpose of data summarization, the value of W should be much smaller than the number of the sample ARGs.

To calculate $f(G|Z)$, we need to know how to evaluate $\{\xi(G|\Phi_w)\}_w$. In the following subsections, we first define the representations for the sample ARGs and Z. Then, we derive the computational forms for $\{\xi(G|\Phi_w)\}_w$.

2.2. The sample ARGs

The sample ARG set is denoted as $\mathbb{G} = \{G_i\}_{i=1}^{S}$, where S is the number of the sample ARGs. The nodes of the sample ARGs are called sample nodes. The relations of the sample ARGs are called sample relations. A sample ARG is represented as $G_i = \langle A_i, R_i \rangle$, which is explained in details as below.

(a) $A_i = \{\langle o_{ik}, \underset{ik}{\vec{a}} \rangle\}_{k=1}^{U_i}$, where o_{ik} is a sample node, $\underset{ik}{\vec{a}}$ is the attribute vector of o_{ik}, and U_i is the number of the sample nodes in G_i.

(b) $R_i = \{\langle r_{icd}, \underset{icd}{\vec{b}} \rangle\}_{c,d=1}^{U_i}$, where r_{icd} represents the relation between o_{ic} and o_{id}, $\underset{icd}{\vec{b}}$ is the attribute vector of r_{icd}. We assume the relationships are directional. If r_{icd} and r_{idc} are directionless, we have $r_{icd} = r_{idc}$ and $\underset{icd}{\vec{b}} = \underset{idc}{\vec{b}}$. If there is no relationship from o_{ic} to o_{id}, both r_{icd} and $\underset{icd}{\vec{b}}$ are void.

2.3. The parametric pattern model

The model components are represented as parametric attributed relational graphs. The nodes of the model components are called model nodes. The relations of the model components are called model relations. Each model component is denoted as $\Phi_w = \langle \Omega_w, \Psi_w \rangle$, where:

(a) $\Omega_w = \{\langle \omega_{wk}, \underset{wk}{\vec{\varphi}}, \beta_{wk} \rangle\}_{k=0}^{N_w}$. ω_{wk} is a model node. Particularly, ω_{w0} is a null model node. N_w is the number of non-null model nodes. To allow occlusions, different model components may have different number of non-null model nodes. Each non-null model node ω_{wk} is associated with a parametric node PDF $p(o_{im}|\omega_{wk})$ whose parameter vector is $\underset{wk}{\overrightarrow{\varphi}}$. The parameter β_{wk} implies the relative frequency of the model node ω_{wk} being observed in the sample ARGs. It is normalized with respect to all the model nodes in Φ_w. We have $\sum_{k=0}^{N_w} \beta_{wk} = 1$. The null model node ω_{w0} does not have physical existence and is used to provide a modelling destination for those sample nodes that represent backgrounds. The node PDF $p(o_{im}|\omega_{w0})$ and the parameter vector $\underset{w0}{\overrightarrow{\varphi}}$ are void.

(b) $\Psi_w = \langle \psi_{w\sigma\tau}, \underset{w\sigma\tau}{\overrightarrow{\vartheta}} \rangle$. $\psi_{w\sigma\tau}$ is a model relation. The model relation $\psi_{w\sigma\tau}$ is a null relation if there is no relation from $\omega_{w\sigma}$ to $\omega_{w\tau}$. Each non-null relation $\psi_{w\sigma\tau}$ is associated with a parametric relation PDF $p(r_{icd}|\omega_{w\sigma\tau})$ whose parameter vector is $\underset{w\sigma\tau}{\overrightarrow{\vartheta}}$. The relation PDF and the parameter vector of a null model relation are void.

Let $\Theta_w = \{\underset{wk}{\overrightarrow{\varphi}}\} \cup \{\beta_{wk}\} \cup \{\underset{w\sigma\tau}{\overrightarrow{\vartheta}}\}$ denote the parameter set of Φ_w. Let $\tilde{\Theta} = \bigcup_w \Theta_w$ denote the parameter set of Z.

2.4. The probability density function of the spatial pattern model

To evaluate $\{\xi(G|\Phi_w)\}_w$ and $f(G|Z)$, the match between G and the model Z is required. Let $\underset{i}{\overrightarrow{y}} = [q_i, y_{i1}, \ldots, y_{iU_i}]$ denote the match between a sample ARG G_i and the model Z. The information in $\underset{i}{\overrightarrow{y}}$ represents two-level match between G_i and Z. The first level information is represented by q_i, which denotes G_i as a whole graph matches the component Φ_{q_i} of Z. The value range of q_i is $[1, W]$. The second level information is represented by $[y_{i1}, \ldots, y_{iU_i}]$, which denotes the match between the sample nodes of G_i and the model nodes of Φ_{q_i}. The element y_{ij} denotes that the sample node o_{ij} matches the model node $\omega_{q_i y_{ij}}$. The value range of y_{ij} is $[0, N_{q_i}]$.

Let $P(y_{ij}|G_i, \Phi_w)$ (i.e., $P(y_{ij}|G_i, \Theta_w)$) denote the matching probability between o_{ij} and $\omega_{wy_{ij}}$. Assuming $P(y_{ij}|G_i, \Phi_w)$ is available (The details about calculating $P(y_{ij}|G_i, \Phi_w)$ will be discussed in Section 4), we have $\xi(G_i|\Phi_w)$ as

$$\begin{aligned}\xi(G_i|\Phi_w) = &\sum_{k=1}^{U_i}\sum_{j=1}^{N_w} P(y_{ik}=j|G_i,\Phi_w)p(o_{ik}|\omega_{wj})\\ &+\sum_{c=1}^{U_i}\sum_{d=1}^{U_i}\sum_{\sigma=1}^{N_w}\sum_{\tau=1}^{N_w} P(y_{ic}=\sigma|G_i,\Phi_w)\\ &\times P(y_{id}=\tau|G_i,\Phi_w)p(r_{icd}|\psi_{w\sigma\tau}).\end{aligned} \tag{2}$$

The probability of G_i matching Φ_w given the model Z is

$$P(q_i = w|G_i, Z) = \frac{\xi(G_i|\Phi_w)}{\sum_{t=1}^{W} \xi(G_i|\Phi_t)}. \tag{3}$$

3. Estimating the parameters of the spatial pattern model via the EM algorithm

The parameter estimation problem becomes straightforward if we known the matching probabilities between the sample ARGs and Z. However, it is tedious and labor intensive to manually specify the matching information for a large set of sample ARGs. We are interested in automatically learning the spatial pattern model without manually specifying the matching information. This section derives the theory for inferring the maximum likelihood parameters for Z using the EM algorithm [5]. The learning procedure simultaneously estimates the parameters of Z and the matching probabilities between the sample ARGs and Z.

3.1. The basic EM algorithm

The EM algorithm is a technique for iteratively finding the maximum-likelihood estimates for the parameters of a underlying distribution from a training data set, which is incomplete or has missing information. The EM algorithm defines a likelihood function

$$Q(H;H^{(n)}) = E[\log p(D_0, D_m|H)|D_0, H^{(n)}], \tag{4}$$

where H is the unknown parameter set, D_0 is the observed data, D_m is the missing information, and n is the number of the iterations of the EM algorithm. The complete data set is $D_0 \cup D_m$. The likelihood function $Q(H;H^{(n)})$ is a function of H under the assumption that $H = H^{(n)}$. The right hand side of (4) denotes that the expected value of the complete data log-likelihood $\log p(D_0, D_m|H)$ with respect to D_m and D_0 while assuming $H = H^{(n)}$.

The EM algorithm starts with an initial value of H, say $H^{(0)}$, and refines the value of H iteratively in two steps: the expectation step (or the E-step) and the maximization step (or the M-step). In the E-step, $Q(H;H^{(n)})$ is computed. In the M-step, the parameter set H is updated by

$$H^{(n+1)} = \arg\max_{H} Q(H;H^{(n)}). \tag{5}$$

The iterative procedure stops when it converges or a pre-defined maximum number of iterations is reached.

3.2. The likelihood function for learning the parameters of the spatial pattern model

In our case, the observed data D_0 is the sample ARG set $\mathbb{G}$. The missing data D_m corresponds to the match between the sample ARGs and Z. Let $\mathbb{Y} = \{\underset{i}{\vec{y}}\}$. The unknown parameter set is $\tilde{\Theta}$. The likelihood function for our problem is

$$\begin{aligned} Q(\tilde{\Theta};\tilde{\Theta}^{(n)}) &= E_f[\log p(\mathbb{G}, \mathbb{Y}|\tilde{\Theta})|\mathbb{G}, \tilde{\Theta}^{(n)}] \\ &= \sum_{\mathbb{Y}} f(\mathbb{G}, \mathbb{Y}|\tilde{\Theta}^{(n)}) \log p(\mathbb{G}, \mathbb{Y}|\tilde{\Theta}) \end{aligned}$$

$$= \sum_{\mathbb{Y}} f(\mathbb{Y}|\mathbb{G}, \tilde{\Theta}^{(n)}) f(\mathbb{G}|\tilde{\Theta}^{(n)}) \log p(\mathbb{G}, \mathbb{Y}|\tilde{\Theta})$$

$$= f(\mathbb{G}|\tilde{\Theta}^{(n)}) \sum_{\mathbb{Y}} f(\mathbb{Y}|\mathbb{G}, \tilde{\Theta}^{(n)}) \log p(\mathbb{G}, \mathbb{Y}|\tilde{\Theta}). \tag{6}$$

We can remove $f(\mathbb{G}|\tilde{\Theta}^{(n)})$ from (6) because it does not depend on either $\tilde{\Theta}$ or $\mathbb{Y}$ and will not affect the final results. We further assume that G_i is independent of each other. Consequently, $\underset{i}{\vec{y}}$ is independent of each other. Hence, (6) can be rewritten as

$$\begin{aligned}
Q(\tilde{\Theta}; \tilde{\Theta}^{(n)}) &= \sum_{\mathbb{Y}} f(\mathbb{Y}|\mathbb{G}, \tilde{\Theta}^{(n)}) \log p(\mathbb{G}, \mathbb{Y}|\tilde{\Theta}) \\
&= \sum_{\underset{1}{\vec{y}}} \cdots \sum_{\underset{S}{\vec{y}}} \sum_{i=1}^{S} \left(\log p(G_i, \underset{i}{\vec{y}}|\tilde{\Theta}) \prod_{j=1}^{S} f(\underset{j}{\vec{y}}|G_j, \tilde{\Theta}^{(n)}) \right) \\
&= \sum_{i=1}^{S} \sum_{\underset{i}{\vec{y}}} f(\underset{i}{\vec{y}}|G_i, \tilde{\Theta}^{(n)}) \log p(G_i, \underset{i}{\vec{y}}|\tilde{\Theta}) \\
&= \sum_{i=1}^{S} \sum_{\underset{i}{\vec{y}}} f(\underset{i}{\vec{y}}|G_i, \tilde{\Theta}^{(n)}) \log(p(G_i|\underset{i}{\vec{y}}, \tilde{\Theta}) p(\underset{i}{\vec{y}}|\tilde{\Theta})) \\
&= \sum_{i=1}^{S} \sum_{\underset{i}{\vec{y}}} f(\underset{i}{\vec{y}}|G_i, \tilde{\Theta}^{(n)}) \log(p(G_i|\underset{i}{\vec{y}}, \tilde{\Theta}) p(\underset{i}{\vec{y}})).
\end{aligned} \tag{7}$$

The term $f(\underset{i}{\vec{y}}|G_i, \tilde{\Theta}^{(n)})$ in (7) is the marginal distribution of $\underset{i}{\vec{y}}$, i.e., the unobserved match between G_i and Z. It is dependent on the observed data $\mathbb{G}$ and the current value of the parameter set $\tilde{\Theta}$. The contextual information of the nodes is fully described in G_i. In other words, the interdependence among $\{y_{ik}\}$ is described by G_i. Hence, we have

$$f(\underset{i}{\vec{y}}|G_i, \tilde{\Theta}^{(n)}) = P(q_i|G_i, \tilde{\Theta}^{(n)}) \prod_{k=1}^{U_i} f(y_{ik}|G_i, \Theta_{q_i}^{(n)}). \tag{8}$$

Since the value space of y_{ik} is uniformly discretized with respect to the number of model nodes in Φ_{qi}, (8) can be rewritten as

$$f(\underset{i}{\vec{y}}|G_i, \tilde{\Theta}^{(n)}) = P(q_i|G_i, \tilde{\Theta}^{(n)}) \prod_{k=1}^{U_i} P(y_{ik}|G_i, \Theta_{q_i}^{(n)}). \tag{9}$$

The term $p(G_i|\underset{i}{\overrightarrow{y}},\tilde{\Theta})$ in (7) is the marginal distribution of G_i given the model Z and the match $\underset{i}{\overrightarrow{y}}$. It can be rewritten as

$$
\begin{aligned}
p(G_i|\underset{i}{\overrightarrow{y}},\tilde{\Theta}) &= p(G_i|[y_{i1}\cdots y_{iU_i}],\Theta_{q_i})\\
&= \prod_{m=1}^{U_i} p(o_{im}|\omega_{q_i y_{im}}) \prod_{c=1}^{U_i}\prod_{d=1}^{U_i} p(r_{icd}|\psi_{q_i y_{ic} y_{id}}),
\end{aligned}
\tag{10}
$$

where $p(o_{im}|\omega_{q_i y_{im}})$ is the node PDF of $\omega_{q_i y_{im}}$ and $p(r_{icd}|\psi_{q_i y_{ic} y_{id}})$ is the relation PDF of $\psi_{q_i y_{ic} y_{id}}$. If the relations are directionless, (10) should be written as

$$
\begin{aligned}
p(G_i|\underset{i}{\overrightarrow{y}},\tilde{\Theta}) &= p(G_i|[y_{i1}\cdots y_{iU_i}],\Theta_{q_i})\\
&= \prod_{m=1}^{U_i} p(o_{im}|\omega_{q_i y_{im}}) \left(\prod_{c=1}^{U_i}\prod_{d=1}^{U_i} p(r_{icd}|\psi_{q_i y_{ic} y_{id}})\right)^{1/2}.
\end{aligned}
\tag{11}
$$

In the following derivation, we use (10). It can be easily shown that only part of the results will be affected by a scale of 1/2 if we use (11).

Expanding the term $P(\underset{i}{\overrightarrow{y}})$ in (7), we have

$$
P(\underset{i}{\overrightarrow{y}}) = P(q_i)\prod_{t=1}^{U_i} P(y_{it}|q_i),
\tag{12}
$$

where $P(q_i=h)=\alpha_h$ and $P(y_{ic}=\eta|q_i=h)=\beta_{h\eta}$.

Substituting (9), (10), and (12) into (7), we have

$$
\begin{aligned}
Q(\tilde{\Theta};\tilde{\Theta}^{(n)}) &= \sum_{i=1}^{S}\sum_{\underset{i}{\overrightarrow{y}}} P(q_i|G_i,\tilde{\Theta}^{(n)})\prod_{k=1}^{U_i} P(y_{ik}|G_i,\Theta_{q_i}^{(n)})\\
&\times \log\left(\prod_{m=1}^{U_i} p(o_{im}|\omega_{q_i y_{im}})\prod_{c=1}^{U_i}\prod_{d=1}^{U_i} p(r_{icd}|\psi_{q_i y_{ic} y_{id}})P(q_i)\right.\\
&\left.\times\prod_{t=1}^{U_i} P(y_{it}|q_i)\right).
\end{aligned}
\tag{13}
$$

Expanding $\sum_{\underset{i}{\overrightarrow{y}}}$ and replacing $\log\prod g(x)$ with $\sum \log g(x)$ in (13), we have

$$
\begin{aligned}
Q(\tilde{\Theta};\tilde{\Theta}^{(n)}) &= \sum_{i=1}^{S}\sum_{q_i=1}^{W}\sum_{y_{i1}=0}^{N_{q_i}}\cdots\sum_{y_{iU_i}=0}^{N_{q_i}} P(q_i|G_i,\tilde{\Theta}^{(n)})\prod_{k=1}^{U_i} P(y_{ik}|G_i,\Theta_{q_i}^{(n)})\\
&\times\left[\log P(q_i) + \sum_{m=1}^{U_i}\log(p(o_{im}|\omega_{q_i y_{im}})P(y_{im}|q_i))\right.\\
&\left. +.\sum_{c=1}^{U_i}\sum_{d=1}^{U_i}\log p(r_{icd}|\psi_{q_i y_{ic} y_{id}})\right].
\end{aligned}
\tag{14}
$$

Eq. (14) can be simplified into (see Appendix A)

$$\begin{aligned} Q(\tilde{\Theta};\tilde{\Theta}^{(n)}) = \sum_{i=1}^{S}\sum_{h=1}^{W} P_{q_i}(h|G_i,\tilde{\Theta}^{(n)}) \Bigg[& \log\alpha_h \\ & + \sum_{m=1}^{U_i}\sum_{\eta=0}^{N_h} P_{y_{im}}(\eta|G_i,\Theta_h^{(n)})\log\beta_{h\eta} \\ & + \sum_{m=1}^{U_i}\sum_{\eta=0}^{N_h} P_{y_{im}}(\eta|G_i,\Theta_h^{(n)})\log p(o_{im}|\omega_{h\eta}) \\ & + \sum_{c=1}^{U_i}\sum_{d=1}^{U_i}\sum_{\sigma=0}^{N_h}\sum_{\tau=0}^{N_h} P_{y_{ic}}(\sigma|G_i,\Theta_h^{(n)}) \\ & \times P_{y_{id}}(\tau|G_i,\Theta_h^{(n)})\log p(r_{icd}|\psi_{h\sigma\tau}) \Bigg], \end{aligned} \tag{15}$$

where $P_{q_i}(h|G_i,\tilde{\Theta}^{(n)})$ denotes $P(q_i = h|G_i,\tilde{\Theta}^{(n)})$ and $P_{y_{im}}(\eta|G_i,\Theta_h^{(n)})$ denotes $P(y_{im} = \eta|G_i,\Theta_h^{(n)})$. The probability $P_{q_i}(h|G_i,\tilde{\Theta}^{(n)})$ can be calculate using (3). The calculation of $P_{y_{im}}(\eta|G_i,\Theta_h^{(n)})$ will be discussed in Section 4.1.

3.3. The expressions for updating the parameters in the M-step

In the Maximization step, $\tilde{\Theta}$ is updated by $\tilde{\Theta}^{(n+1)} = \arg\max_{\Theta} Q(\tilde{\Theta};\tilde{\Theta}^{(n)})$. The expressions for updating α_h and $\beta_{h\eta}$ can be obtained as below regardless the forms of the node PDFs and those of the relation PDFs (see Appendix B)

$$\alpha_h^{(n+1)} = \frac{\sum_{i=1}^{S} P_{q_i}(h|G_i,\tilde{\Theta}^{(n)})}{S}, \tag{16}$$

$$\beta_{h\eta}^{(n+1)} = \frac{\sum_{i=1}^{S}\sum_{m=1}^{U_i} P_{y_{im}}(\eta|G_i,\Theta_h^{(n)}) P_{q_i}(h|G_i,\tilde{\Theta}^{(n)})}{\sum_{i=1}^{S} P_{q_i}(h|G_i,\tilde{\Theta}^{(n)})U_i}. \tag{17}$$

Both the parameters of the node PDFs and those of the relation PDFs are decided by the forms of the PDFs, and so are their updating expressions.

If the node PDFs and relation PDFs are Gaussian PDFs, analytical expressions can be derived for updating the parameters of the PDFs in the M-step of the EM algorithm.

Assume the node PDF is Gaussian

$$p(o_{im}|\omega_{h\eta}) = \frac{\exp(-\frac{1}{2}(\underset{im}{\vec{a}} - \underset{h\eta}{\vec{\mu}})^{\mathrm{T}}\Sigma_{h\eta}^{-1}(\underset{im}{\vec{a}} - \underset{h\eta}{\vec{\mu}}))}{(2\pi)^{\varsigma/2}|\Sigma_{h\eta}|^{1/2}}, \tag{18}$$

where $\underset{h\eta}{\vec{\mu}}$ and $\Sigma_{h\eta}$ are the mean and covariance matrix of the node PDF of the model node $\omega_{h\eta}$ respectively, and ς is the dimension of $\underset{h\eta}{\vec{\mu}}$. We can obtain the expressions for updating $\underset{h\eta}{\vec{\mu}}$ and $\Sigma_{h\eta}$ as below (see Appendix C)

$$\underset{h\eta}{\overset{(n+1)}{\vec{\mu}}} = \frac{\sum_{i=1}^{S} \sum_{m=1}^{U_i} \underset{im}{\vec{a}} P_{y_{im}}(\eta|G_i, \Theta_h^{(n)}) P_{q_i}(h|G_i, \tilde{\Theta}^{(n)})}{\sum\limits_{i=1}^{S} \sum\limits_{m=1}^{U_i} P_{y_{im}}(\eta|G_i, \Theta_h^{(n)}) P_{q_i}(h|G_i, \tilde{\Theta}^{(n)})}, \tag{19}$$

$$\Sigma_{h\eta}^{\rightarrow(n+1)} = \frac{\sum_{i=1}^{S} \sum_{m=1}^{U_i} \underset{im}{\overset{(n)}{\vec{x}}} \underset{im}{\overset{(n)}{\vec{x}}}^{\mathrm{T}} P_{y_{im}}(\eta|G_i, \Theta_h^{(n)}) P_{q_i}(h|G_i, \tilde{\Theta}^{(n)})}{\sum_{i=1}^{S} \sum_{m=1}^{U_i} P_{y_{im}}(\eta|G_i, \Theta_h^{(n)}) P_{q_i}(h|G_i, \tilde{\Theta}^{(n)})}, \tag{20}$$

where $\underset{im}{\overset{(n)}{\vec{x}}} = \underset{im}{\vec{a}} - \underset{h\eta}{\overset{(n+1)}{\vec{\mu}}}$.

Assume the relation PDF is Gaussian

$$p(r_{icd}|\psi_{h\sigma\tau}) = \frac{\exp\left(-\frac{1}{2}\left(\underset{icd}{\vec{b}} - \underset{h\sigma\tau}{\vec{\gamma}}\right)^{\mathrm{T}} \Lambda_{h\sigma\tau}^{-1} \left(\underset{icd}{\vec{b}} - \underset{h\sigma\tau}{\vec{\gamma}}\right)\right)}{(2\pi)^{\kappa/2}|\Lambda_{h\sigma\tau}|^{1/2}}, \tag{21}$$

where $\underset{h\sigma\tau}{\vec{\gamma}}$ and $\Lambda_{h\sigma\tau}$ are the mean and covariance matrix of the relation PDF of $\psi_{h\sigma\tau}$, and κ is the dimension of $\underset{h\sigma\tau}{\vec{\gamma}}$. We can obtain the expressions for updating $\underset{h\sigma\tau}{\vec{\gamma}}$ and $\Lambda_{h\sigma\tau}$ as below (see Appendix C)

$$\underset{h\sigma\tau}{\overset{(n+1)}{\vec{\gamma}}} = \frac{\sum_{i=1}^{S} \sum_{c=1}^{U_i} \sum_{d=1}^{U_i} \underset{icd}{\vec{b}} \ell_h(y_{ic}, y_{id}, \sigma, \tau) P_{q_i}(h|G_i, \tilde{\Theta}^{(n)})}{\sum_{i=1}^{S} \sum_{c=1}^{U_i} \sum_{d=1}^{U_i} \ell_h(y_{ic}, y_{id}, \sigma, \tau) P_{q_i}(h|G_i, \tilde{\Theta}^{(n)})}, \tag{22}$$

$$\underset{h\sigma\tau}{\overset{(n+1)}{\vec{\Lambda}}} = \frac{\sum_{i=1}^{S} \sum_{c=1}^{U_i} \sum_{d=1}^{U_i} \underset{icd}{\overset{(n)}{\vec{z}}} \underset{icd}{\overset{(n)}{\vec{z}}}^{\mathrm{T}} \ell_h(y_{ic}, y_{id}, \sigma, \tau) P_{q_i}(h|G_i, \tilde{\Theta}^{(n)})}{\sum_{i=1}^{S} \sum_{c=1}^{U_i} \sum_{d=1}^{U_i} \ell_h(y_{ic}, y_{id}, \sigma, \tau) P_{q_i}(h|G_i, \tilde{\Theta}^{(n)})}, \tag{23}$$

where $\ell_h(y_{ic}, y_{id}, \sigma, \tau) = P_{y_{ic}}(\sigma|G_i, \Theta_h^{(n)}) P_{y_{id}}(\tau|G_i, \Theta_h^{(n)})$ and $\underset{icd}{\overset{(n)}{\vec{z}}} = \underset{icd}{\vec{b}} - \underset{h\sigma\tau}{\overset{(n+1)}{\vec{\gamma}}}$.

4. Implementation issues

4.1. Register the sample ARGs with the spatial pattern model

Given the current value of the parameter set of Φ_w, we can calculate the matching probabilities between G_i and Φ_w using inexact two-graph matching techniques.

Inexact two-graph matching is a fundamental combinatorial problem and is an NP problem [8]. It has been widely investigated for finding a local optimum inexact match between two graphs [1,3,4,11,14,17,15,18,19]. We use an implementation of the probabilistic relaxation graph matching algorithm [4] to match each sample ARG with every component of Z. The matching results are local maximum approximations to $\{P_{y_{im}}(\eta|G_i,\Theta_h^{(n)})\}$.

4.2. Initialize the spatial pattern model

Initializing the spatial pattern model is the first step of the learning procedure and is very important. The number of the model components is decided by the user or the applications. We initialize the model components one by one. First, the average number of the nodes of the sample ARGs is calculated. We select a sample ARG, say G_p, so that the number of sample nodes in G_p is the closest to the average node number. The geometric structure of G_p is used to initialize that of the first model component Φ_1. If the node PDFs and relation PDFs are assumed to be Gaussian, the feature vectors of the nodes and relations of G_p are used to initialize the corresponding means of the node PDFs and relation PDFs of Φ_1. The covariance matrixes of the node PDFs and relation PDFs are initialized as identical matrixes.

The rest of the model components are initialized using the following algorithm. The idea is to initialize the model components by some sample ARGs which are as different from each other as possible.

Algorithm 1 (Initialize the spatial pattern model).

```
(a) for $w = 2$ to $W$
(b) Select a sample $G_p = \arg\min_{G_i}(\max_{\Phi_h}\{\xi(G_i|\Phi_h)\})$.
(c) Initialize the model component $\Phi_w$ using $G_p$.
(d) $\alpha_w = 1$
(e) $\beta_{wk} = 1$ $(0 \leqslant k \leqslant N_w)$
(d) endfor
```

Before beginning the iterative procedure of the EM algorithm, the K-means algorithm is used to pre-adjust the parameters of the spatial pattern model.

4.3. Modify the structure of the spatial pattern model

Since we select a subset of the sample ARGs to initialize the components of the model, it is very likely that the model components have spurious nodes which represent backgrounds. During the iterations of the EM algorithm, we calculate the average probability of being matched for each model node ω_{wk} as

$$\varrho_{wk} = \frac{\sum_{i=1}^{S} P_{q_i}(w|G_i,\tilde{\Theta}^{(n)})\left(\sum_{m=1}^{U_i} P_{y_{im}}(k|G_i,\Theta_w^{(n)})\right)}{\sum_{k=1}^{S} P_{q_k}(w|G_k,\tilde{\Theta}^{(n)})}. \tag{24}$$

If ϱ_{wk} is smaller than a threshold ε, the model node ω_{wk} and its relations will be removed. The threshold ε can be a constant or an ascendant function of the iteration number of the EM algorithm (e.g., we choose $\varepsilon = 1 - 0.5^n$).

4.4. The computational complexity

The computational complexity of the learning procedure is O (the number of the EM iterations $\times$ $\sum_{i=1}^{S}\sum_{w=1}^{W}$ (the computational complexity of matching G_i to Φ_w)). Since it might take too long for the EM algorithm to converge, a maximum number of iterations T is empirically set for the EM algorithm (e.g., we set T to 50). The graph matching algorithm is used by the EM algorithm to deal with the hidden variables, i.e., the match between the sample ARGs and the spatial pattern model. Without additional constraints or prior knowledge, the complexity of the value space of the match between the sample nodes of G_i and the model nodes of Φ_w is $\mathrm{O}((N_w+1)^{U_i})$. To deal with such a huge searching space, we chose a bottom-up graph matching approach (see Section 4.1), which finds a local optimum solution by fusing the low-level information. The computational complexity of our implementation of the graph matching algorithm is $\mathrm{O}(N_w^2 U_i^2)$. The overall computational complexity of the implemented learning procedure is $\mathrm{O}\big(T\sum_i \sum_w (N_w^2 U_i^2)\big)$.

5. Detect the spatial pattern

The learned model captures the statistical characteristics of a spatial pattern observed under various conditions. It can be used to detect whether the pattern appears in a new sample ARG, say $G_x = \langle O_x, R_x\rangle$. The similarity between G_x and the model Z is calculated as $f(G_x|Z)$. An instance of the pattern is said to be found in G_x if $f(G_x|Z)$ is larger than a predefined threshold ϵ_1, which depends on applications. A choice of ϵ_1 could be $\min_{G_i\in\mathbb{G}} f(G_i|Z)$ if each sample ARG G_i has at least one instance of the spatial pattern.

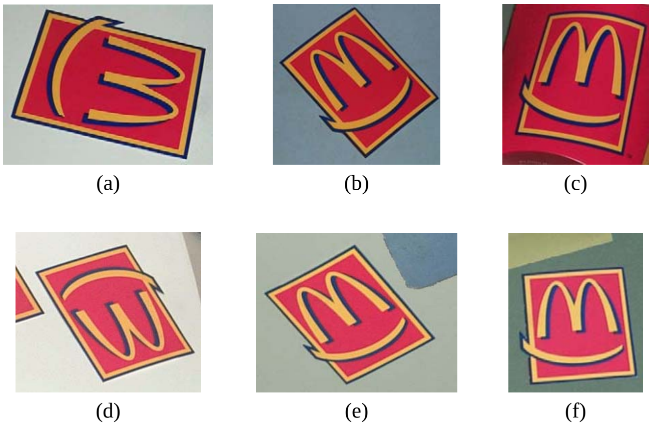
The likelihood of each sample node o_{xk} is calculated as

$$\sum_{h=1}^{W} \alpha_h P(G_x = \Phi_h | G_x, Z) \sum_{\eta=1}^{N_h} \beta_{h\eta} P(o_{xk} = \omega_{h\eta} | G_x, \Phi_h). \tag{25}$$

Those sample nodes whose likelihood is larger than a predefined threshold ϵ_2 are selected. A choice of ϵ_2 could be $0.95S/\left(W\sum_{i=1}^{S} U_i\right)$. The relations among the selected sample nodes are preserved. The selected nodes and relations form an instance of the pattern in G_x.

6. Experimental results

We applied the proposed approach to the problem of unsupervised visual pattern extraction. The image samples are segmented using a segmentation algorithm [6] and



(a) (b) (c)

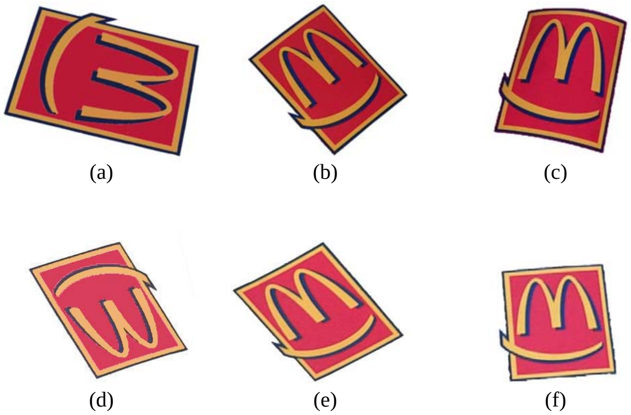
(d) (e) (f)

Fig. 3. The *McDonald's* logo. The first row lists three images captured under the first lighting condition. The second row shows three images captured under the second lighting condition.

are represented as ARGs. Each image segment is represented as a node. The attribute of a node denotes the mean and variance of the color (RGB) features of the corresponding image segment. The adjacent relationships between the image segments are considered. The attributes of the relationships in the sample ARGs are either 1 (adjacent) or 0 (non-adjacent). During the learning process, the attributes of relationships are updated as continuous variables in the range of [0, 1]. When the learning procedure stops, a threshold of 0.5 is used to decide whether a relationship should be kept. A model node without any neighbor will be deleted.

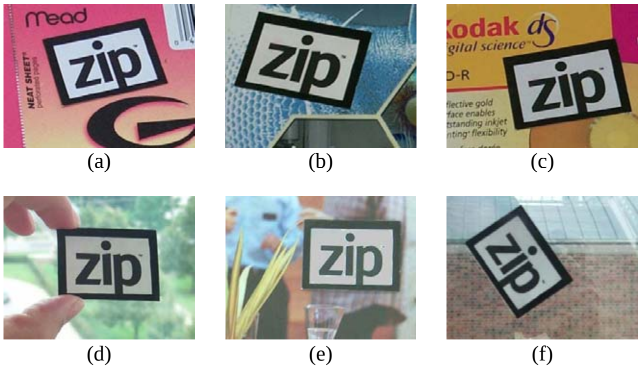
We first show a simple example. The pictures of the *McDonald's* logo were taken in various backgrounds, from different viewpoints, and under two different lighting conditions. Ten images were captured under each lighting condition. Some of them are shown in Fig. 3. The observed color features of the *McDonald's* logo are different in the samples due to different lighting conditions, different viewpoints, and noise. Take '**m**' in the middle of the logo as an example. The images shown in Fig. 3(b) and (e) are captured under different lighting conditions. The means of the color features of '**m**' are (202.4, 138.2, 59.8) and (240.3, 180.1, 109.4) in Fig. 3(b) and (e) respectively. The images shown in Fig. 3(a)–(c) are captured under the same lighting condition. The means of the color features of '**m**' are (208.2, 149.7, 69.1), (202.4, 138.2, 59.8), and (205.7, 144.3, 71.2) in Fig. 3(a), (b), and (c), respectively.

We made two assumptions. First, the spatial model has two model components. Second, the node PDFs and the relation PDFs are Gaussian with fixed covariance matrixes as identical matrixes. Both components of the learned model have 8 nodes. The means of the color attributes of the model nodes, which correspond to '**m**', are (207.5, 140.3, 68.6) and (240.2, 179.7, 117.1), respectively. The learning results already include the detection results of the *McDonald's* logo in the training images (see Fig. 4).


(a) (b) (c)

(d) (e) (f)

Fig. 4. Detect the *McDonald's* logo in the training images.



(a) (b) (c)

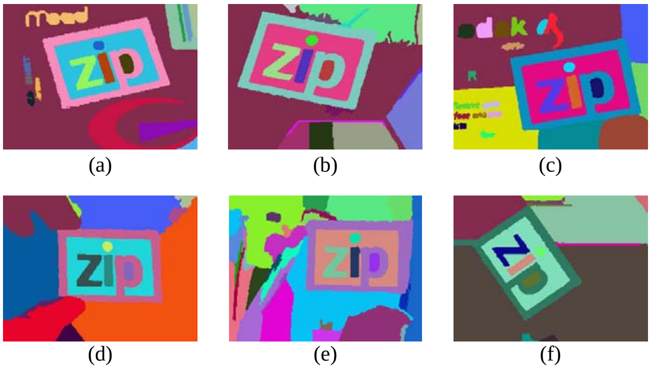
(d) (e) (f)

Fig. 5. The ZIP logo images.

In another experiment, we used the images of the ZIP logo in various backgrounds. The sample set has 20 images. Some of them are shown in Fig. 5. The backgrounds in this experiment are more complicated than those in the previous one. More intermediate results of the computation are provided.

The images are segmented (see Fig. 6) and are represented as ARGs (see Fig. 7). The spatial pattern model is assumed to have one component. The node PDFs and the relation PDFs are assumed to be Gaussian with fixed covariance matrixes as identical matrixes. Fig. 8 shows the detection results on the sample ARGs, which are shown in Fig. 7. Fig. 9 shows the original image regions that correspond to the detected subgraphs in Fig. 8. We also used the learned model to detect the ZIP logo in a new image (see Fig. 10).

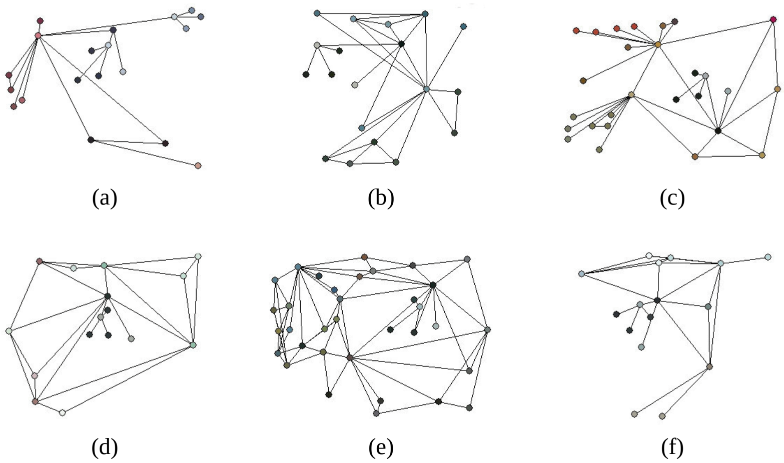
As shown in Fig. 9(d)–(f), the final results depend on the quality of the image segmentation results. In fact, if each image pixel is represented as a node in an



(a) (b) (c)

(d) (e) (f)

Fig. 6. The segmentation results of the images are shown in Fig. 5. The image segments are automatically painted in pseudo colors by the segmentation program [6].



(a) (b) (c)

(d) (e) (f)

Fig. 7. The ARG representations of the images shown in Fig. 5. The nodes represent the image segments. The 2D coordinates of a node in the image plane are decided by the coordinates of a randomly selected image pixel in the corresponding image segment. The coordinates of the nodes are used for visualization only. An edge is drawn to connect two nodes if the corresponding image segments are adjacent.

ARG, our theory can be directly applied to image pixels so that we can avoid using corrupted information generated by the low-level image preprocessing step (e.g., image segmentation, edge detection, etc.). Nonetheless, this will result in high computational complexity if the sample images have large numbers of image pixels. Image segmentation was just used to reduce the computational complexity in our experiments.

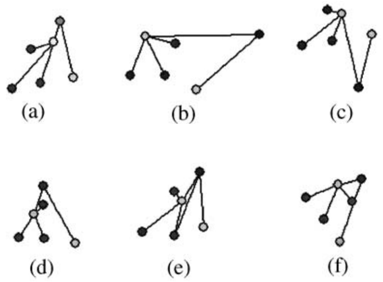


Fig. 8. The detected subgraphs that correspond to the instances of the learned spatial pattern model.

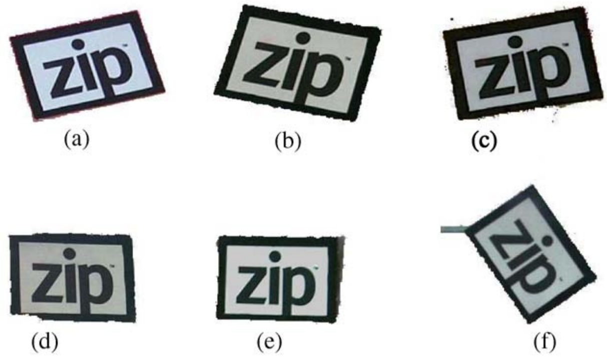


Fig. 9. The original image segments that correspond to the subgraphs in Fig. 8.

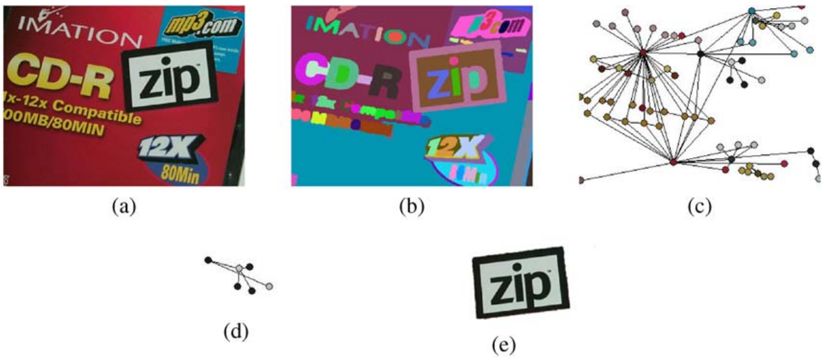


Fig. 10. Detect the ZIP logo in a new image. (a) The image, (b) the segmentation results, (c) the ARG representation, (d) the detected subgraph, and (e) the image segments corresponding to the detected subgraph.

7. Summary and discussions

We present a statistic learning approach that discovers frequently observed structured information by simultaneously examining multiple samples. We assume that the structured information is governed by a PDF which is represented as a probabilistic parametric graph model. The model consists of a set of parametric attributed relational graphs. The learning procedure iteratively finds a local optimum estimate for the para meter set of the model. The learned model summarizes the samples and can be used for pattern detection. We demonstrated the approach by applying it to unsupervised 2D visual spatial pattern extraction. The experimental results show that the learning procedure is able to distinguish the instances of the spatial pattern from their backgrounds if similar backgrounds are not always observed in the samples.

Although the proposed approach was only applied to two dimensional images in the experiments, it is suitable for general spatial pattern learning and discovery. This is because ARG can be used to represent data in any dimensional space. In addition, our approach can be used for feature selection. Representing the instantiations of feature elements as the nodes of sample ARGs, our approach is not only able to discover the dominant feature space but also capture the relationships between the selected features. This is important when the features are not independent.

Future work will expand the proposed methodology and theory for temporal-spatial pattern modelling and incremental learning. We will investigate the applications of our approach to real applications (e.g., gene function modelling and detection, network flow modelling, multimodal human-computer interaction, content-based image retrieval, depth information recovery from multiple images, face detection and recognition, etc.).

Appendix A. Simplify the maximum-likelihood function

We rewrite (14) as

$$Q(\tilde{\Theta};\tilde{\Theta}^{(n)})=\sum_{i=1}^{S}\sum_{q_i=1}^{M}P(q_i|G_i,\tilde{\Theta}^{(n)})(L_1+L_2+L_3), \tag{A.1}$$

where

$$L_1=\sum_{y_{i1}=0}^{N_{q_i}}\cdots\sum_{y_{iU_i}=0}^{N_{q_i}}\prod_{k=1}^{U_i}P(y_{ik}|G_i,\Theta_{q_i}^{(n)})\log P(q_i), \tag{A.2}$$

$$L_2=\sum_{y_{i1}=0}^{N_{q_i}}\cdots\sum_{y_{iU_i}=0}^{N_{q_i}}\prod_{k=1}^{U_i}P(y_{ik}|G_i,\Theta_{q_i}^{(n)})\sum_{m=1}^{U_i}\log\left(p(o_{im}|\omega_{q_i y_{im}})P(y_{im}|q_i)\right), \tag{A.3}$$

$$L_3 = \sum_{y_{i1}=0}^{N_{q_i}} \cdots \sum_{y_{iU_i}=0}^{N_{q_i}} \prod_{k=1}^{U_i} P(y_{ik}|G_i, \Theta_{q_i}^{(n)}) \sum_{c=1}^{U_i} \sum_{d=1}^{U_i} \log p(r_{icd}|\psi_{q_i y_{ic} y_{id}}). \tag{A.4}$$

We then simplify the above three terms one-by-one. From time to time, we will use the fact that $\sum_{y_{ik}=0}^{N_{q_i}} P(y_{ik}|G_i, \Theta_{q_i}^{(n)}) = 1$.

$$\begin{aligned} L_1 &= \log P(q_i) \sum_{y_{i1}=0}^{N_{q_i}} \cdots \sum_{y_{iU_i}=0}^{N_{q_i}} \prod_{k=1}^{U_i} P(y_{ik}|G_i, \Theta_{q_i}^{(n)}) \\ &= \log P(q_i) \prod_{k=1}^{U_i} \sum_{y_{ik}=0}^{N_{q_i}} P(y_{ik}|G_i, \Theta_{q_i}^{(n)}) \\ &= \log P(q_i), \end{aligned} \tag{A.5}$$

$$\begin{aligned} L_2 &= \sum_{y_{i1}=0}^{N_{q_i}} \cdots \sum_{y_{iU_i}=0}^{N_{q_i}} \prod_{k=1}^{U_i} P(y_{ik}|G_i, \Theta_{q_i}^{(n)}) \sum_{m=1}^{U_i} \log(p(o_{im}|\omega_{q_i y_{im}}) P(y_{im}|q_i)) \\ &= \sum_{m=1}^{U_i} \sum_{y_{im}=0}^{N_{q_i}} \log(p(o_{im}|\omega_{q_i y_{im}}) P(y_{im}|q_i)) P(y_{im}|G_i, \Theta_{q_i}^{(n)}) \\ &\quad \times \left[\sum_{y_{i1}=0}^{N_{q_i}} \cdots \sum_{y_{im-1}=0}^{N_{q_i}} \sum_{y_{im+1}=0}^{N_{q_i}} \cdots \sum_{y_{iU_i}=0}^{N_{q_i}} \prod_{k=1, k \neq m}^{U_i} P(y_{ik}|G_i, \Theta_{q_i}^{(n)}) \right] \\ &= \sum_{m=1}^{U_i} \sum_{y_{im}=0}^{N_{q_i}} \log(p(o_{im}|\omega_{q_i y_{im}}) P(y_{im}|q_i)) P(y_{im}|G_i, \Theta_{q_i}^{(n)}) \\ &\quad \times \prod_{k=1, k \neq m}^{U_i} \sum_{y_{ik}=0}^{N_{q_i}} P(y_{ik}|G_i, \Theta_{q_i}^{(n)}) \\ &= \sum_{m=1}^{U_i} \sum_{y_{im}=0}^{N_{q_i}} \log(p(o_{im}|\omega_{q_i y_{im}}) P(y_{im}|q_i)) P(y_{im}|G_i, \Theta_{q_i}^{(n)}) \\ &= \sum_{m=1}^{U_i} \sum_{y_{im}=0}^{N_{q_i}} P(y_{im}|G_i, \Theta_{q_i}^{(n)}) [\log p(o_{im}|\omega_{q_i y_{im}}) + \log P(y_{im}|q_i)], \end{aligned} \tag{A.6}$$

$$
\begin{aligned}
L_3 &= \sum_{y_{i1}=0}^{N_{q_i}} \cdots \sum_{y_{iU_i}=0}^{N_{q_i}} \prod_{k=1}^{U_i} P(y_{ik}|G_i,\Theta_{q_i}^{(n)}) \sum_{c=1}^{U_i}\sum_{d=1}^{U_i} \log p(r_{icd}|\psi_{q_i y_{ic} y_{id}}) \\
&= \sum_{c=1}^{U_i}\sum_{d=1}^{U_i}\sum_{y_{ic}=0}^{N_{q_i}}\sum_{y_{id}=0}^{N_{q_i}} P(y_{ic}|G_i,\Theta_{q_i}^{(n)})P(y_{id}|G_i,\Theta_{q_i}^{(n)}) \log(p(r_{icd}|\psi_{q_i y_{ic} y_{id}})) \\
&\quad\times \left[\sum_{y_{i1}=0}^{N_{q_i}} \cdots \sum_{y_{ic-1}=0}^{N_{q_i}} \sum_{y_{ic+1}=0}^{N_{q_i}} \cdots \sum_{y_{id-1}=0}^{N_{q_i}} \sum_{y_{id+1}=0}^{N_{q_i}} \cdots \sum_{y_{iU_i}=0}^{N_{q_i}} \prod_{k=1,k\neq c,d}^{U_i} P(y_{ik}|G_i,\Theta_{q_i}^{(n)}) \right] \\
&= \sum_{c=1}^{U_i}\sum_{d=1}^{U_i}\sum_{y_{ic}=0}^{N_{q_i}}\sum_{y_{id}=0}^{N_{q_i}} P(y_{ic}|G_i,\Theta_{q_i}^{(n)})P(y_{id}|G_i,\Theta_{q_i}^{(n)}) \log(p(r_{icd}|\psi_{q_i y_{ic} y_{id}})) \\
&\quad\times \prod_{k=1,k\neq c,d}^{U_i} \sum_{y_{ik}=0}^{N_{q_i}} P(y_{ik}|G_i,\Theta_{q_i}^{(n)}) \\
&= \sum_{c=1}^{U_i}\sum_{d=1}^{U_i}\sum_{y_{ic}=0}^{N_{q_i}}\sum_{y_{id}=0}^{N_{q_i}} P(y_{ic}|G_i,\Theta_{q_i}^{(n)})P(y_{id}|G_i,\Theta_{q_i}^{(n)}) \log p(r_{icd}|\psi_{q_i y_{ic} y_{id}}). \qquad \text{(A.7)}
\end{aligned}
$$

Finally, we can obtain

$$
\begin{aligned}
Q(\tilde{\Theta};\tilde{\Theta}^{(n)}) &= \sum_{i=1}^{S}\sum_{q_i=1}^{W} P(q_i|G_i,\tilde{\Theta}^{(n)}) \Bigg[\log P(q_i) \\
&\quad + \sum_{m=1}^{U_i}\sum_{y_{im}=0}^{N_{qi}} P(y_{im}|G_i,\Theta_{q_i}^{(n)}) \log P(y_{im}|q_i) \\
&\quad + \sum_{m=1}^{U_i}\sum_{y_{im}=0}^{N_{qi}} P(y_{im}|G_i,\Theta_{q_i}^{(n)}) \log p(o_{im}|\omega_{q_i y_{im}}) \\
&\quad + \sum_{c=1}^{U_i}\sum_{d=1}^{U_i}\sum_{y_{ic}=0}^{N_{q_i}}\sum_{y_{id}=0}^{N_{q_i}} P(y_{ic}|G_i,\Theta_{q_i}^{(n)}) \\
&\quad \times P(y_{id}|G_i,\Theta_{q_i}^{(n)}) \log p(r_{icd}|\psi_{q_i y_{ic} y_{id}}) \Bigg] \\
&= \sum_{i=1}^{S}\sum_{h=1}^{W} P_{q_i}(h|G_i,\tilde{\Theta}^{(n)}) \Bigg[\log \alpha_h
\end{aligned}
$$

$$+\sum_{m=1}^{U_i}\sum_{\eta=0}^{N_h} P_{y_{im}}(\eta|G_i,\Theta_h^{(n)})\log\beta_{h\eta}$$

$$+\sum_{m=1}^{U_i}\sum_{\eta=0}^{N_h} P_{y_{im}}(\eta|G_i,\Theta_h^{(n)})\log p(o_{im}|\omega_{h\eta})$$

$$+\sum_{c=1}^{U_i}\sum_{d=1}^{U_i}\sum_{\sigma=0}^{N_h}\sum_{\tau=0}^{N_h} P_{y_{ic}}(\sigma|G_i,\Theta_h^{(n)})$$

$$\left.\times P_{y_{id}}(\tau|G_i,\Theta_h^{(n)})\log p(r_{icd}|\psi_{h\sigma\tau})\right], \tag{A.8}$$

where $P_{q_i}(h|G_i,\tilde{\Theta}^{(n)}) = P(q_i = h|G_i,\tilde{\Theta}^{(n)})$ and $P_{y_{im}}(\eta|G_i,\Theta_h^{(n)}) = P(y_{im}=\eta|G_i,\Theta_h^{(n)})$, $P(q_i = h) = \alpha_h$, and $P(y_{im}=\eta|q_i = h) = \beta_{h\eta}$.

Appendix B. Derive expressions for updating α_h and $\beta_{h\eta}$

First, we derive the updating expression for α_h. We introduce the Lagrange multiplier λ with the constraint that $\Sigma_h\,\alpha_h = 1$, and solve the following equation

$$\frac{\partial}{\partial\alpha_h}\left[Q(\tilde{\Theta};\tilde{\Theta}^{(n)}) + \lambda\left(\sum_{h=1}^{W}\alpha_h - 1\right)\right]$$

$$=\frac{\partial}{\partial\alpha_h}\left[\sum_{i=1}^{S}\sum_{h=1}^{W}P_{q_i}(h|G_i,\tilde{\Theta}^{(n)})\log\alpha_h + \lambda\left(\sum_{h=1}^{M}\alpha_h - 1\right)\right]$$

$$=\sum_{i=1}^{S}\frac{1}{\alpha_h}P_{q_i}(h|G_i,\tilde{\Theta}^{(n)}) + \lambda = 0$$

$$\Rightarrow \sum_{h=1}^{W}\left[\sum_{i=1}^{S}\frac{1}{\alpha_h}P_{q_i}(h|G_i,\tilde{\Theta}^{(n)}) + \lambda\right] = 0 \Rightarrow \lambda = -S$$

$$\Rightarrow \alpha_h = \frac{\sum_{i=1}^{S}P_{q_i}(h|G_i,\tilde{\Theta}^{(n)})}{S}. \tag{B.1}$$

Second, we derive the updating expression for $\beta_{h\eta}$. We introduce the Lagrange multiplier λ with the constraint that $\Sigma_\eta\beta_{h\eta} = 1$, and solve the following equation:

$$\frac{\partial}{\partial\beta_{h\eta}}\left[Q(\tilde{\Theta};\tilde{\Theta}^{(n)}) + \lambda\left(\sum_{\eta=0}^{N_h}\beta_{h\eta} - 1\right)\right]$$

$$= \frac{\partial}{\partial \beta_{h\eta}} \left[\sum_{i=1}^{S} \sum_{h=1}^{W} \sum_{m=1}^{U_i} \sum_{\eta=0}^{N_h} P_{y_{im}}(\eta | G_i, \Theta_h^{(n)}) P_{q_i}(h | G_i, \tilde{\Theta}^{(n)}) \log \beta_{h\eta} \right.$$

$$\left. + \lambda \left(\sum_{\eta=0}^{N_h} \beta_{h\eta} - 1 \right) \right]$$

$$= \sum_{i=1}^{S} \sum_{m=1}^{U_i} \frac{1}{\beta_{h\eta}} P_{y_{im}}(\eta | G_i, \Theta_h^{(n)}) P_{q_i}(h | G_i, \tilde{\Theta}^{(n)}) + \lambda = 0$$

$$\Rightarrow \sum_{\eta=0}^{N_h} \left[\sum_{i=1}^{S} \sum_{m=1}^{U_i} P_{y_{im}}(\eta | G_i, \Theta_h^{(n)}) P_{q_i}(h | G_i, \tilde{\Theta}^{(n)}) + \lambda \beta_{h\eta} \right] = 0$$

$$\Rightarrow \lambda = - \sum_{i=1}^{S} P_{q_i}(h | G_i, \tilde{\Theta}^{(n)}) U_i$$

$$\Rightarrow \beta_{h\eta} = \frac{\sum_{i=1}^{S} \sum_{m=1}^{U_i} P_{y_{im}}(\eta | G_i, \Theta_h^{(n)}) P_{q_i}(h | G_i, \tilde{\Theta}^{(n)})}{\sum_{i=1}^{S} P_{q_i}(h | G_i, \tilde{\Theta}^{(n)}) U_i}. \tag{B.2}$$

Appendix C. Derive the updating expressions for Gaussian node PDFs and Gaussian relation PDFs

If the node PDFs and relation PDFs are Gaussian, we can obtain analytical expressions for updating the parameters of the PDFs in the M-step. Basically, we take the derivatives of $Q(\tilde{\Theta}, \tilde{\Theta}^{(n)})$ with respect to the parameters of the PDFs, set the derivatives to zero, and solve the equations.

Only the third term of $Q(\tilde{\Theta}, \tilde{\Theta}^{(n)})$ is related to the node PDFs. Substituting the Gaussian node PDF (18) into the third term of $Q(\tilde{\Theta}, \tilde{\Theta}^{(n)})$, we obtain

$$\sum_{i=1}^{S} \sum_{m=1}^{U_i} \sum_{h=1}^{W} \sum_{\eta=0}^{N_h} P_{q_i}(h | G_i, \tilde{\Theta}^{(n)}) P(\eta | G_i, \Theta_h^{(n)}) \log p(o_{im} | \omega_{h\eta})$$

$$= \sum_{i=1}^{S} \sum_{m=1}^{U_i} \sum_{h=1}^{W} \sum_{\eta=0}^{N_h} P_{q_i}(h | G_i, \tilde{\Theta}^{(n)}) P(\eta | G_i, \Theta_h^{(n)})$$

$$\times \left[\frac{-\varsigma}{2} \log 2\pi + \log |\Sigma_{h\eta}| + (\underset{im}{\vec{a}} - \underset{h\eta}{\vec{\mu}}) \Sigma_{h\eta}^{-1} (\underset{im}{\vec{a}} - \underset{h\eta}{\vec{\mu}}) \right]. \tag{C.1}$$

The above expression is quadratic. It is a typical optimization problem to solve a equation that is obtained by taking the derivative of (C.1) with respect to its parameter and setting the derivative to zero [2]. We first take the derivative of (C.1) with respect to $\underset{h\eta}{\vec{\mu}}$, set it equal to zero, and obtain the updating expression of $\underset{h\eta}{\vec{\mu}}$ as (19). Then, we take the derivative of (C.1) with respect to $\Sigma_{h\eta}$, set it equal to zero, and obtain the updating expression of $\Sigma_{h\eta}$ as (20).

Similarly, only the forth term of $Q(\tilde{\Theta}, \tilde{\Theta}^{(n)})$ is related to the relation PDFs. Substituting the Gaussian relation PDF (18) into the forth term of $Q(\tilde{\Theta}, \tilde{\Theta}^{(n)})$, we obtain a quadratic expression with respect to the parameters of the Gaussian relation PDFs. Using the same method described above, we can obtain the updating expressions of the parameters of the Gaussian relation PDFs as (22) and (23), respectively.

References

[1] H.A. Almohamad, S.O. Duffuaa, A linear programming approach for the weighted graph matching problem, IEEE Trans. Pattern Anal. Mach. Intell. 15 (1993) 522–525.

[2] D.P. Bertsekas, Nonlinear Programming, Athena Scientific, Belmonte, MA, USA, 1995.

[3] B. Bhanu, O.D. Faugeras, Shape matching of two-dimensional objects, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1984) 137–156.

[4] W.J. Christmas, J. Kittler, M. Petrou, Structural matching in computer vision using probabilistic relaxation, IEEE Trans. Pattern Anal. Mach. Intell. 17 (8) (1995) 749–764.

[5] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Stat. Soc. Ser. B 39 (1) (1977) 1–38.

[6] P.F. Felzenszwalb, D.O. Huttenlocher, Image segmentation using local variation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1998, pp. 98–104.

[7] B.J. Frey, N. Jojic, Transformed component analysis: Joint estimation of spatial transformations and image components, International Conference on Comput. Vision, 1999.

[8] M.R. Garey, D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W. H. Freeman and Company, New York, 1979.

[9] P. Hong, T.S. Huang, Extracting the recurring patterns from image, The 4th Asian Conference on Computer Vision, Taipei, Taiwan, 2000.

[10] P. Hong, R. Wang, T.S. Huang, Learning patterns from images by combining soft decisions and hard decisions, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA, 2000.

[11] S.Z. Li, Matching: Invariant to translations, rotations and scale changes, Pattern Recognition 25 (1992) 583–594.

[12] O. Maron, T. Lozano-Prez, A framework for multiple-instance learning, Neural Inform. Process. Systems, 1998.

[13] A.L. Ratan, O. Maron, et al., A framework for learning query concepts in image classification, In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Ft. Collins, CO, USA, 1999, pp. 423–429.

[14] A. Rosenfeld, R. Hummel, S. Zucker, Scene labeling by relaxation operations, IEEE Trans. Systems, Man Cybernet. 6 (1976) 420–433.

[15] L.G. Shapiro, R.M. Haralick, Structural descriptions and inexact matching, IEEE Trans. Pattern Anal. Mach. Intell. 3 (1981) 504–519.

[16] M.A. Tanner, W.H. Wong, The calculation of posterior distributions by data augmentation (with discussion), J. Amer. Statist. Assoc. 82 805–811.

[17] W.H. Tsai, K.S. Fu, Error-correcting isomorphism of attributed relational graphs for pattern analysis, IEEE Trans. Systems Man Cybernet. 9 (1979) 757–768.

[18] S. Umeyama, An eigen-decomposition approach to weighted graph matching problems, IEEE Trans. Pattern Anal. Mach. Intell. 10 (1988) 695–703.

[19] R.C. Wilson, E.R. Hancock, Structural matching by discrete relaxation, IEEE Trans. Pattern Anal. Mach. Intell. 19 (6) (1997) 634–648.

[20] S.C. Zhu, C.E. Guo, Conceptualization and modeling of visual patterns, International Workshop on Perceptual Organization in Computer Vision, Vancouver, Canada, 2001.

[21] S.C. Zhu, Y.N. Wu, D.B. Mumford, Minimax entropy principle and its applications to texture modeling, Neural Comput. 9 (1997) 1627–1660.