Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Exploiting multi-context analysis in semantic image classification*

TIAN Yong-hong (田永鸿)[1], HUANG Tie-jun (黄铁军)[1,2], GAO Wen (高 文)[1,2]

([1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China)

([2]Graduate School of Chinese Academy of Sciences, Beijing 100039, China)

E-mail: yhtian@ict.ac.cn; tjhuang@ict.ac.cn; wgao@ict.ac.cn

Received Aug. 5, 2005; revision accepted Sept. 10, 2005

Abstract: As the popularity of digital images is rapidly increasing on the Internet, research on technologies for semantic image classification has become an important research topic. However, the well-known content-based image classification methods do not overcome the so-called semantic gap problem in which low-level visual features cannot represent the high-level semantic content of images. Image classification using visual and textual information often performs poorly since the extracted textual features are often too limited to accurately represent the images. In this paper, we propose a semantic image classification approach using multi-context analysis. For a given image, we model the relevant textual information as its multi-modal context, and regard the related images connected by hyperlinks as its link context. Two kinds of context analysis models, i.e., cross-modal correlation analysis and link-based correlation model, are used to capture the correlation among different modals of features and the topical dependency among images induced by the link structure. We propose a new collective classification model called relational support vector classifier (RSVC) based on the well-known Support Vector Machines (SVMs) and the link-based correlation model. Experiments showed that the proposed approach significantly improved classification accuracy over that of SVM classifiers using visual and/or textual features.

Key words: Image classification, Multi-context analysis, Cross-modal correlation analysis, Link-based correlation model, Linkage semantic kernels, Relational support vector classifier

doi:10.1631/jzus.2005.A1268 **Document code:** A **CLC number:** TP391

INTRODUCTION

The popularity of digital images is rapidly increasing due to improving digital imaging technologies, and convenient availability facilitated by the Internet. Organizing these images into categories and providing effective indexing is imperative for real-time browsing and retrieval. Typically, existing image classification work such as that of Vailaya *et al.*(2001) follows the paradigm of content-based image retrieval (CBIR) technologies, i.e., representing images using a set of low-level visual features such as colour, texture and shape, and grouping visually similar images as training images. An image may supply much information from which many different concepts or ideas can be extracted. Users typically do not think in terms of low-level features. As a result, most of these systems have poor classification performance since low-level visual features cannot represent the high-level semantic content of images.

To overcome the so-called semantic gap, some current research efforts (e.g., Chen *et al.*, 2001; Zhao and Grosky, 2002; Paek *et al.*, 1999) focus on combining low-level features and high-level features for semantic image classification and retrieval, in which the text information (e.g., image annotations, or surrounding texts on the Web pages that contain the images) can be used as potential high-level semantic features to represent the images. However, a pure combination of traditional text-based and content-based approaches is not adequate for dealing with the problem of image classification and retrieval

* Project supported by the Hi-Tech Research and Development Program (863) of China (No. 2003AA119010), and China-American Digital Academic Library (CADAL) Project (No. CADAL2004002)

on the WWW (Chen *et al.*, 2001), mainly because of the following difficulties:

(1) The textual feature source problem. How to obtain high-level textual features is a key issue. Clearly, image annotation is a tedious process. Moreover, it is often difficult to make exactly the annotations on the images (Chen *et al.*, 2001). Many works (Chen *et al.*, 2001; Zhao and Grosky, 2002; Cai *et al.*, 2004; Wang *et al.*, 2004) use the text content from the document that contains an image as the semantic features of that image. For example, the image URLs and filenames, page titles, ALT text, and surrounding text on the Web pages can be extracted to represent the images on the same pages. However, the available textual features are usually less accurate than annotating text since there is already too much clutter and irrelevant information on the Web pages (Chen *et al.*, 2001). Moreover, some Web images have few or even no surrounding texts. Therefore, we should resolve the problem of noisy information and few surrounding text in the textual feature extraction.

(2) The very high feature dimensionality. In general, the dimensionality of textual features is often much higher (even up to 2000~5000). Zhao and Grosky (2002) utilized latent semantic indexing (LSI) technique to reduce the dimensionality of textual features and to improve the retrieval performance of Web documents. However, their conclusion cannot be directly applied in other Web image collections since in their experiments only 43 keywords are extracted to represent the images.

(3) The cross-modal correlation. Once we extract the visual and textual features of images, we can combine the two kinds of feature vectors into a high-dimensional vector (Zhao and Grosky, 2002), or calculate the similarities based on the visual and textual features separately and then use the linear combination of these two similarities for image retrieval (Chen *et al.*, 2001). In these approaches, the textual features are treated as additional features, and the different types of features remain unchanged during the learning process. Therefore, the correlations among different modals of features are not fully explored.

Alternatively, some researchers investigated how to exploit link information to improve the performance of image clustering and retrieval. The underlying fundamental premise is that images which are co-contained in pages are likely to be related to the same topic, and images which are contained in pages that are co-cited by a certain page are likely related to the same topic (Lempel and Soffer, 2002). PicASHOW is such a Web image retrieval system that is based on several link analysis algorithms (Lempel and Soffer, 2002). It can retrieve relevant images even when those are stored in files with meaningless names. Cai *et al.*(2004) exploited visual, textual and link information to hierarchically cluster Web image search results. By exploiting link information, one can also explore the inter-relationships between Web images and their textual annotations to improve Web image retrieval (Wang *et al.*, 2004).

However, the link regularities in many real-world link data such as Web pages are very complex. For example, pages with the same class tend to link to pages that are topically similar to each other, but also link to a wide variety of other pages without semantic reason (Yang *et al.*, 2002). The presence (or absence) of such a complex regularity may significantly influence the optimal design of a link-based Web image classification or retrieval model. However, the above approaches do not automatically identify which links are most relevant to the task. As a result, this lack of selectivity will make the models more difficult to be practically applicable (Neville and Jensen, 2003). Therefore, what is important is to be able to effectively capture such complex regularity in link-based image classification models so that they can be robust in the real-world environment.

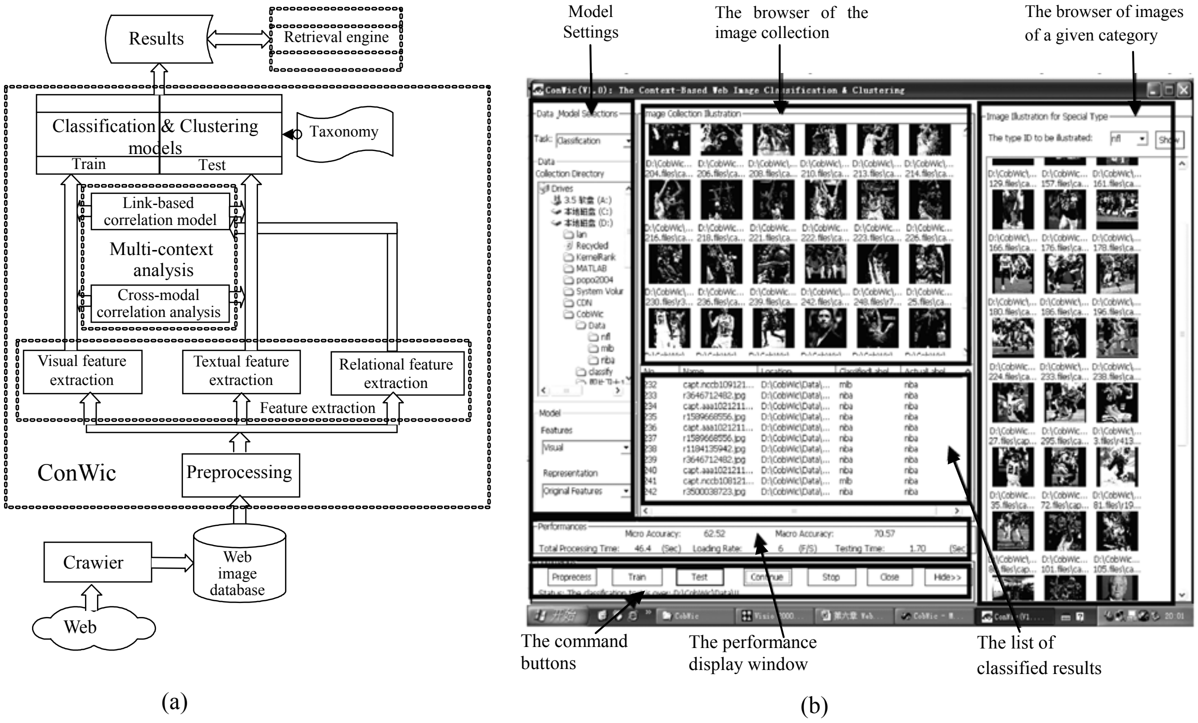
In this paper, we propose a semantic image classification approach using multi-context analysis. A fundamental aspect of our approach is the explicit use of the context. For a given image, it models the relevant textual information as its multi-modal context, and regards the related images connected by hyperlinks as its link context. Two kinds of context analysis models, i.e., cross-modal correlation analysis (CMC) and link-based correlation model (LCM), are used to capture the correlation among different modals of features and the topical dependency among images that is induced by the link structure. Specifically, instead of directly using link information for classification, this paper explores how to use linkage semantic kernels to reveal the semantic relationships underlying the link structure. We propose a new collective classification model called relational support vector classifier (RSVC), based on the well-known Support Vector Machines (SVMs) and the linkage semantic kernels. On a sports Web image collection crawled from Yahoo!, the experiments showed that the proposed approach achieved significant improvement in classification accuracy over SVM classifiers using visual and/or textual features. The proposed approach has been implemented in a Web image classification prototype, ConWic.

The paper is organized as follows. Section 2 presents the system architecture of the ConWic system. We describe the visual, textual and relational representation of Web images in Section 3, and then present the multi-context analysis in Section 4. The RSVC models are described in Section 5. Experiments and results are presented in Section 6. Finally, we conclude the paper in Section 7.

OVERVIEW OF THE CONWIC SYSTEM

The ConWic system (Context-Based Web Image Classification & Clustering system) was developed to exploit visual, textual and relational information to aid classification, clustering and semantic-sensitive retrieval of Web images. The architecture of the ConWic system is shown in Fig.1a. As can be seen from this figure, the system consists of four components: the preprocessing component, the feature extraction component, the multi-context analysis component, and the classification and clustering component. The preprocessing component performs several preprocessing tasks such as the mapping between local file names and pages' URLs, the extraction of images' properties from pages, and the splitting of train/test images. In the feature extraction component, we have three modules, namely, the visual, textual and relational feature extractors. In the multi-context analysis component, cross-modal correlation analysis is used to reveal the correlationship among different types of features, and link-based correlation analysis is used to capture the topical correlation among images that is induced by the link structure. Finally, the classification and clustering component exploits some traditional machine learning algorithms such as SVMs or the link-based models such as RSVCs to classify or cluster the given image collection.

Fig.1b shows the main interface of the ConWic system. As an experimental prototype, the system currently does not include the image retrieval component and the user feedback interface.


Results
Retrieval engine
Classification & Clustering models
Train
Test
Taxonomy
Link-based correlation model
Multi-context analysis
Cross-modal correlation analysis
Visual feature extraction
Textual feature extraction
Relational feature extraction
Feature extraction
ConWic
Preprocessing
Crawier
Web image database
Web
(a)
Model Settings
The browser of the image collection
The browser of images of a given category
ConWic(V1.0): The Context-Based Web Image Classification & Clustering
Data Model Selections
Task: Classification
Data Collection Directory
Image Collection Illustration
Image Illustration for Special Type
The type ID to be illustrated: nfl
Show
Model
Features
Visual
Representation
Original Features
Performances
Preprocess
Train
Test
Continue
Stop
Close
Hide>>
The command buttons
The performance display window
The list of classified results
(b)

No.	Name	Location	Classified Label	Actual Label
232	capt.nccb109121...	D:\CobWic\Data\...	mlb	nba
233	r3646712482.jpg	D:\CobWic\Data\...	nba	nba
234	capt.aaa1021211...	D:\CobWic\Data\...	nba	nba
235	r1589668556.jpg	D:\CobWic\Data\...	nba	nba
236	capt.aaa1021211...	D:\CobWic\Data\...	nba	nba
237	r1589668556.jpg	D:\CobWic\Data\...	nba	nba
238	r1184135942.jpg	D:\CobWic\Data\...	nba	nba
239	r3646712482.jpg	D:\CobWic\Data\...	nba	nba
240	capt.aaa1021211...	D:\CobWic\Data\...	nba	nba
241	capt.nccb108121...	D:\CobWic\Data\...	mlb	nba
242	r3500038723.jpg	D:\CobWic\Data\...	nba	nba

Micro Accuracy: 62.52 Macro Accuracy: 70.57
Total Processing Time: 46.4 (Sec) Loading Rate: 6 (F/S) Testing Time: 1.70 (Sec)

Fig.1 The architecture (a) and main interface (b) of the ConWic system

FEATURE REPRESENTATIONS OF IMAGES

In this section, we present in detail how to represent Web images using the visual, textual and relational features. Fig.2 depicts the basic idea for the Web image representation models.

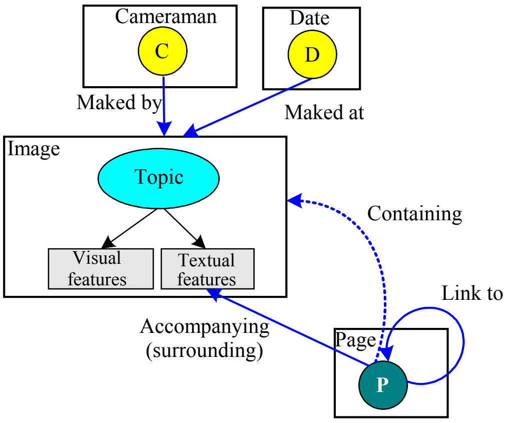


Fig.2 The basic idea of the representation model for Web images

Most existing Web mining algorithms usually treat the whole page as an indivisible node with no internal structure. As mentioned in (Tian *et al*., 2004), pages should be further divided into some logic snippets with a single topic, and this kind of logic snippets, e.g., DOM nodes (Document Object Model, http://www.w3c.org/DOM/), should be treated as the basic analysis units in Web mining tasks so as to effectively reduce the influence of noisy information in pages. Similarly, here we also need to segment each page into several finer-grain blocks, each of which contains an image and its surrounding texts. As in (Cai *et al*., 2004), we also refer to them as image blocks.

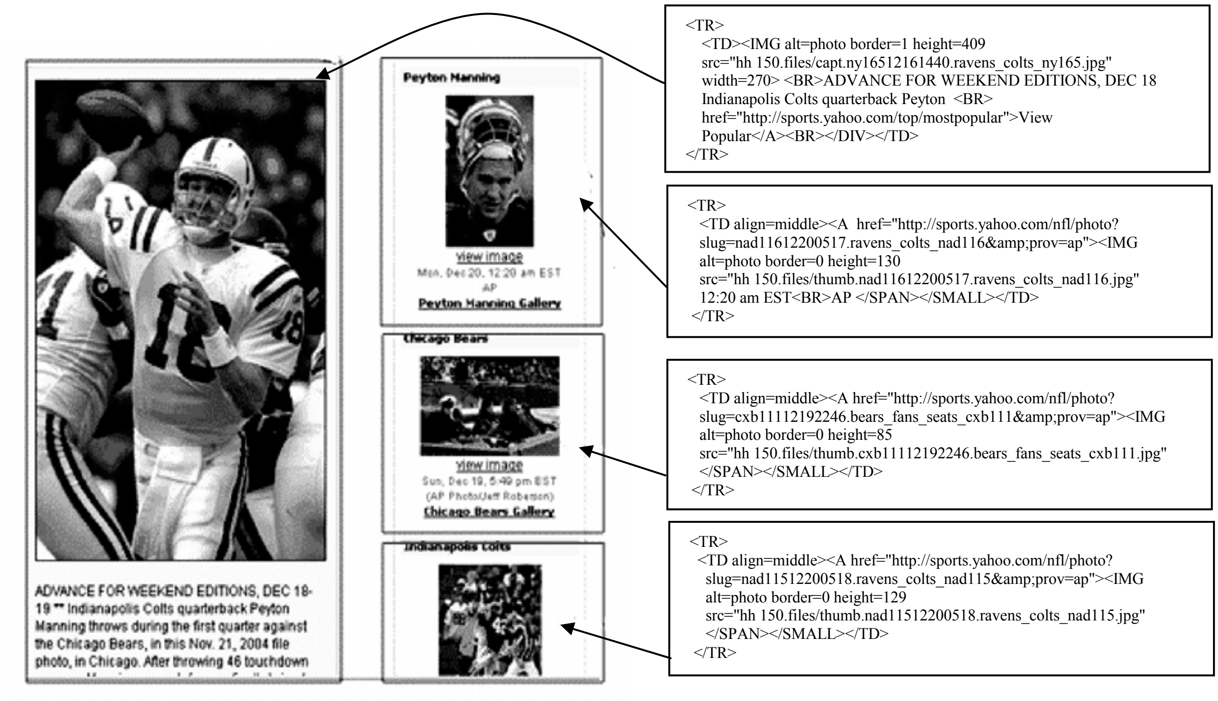
The often-used page segmentation method is directly based on DOM trees. In the HTML DOM tree, an image is always a leaf node, and thus the text of its sibling nodes can be used as the surrounding text of the image. Naturally, we can use the DOM based method to segment pages into different image blocks. Alternatively, Cai *et al*.(2003) proposed a vision-based page segmentation (VIPS) algorithm to extract the semantic structure of a Web page based on its visual presentation. The VIPS algorithm has been successfully applied in Web image clustering and retrieval systems (Cai *et al*., 2004; Wang *et al*., 2004). However, it also suffers from high complexity. For simplicity, here we use the DOM based method, which yields very satisfactory results on our gathered Web page collection. Several heuristic rules can also be used to remove some "noisy" images such as navigational bars and advertisement icons. Fig.3 shows a simple example of DOM page partitioning, in which each image block corresponds to a DOM subtree between tag <TR> and tag </TR>.

Visual feature representation

Each image $I_i \in I$ can be represented by a visual feature vector $\boldsymbol{f}_i^{(\mathrm{V})}$, where I denotes the image collection. The most widely used visual features include (Ma and Zhang, 1998): (1) color features such as color histogram, color correlogram, color moment, color coherence vector; (2) texture features such as edge histogram, co-occurrence matrix and Gabor wavelet feature; (3) shape features such as Fourier descriptor and moment invariant. In addition, some combined features such as color texture moments (Yu *et al*., 2002) can also be used. However, it is often very difficult to find one or several visual features that are robust for all types of images or for all image analysis tasks. Thus in practice, different sets of visual features may be used to represent different types of images. In our approach, we use eight classes of visual features.

Textual feature representation

For a Web image, the texts extracted from the Web page that contains that image, such as the filename and URL of the image, the image ALT in page source, the page title, the surrounding text, are usually very useful for revealing the semantic meaning of that image (Chen *et al*., 2001; Cai *et al*., 2004). However, none of all these texts are semantically related to the image. For example, we can find Web images with meaningless filenames such as "myimages/image1", or with ALT such as "photo" that do not indicate explicit semantics. So in our approach, the text features are extracted only from the surrounding text of images. When we use the DOM based method to segment each Web page into several blocks, the text of the sibling nodes can be treated as the surrounding text of the image.



Peyton Manning
View image
Mon, Dec 20, 12:20 am EST
AP
Peyton Manning Gallery
Chicago Bears
View image
Sun, Dec 19, 5:49 pm EST
(AP Photo/Jeff Roberson)
Chicago Bears Gallery
Indianapolis Colts
ADVANCE FOR WEEKEND EDITIONS, DEC 18-19 ** Indianapolis Colts quarterback Peyton Manning throws during the first quarter against the Chicago Bears, in this Nov. 21, 2004 file photo, in Chicago. After throwing 46 touchdown

```
<TR>
  <TD><IMG alt=photo border=1 height=409
  src="hh 150.files/capt.ny16512161440.ravens_colts_ny165.jpg"
  width=270> <BR>ADVANCE FOR WEEKEND EDITIONS, DEC 18
  Indianapolis Colts quarterback Peyton <BR>
  href="http://sports.yahoo.com/top/mostpopular">View
  Popular</A><BR></DIV></TD>
</TR>
```

```
<TR>
  <TD align=middle><A href="http://sports.yahoo.com/nfl/photo?
  slug=nad11612200517.ravens_colts_nad116&amp;prov=ap"><IMG
  alt=photo border=0 height=130
  src="hh 150.files/thumb.nad11612200517.ravens_colts_nad116.jpg"
  12:20 am EST<BR>AP </SPAN></SMALL></TD>
</TR>
```

```
<TR>
  <TD align=middle><A href="http://sports.yahoo.com/nfl/photo?
  slug=cxb11112192246.bears_fans_seats_cxb111&amp;prov=ap"><IMG
  alt=photo border=0 height=85
  src="hh 150.files/thumb.cxb11112192246.bears_fans_seats_cxb111.jpg"
  </SPAN></SMALL></TD>
</TR>
```

```
<TR>
  <TD align=middle><A href="http://sports.yahoo.com/nfl/photo?
   slug=nad11512200518.ravens_colts_nad115&amp;prov=ap"><IMG
   alt=photo border=0 height=129
   src="hh 150.files/thumb.nad11512200518.ravens_colts_nad115.jpg"
   </SPAN></SMALL></TD>
</TR>
```

Fig.3 An example page segmentation based on DOM tree

After the extraction of surrounding text, a textual term vector is used to represent the textual features for each image $I_i \in I$. That is, each term in the vector is weighted by its term frequency (TF),

$$\boldsymbol{f}_i^{(\mathrm{T})} = [t_{i,1}, \ldots, t_{i,j}, \ldots, t_{i,m}] \tag{1}$$

where $t_{i,j}$ is the frequency of term j appearing in the text description of image I_i, and m is the size of the term dictionary constructed from the training set. Note that for constructing the term dictionary, the stopwords and rare keywords are removed. Similar to the TFIDF method used in information retrieval, here each term can also be weighted by the factor TFIIF (Term Frequency Inverse Image Frequency),

$$\boldsymbol{f}_i^{(\mathrm{T})} = [t_{i,1}\log(N/n_1), \ldots, t_{i,j}\log(N/n_j), \ldots, t_{i,m}\log(N/n_m)] \tag{2}$$

where n_j stands for the number of images characterized by term j, and N is the total number of images.

Relational feature representation

To derive the relational feature representation of images, we first describe how to construct an image graph whose weights defined on the edges reflect structural relationships between images. To further elucidate the structural relations between images, we introduce some notations. Let $p_k \in P$ denote the kth Web page or document, and $I_i, I_j \in I$ denote two images numbered i and j, where P and I are the sets of all Web pages and all the images, $K=|P|$, $N=|I|$.

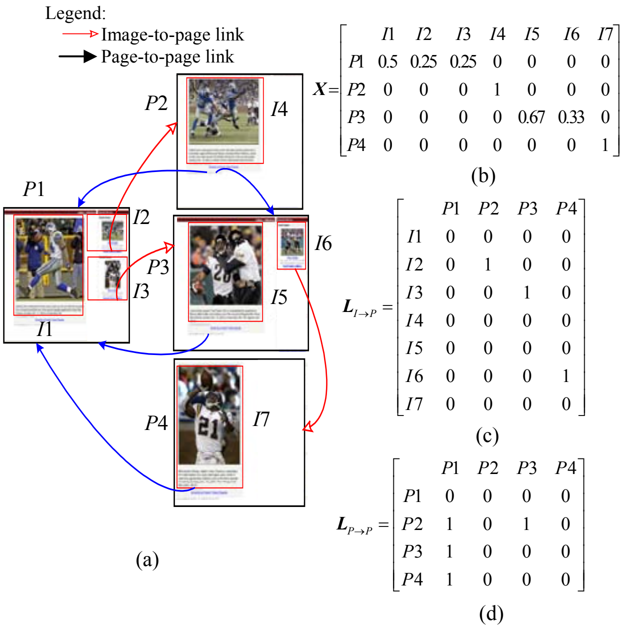
Usually, there are three kinds of structural relations in the Web image domain:

(1) $I_i \in p_k$: A page p_k contains an image I_i. For example, an image pbhp_blu.gif is contained in a file hh296.html,

<IMG border=0 height=28 src="hh296.files/pbhp_blu.gif" width=84>

Accordingly, an image-in-page matrix $\boldsymbol{X} = [x_{k,i}]_{K,N}$ can be derived to represent the relation $I_i \in p_k$. As in (Cai *et al*., 2004), $x_{k,i}$ can be set to an importance value of image I_i in page p_k. For simplicity, here we group images into two types: central images (e.g., $I1$, $I4$, $I5$, $I7$ in Fig.4) and marginal images (e.g., $I2$, $I3$, $I6$ in Fig.4). Thus, $x_{k,i}$ can be defined as:

$$x_{k,i} = \begin{cases} 2a_{p_k}, & \text{if } I_i \in p_k \text{ and } I_i \text{ is a cental image;} \\ a_{p_k}, & \text{if } I_i \in p_k \text{ and } I_i \text{ is a marginal image;} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$



$$X=\begin{bmatrix} & I1 & I2 & I3 & I4 & I5 & I6 & I7 \\ P1 & 0.5 & 0.25 & 0.25 & 0 & 0 & 0 & 0 \\ P2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ P3 & 0 & 0 & 0 & 0 & 0.67 & 0.33 & 0 \\ P4 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

(b)

$$\boldsymbol{L}_{I\to P}=\begin{bmatrix} & P1 & P2 & P3 & P4 \\ I1 & 0 & 0 & 0 & 0 \\ I2 & 0 & 1 & 0 & 0 \\ I3 & 0 & 0 & 1 & 0 \\ I4 & 0 & 0 & 0 & 0 \\ I5 & 0 & 0 & 0 & 0 \\ I6 & 0 & 0 & 0 & 1 \\ I7 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(c)

$$\boldsymbol{L}_{P\to P}=\begin{bmatrix} & P1 & P2 & P3 & P4 \\ P1 & 0 & 0 & 0 & 0 \\ P2 & 1 & 0 & 1 & 0 \\ P3 & 1 & 0 & 0 & 0 \\ P4 & 1 & 0 & 0 & 0 \end{bmatrix}$$

(d)

Fig.4 An example link graph of Web images and its relational matrices
(a) The example link graph of Web images; (b) The image-in-page matrix; (c) The image-to-page matrix; (d) The page-to-page matrix

where a_{p_k} is a normalization factor such that $\sum_{\forall I_i \in p_k} x_{k,i}=1$. Fig.4b shows the image-in-page matrix that corresponds to the link graph shown in Fig.4a.

(2) $I_j \leftrightarrow p_k$: The image block I_j has links to page p_k, or page p_k points to I_j's image file. For example, an image block in page hh296.html has a link to page hh39.html:

```
<TR>
  <TD align=middle><A href="hh39.html"><IMG
    alt=photo border=0 height=85 src="hh296.files/
    bears_fans_seats_cxb111.jpg">
  </TD>
</TR>
```

Accordingly, an image-to-page matrix $\boldsymbol{L}_{I\leftrightarrow P}=[l_{i,k}]_{N,K}$ can be derived to represent the relation $I_j \leftrightarrow p_k$: if the relation exists, $l_{i,k}$=1; otherwise $l_{i,k}$=0. Fig.4c shows an example of the image-to-page matrix.

(3) $p_j \leftrightarrow p_k$: Page p_j points to page p_k. Note that here the relation $p_j \leftrightarrow p_k$ includes neither the links in the image blocks of page p_j that point to page p_k, nor the navigational hyperlinks and advertisement hyperlinks in page p_j. For example, page hh296.html has a hyperlink to page hh291.html:

```
<A class=yspmore href="hh291.html">David Terrell
Gallery </A>
```

Accordingly, a page-to-page matrix $\boldsymbol{L}_{P\to P}=[l_{j,k}]_{K,K}$ can be derived to represent the relation $p_j \leftrightarrow p_k$: if the relation exists, $l_{i,k}$=1; otherwise $l_{i,k}$=0. Fig.4d shows an example of the page-to-page matrix.

The three relational matrices can be used as basis to facilitate construction of the page-to-image adjacency matrix among the pages in P and the images in I:

$$\boldsymbol{A}_{I\sim P}(\varepsilon,\delta)=[\varepsilon\boldsymbol{L}_{P\to P}+(1-\varepsilon)\boldsymbol{I}_K][\delta\boldsymbol{L}'_{I\leftrightarrow P}+(1-\delta)\boldsymbol{X}], \tag{4}$$

where $\boldsymbol{L}'$ denotes the transpose of $\boldsymbol{L}$, $\boldsymbol{I}_K$ is a $K\times K$ identity matrix, ε and δ are the weights, $1\geq\varepsilon\geq0$, $1\geq\delta\geq0$. In general, choosing small values of ε and δ will boost the relationship between pages and the images which they themselves contain (i.e., $I_i \in p_k$), while choosing large value of ε will introduce bias towards relations between pages and images contained in pages linked from them (i.e., those that correspond to $p_j \leftrightarrow p_k$), and choosing large value of δ will introduce bias towards relations $I_j \leftrightarrow p_k$. Clearly, $\boldsymbol{A}_{I\sim P}(0,0)=\boldsymbol{X}$, $\boldsymbol{A}_{I\sim P}(0,1)=\boldsymbol{L}'_{I\leftrightarrow P}$, $\boldsymbol{A}_{I\sim P}(1,0)=\boldsymbol{L}_{P\to P}\boldsymbol{X}$, $\boldsymbol{A}_{I\sim P}(1,1)=\boldsymbol{L}_{P\to P}\boldsymbol{L}'_{I\leftrightarrow P}$.

Naturally, we can derive the image adjacency matrix among the images in I, which naturally defines an image graph G_I.

$$\boldsymbol{A}_{I\sim I}=\boldsymbol{A}'_{I\sim P}\boldsymbol{A}_{I\sim P} \tag{5}$$

The matrix $\boldsymbol{A}_{I\sim I}$ can be used as the basis for calculation of semantically richer representations of linked images. In the following, we will develop two relational feature representation models for Web images.

The first model is referred to as linkage relationship vector (LRV) model, which is derived directly from the matrix $\boldsymbol{A}_{I\sim I}$. For a given image graph G_I, the image set I can be further divided into two parts: the set of target images $I_{(\mathrm{T})}$ and the set of background entities $I_{(\mathrm{B})}$. Usually, the set of background entities $I_{(\mathrm{B})}$ may consist of the training images and other images for which the classification is known but that are not crawled into local machines or are practically unavailable (thus cannot be used as the training images). Let $N_{(\mathrm{T})}=|I_{(\mathrm{T})}|$ and $N_{(\mathrm{B})}=|I_{(\mathrm{B})}|$. Obviously, $N= N_{(\mathrm{T})}+ N_{(\mathrm{B})}$.

Definition (LRV model)

Given an image graph G_I, each image $I_j\in I_{(\mathrm{T})}$ is represented by a weighted vector

$$\boldsymbol{f}_i^{(\mathrm{L})}=[w_{i,1},w_{i,2},\ldots,w_{i,N_{(\mathrm{T})}},w_{i,N_{(\mathrm{T})}+1},\ldots,w_{i,N}] \tag{6}$$

where $w_{i,k}$ ($1\le k\le N_{(\mathrm{T})}$) is the weight of the relationship between image I_i and target image $I_k\in I_{(\mathrm{T})}$, and $w_{i,k}$ ($(N_{(\mathrm{T})}+1)\le k\le N$) is the weight of the relationship between I_i and background entity $I_k\in I_{(\mathrm{B})}$.

There are many feasible alternatives to define the weight $w_{i,k}$ ($1\le k\le N_{(\mathrm{T})}$) from the image graph G_I. $w_{i,k}$ may be binary, representing the presence/absence of a link between image I_i and image I_k. $w_{i,k}$ may be set to $w_{i,k}=\omega_{i,k}+\omega_{k,i}$, where $\omega_{i,k}$ indicates the number of links from I_i to I_k, or represents the significance of the link $I_i\rightarrow I_k$. In this paper, $w_{i,k}$ is set to be the number (or the frequency) of linkage modes between I_i and I_k, where the linkage modes denote the important link relations that are likely to convey explicit semantic meaning, such as co-containedness, in-link, out-link, co-citation and co-reference.

The second model is derived by exploiting aggregated link features. The motivated observation is that link features that are computed based on statistics from the categories of different sets of linked objects may be more robust to irrelevant links. Thus we have the following definition of the class-based linkage relationship vector (CLRV) model.

Definition (CLRV model)

Given an image graph G_I, each image $I_i\in I_{(\mathrm{T})}$ is represented by a class-based weighted vector

$$\dddot{\boldsymbol{f}}_i^{(\mathrm{L})}=[w_{i,1},w_{i,2},\cdots,w_{i,5|C|}] \tag{7}$$

where $w_{i,k}$ ($1\le k\le 5|C|$) are the weighted frequencies of the five important link relations (i.e., co-containedness, in-link, out-link, co-citation and co-reference) between image I_i and its neighboring images of different classes, $|C|$ is the number of the class taxonomy C.

Two additional advantages of the CLRV representation are as follows: it can be significantly more compact than storing the LRV matrix; and it can accommodate the introduction of new images, and thus is applicable in a wider range of situations.

MULTI-CONTEXT ANALYSIS

In the ConWic system, the main tasks of multi-context analysis are: (1) to reveal the correlation among different modals of image features; (2) to capture the topical dependency among linked images. The two kinds of correlation can be then exploited to improve the performance of image classification or retrieval.

Cross-modal correlation (CMC) analysis

Among the three kinds of representations for each image, the visual and textual features can be combined into high-dimensional vectors and then be directly used for image classification or retrieval. Let n stand for the dimensions of visual features, and m stand for the dimensions of textual features, then in the joint visual-textual feature space, each image can be represented as

$$\begin{aligned}\boldsymbol{f}_i^{(\mathrm{C})}&=[\boldsymbol{f}_i^{(\mathrm{V})},\boldsymbol{f}_i^{(\mathrm{T})}]\\&=[v_{i,1},\cdots,v_{i,j},\cdots,v_{i,n},t_{i,1},\cdots,t_{i,k},\cdots,t_{i,m}]\end{aligned} \tag{8}$$

Note that since various visual and textual features can have quite different variations, we also need to normalize each feature in the joint space according to its maximum elements (or certain other statistical measurements).

However, the dimensionality of the visual or textual feature vectors is very high. And the extracted textual features are usually companied with some noisy or irrelevant information. For reducing the feature dimensionality and removing noise, an often-used method is the so-called latent semantic indexing (LSI) technique (Deerwester *et al*., 1990), which relies on singular value decomposition (SVD) of the feature matrix to capture the latent semantic structure among the matrix elements. Thus we can apply the LSI technique to reveal the latent semantic structure in the joint visual-textual feature space, as in

(Zhao and Grosky, 2002). However, LSI does not distinguish features from different modalities in the joint space, thus the optimal solution based on overall distribution may not best represent semantic relationships between features of different modalities. In (Li *et al.*, 2003), two cross-modal association analysis methods, i.e., cross-modal factor analysis (CFA) and canonical correlation analysis (CCA), were introduced to identify and measure intrinsic associations between visual and audio features. Here we adopt the CFA method to capture the best coupled patterns between visual and textual features.

The key idea underlying the CFA method is to find two orthogonal transformation matrices so that the coupled data in the two subsets of features can be projected as close to each other as possible (Li *et al.*, 2003). Let $\boldsymbol{F}_\mathrm{V}$ and $\boldsymbol{F}_\mathrm{T}$ be the visual and textual feature matrices for the images in *I*, then the transformation matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ can be obtained by solving the following optimisation:

$$\min \left\| \boldsymbol{F}_\mathrm{V}\boldsymbol{A} - \boldsymbol{F}_\mathrm{T}\boldsymbol{B} \right\|_F^2 \ \text{s.t.}\ \boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I},\ \boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}, \tag{9}$$

where $\|\cdot\|_F$ denotes Frobenius norm. According to the orthogonality of $\boldsymbol{A}$ and $\boldsymbol{B}$, we have

$$\left\| \boldsymbol{F}_\mathrm{V}\boldsymbol{A} - \boldsymbol{F}_\mathrm{T}\boldsymbol{B} \right\|_F^2 = \mathrm{tr}(\boldsymbol{F}_\mathrm{V}\boldsymbol{F}_\mathrm{V}') + \mathrm{tr}(\boldsymbol{F}_\mathrm{T}\boldsymbol{F}_\mathrm{T}') - 2\mathrm{tr}(\boldsymbol{F}_\mathrm{V}\boldsymbol{A}\boldsymbol{B}'\boldsymbol{F}_\mathrm{T}') \tag{10}$$

where tr(·) denotes the matrix trace. Thus Eq.(9) is equivalent to maximize the term $2\mathrm{tr}(\tilde{\boldsymbol{V}}\boldsymbol{A}\boldsymbol{B}'\tilde{\boldsymbol{T}}^\mathrm{T})$. It can be shown (Li *et al.*, 2003) that such matrices are given by the SVD decomposition of $\boldsymbol{F}_\mathrm{V}'\boldsymbol{F}_\mathrm{T}$, i.e., $\boldsymbol{F}_\mathrm{V}'\boldsymbol{F}_\mathrm{T} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{B}$ where $\boldsymbol{D}$ is the singular value matrix. Thus with the optimal transformation matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{F}_\mathrm{V}$ and $\boldsymbol{F}_\mathrm{T}$ can be transformed by the following equation:

$$\begin{cases} \tilde{\boldsymbol{F}}_\mathrm{V} = \boldsymbol{F}_\mathrm{V}\boldsymbol{A} \\ \tilde{\boldsymbol{F}}_\mathrm{T} = \boldsymbol{F}_\mathrm{T}\boldsymbol{B} \end{cases} \tag{11}$$

And $\tilde{\boldsymbol{F}}_\mathrm{V}$ and $\tilde{\boldsymbol{F}}_\mathrm{T}$ can then be combined into a joint feature matrix $\tilde{\boldsymbol{F}}_\mathrm{C}$. Similarly to those in LSI, the first and most important *k* vectors in $\tilde{\boldsymbol{F}}_\mathrm{V}$ and $\tilde{\boldsymbol{F}}_\mathrm{T}$ can be used to preserve the principal coupled patterns in much lower dimensions, and correspondingly irrelevant noise is removed.

A significant advantage of the CFA method is in favour of coupled patterns with high variations. However, the CFA method is based on some naive techniques such as the linear correlation model and the projected distance, which would limit its application in more complex situations. Moreover, such an approach can only be applied in the cases where the feature matrix for all the testing images is constructed offline. Instead, in this paper the two optimal transformation matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are learned from the training images in $I_{(\mathrm{B})}$, and then used as two semantic matrices to map the testing images in $I_{(\mathrm{T})}$ into another semantic space. That is, for a target image $I_i \in I_{(\mathrm{T})}$, the transformed feature vector can be represented as $\boldsymbol{f}_i^{(\mathrm{C})} = [\tilde{\boldsymbol{f}}_i^{(\mathrm{V})}, \tilde{\boldsymbol{f}}_i^{(\mathrm{T})}]$, where $\tilde{\boldsymbol{f}}_i^{(\mathrm{V})} = \boldsymbol{f}_i^{(\mathrm{V})}\boldsymbol{A}$ and $\tilde{\boldsymbol{f}}_i^{(\mathrm{T})} = \boldsymbol{f}_i^{(\mathrm{T})}\boldsymbol{B}$. This is similar to the semantic smoothing method used in (Cristianini *et al.*, 2002; Siolas and d'Alché-Buc, 2000). Without risk of confusion, this paper refers to this version of the CFA method as cross-modal correlation (CMC) analysis.

Thus if we consider the visual-textual joint space, the corresponding kernel is given by $\boldsymbol{K}_\mathrm{C} = [k(\boldsymbol{f}_i^{(\mathrm{C})}, \boldsymbol{f}_j^{(\mathrm{C})})]$, where $k(\boldsymbol{f}_i^{(\mathrm{C})}, \boldsymbol{f}_j^{(\mathrm{C})})$ is a kernel function. Here we refer to $\boldsymbol{K}_\mathrm{C}$ as the content kernel. Naturally, this kernel can be used by any "kernelized" algorithm such as SVM for Web image classification using visual and textual features.

Link-based correlation model (LCM)

In the preceding section, we derived two relational feature representations of images. Typically, the data contain informative, high-order features of some complex, non-linear relationships that may not be apparent in the raw data (Schölkopf, 2000). For example, while each edge of a link graph contains only local information about neighboring vertices, the set of all edges, i.e., the graph itself contains information about the global structure of the instance space which can be exploited to improve the classification accuracy of relational models (Gärtner, 2003). Thus to extract much more information from the link structure, kernel methods are introduced here. Such information is encoded in the linkage kernels, and defines a new metric in the original feature space, or equivalently a further mapping of the objects into another space.

Using the LRV representation of images, the simplest linkage kernel is constructed directly by the dot product $<\boldsymbol{f}_i^{(\mathrm{L})}, \boldsymbol{f}_j^{(\mathrm{L})}>$ and the corresponding linkage kernel matrix is the co-citation matrix $\boldsymbol{K}_{\mathrm{L}}^{(\mathrm{C})} = \boldsymbol{F}_{\mathrm{L}}'\boldsymbol{F}_{\mathrm{L}}$ where $\boldsymbol{F}_{\mathrm{L}}$ is the LRV link feature matrix for $I_{(\mathrm{T})}$ with dimension $N\times N_{(\mathrm{T})}$. We can also use the three popular kernel functions (Schölkopf, 2000), i.e., polynomial kernel, Gaussian kernel and sigmoid kernel, to construct the linkage kernels. However, these kernels cannot take advantage of the "natural" structure of the link data. Instead, several kernels are directly defined on the structure of the instances or on the structure of the instance space (Gärtner, 2003). For graph-like structures such as link data, the best-known kernel for this purpose is the diffusion kernel proposed by Kondor and Lafferty (2002). Following the spirit of the diffusion kernel, we proposed a semantic diffusion kernel (SDK) (Tian *et al.*, 2005). That is, a semantic proximity matrix is introduced to capture the semantic correlations among linked objects, and we perform the diffusion process on the semantic proximity matrix rather than directly on the linkage kernels. Let $\boldsymbol{S}_\lambda$ be a semantic proximity matrix that approximately captures the semantic relationships between the coordinate entities of the N-dimensional space spanned by the $N_{(\mathrm{T})}$ target images in $I_{(\mathrm{T})}$ and the $N_{(\mathrm{B})}$ background images in $I_{(\mathrm{B})}$, then under the semantic diffusion process, $\boldsymbol{S}_\lambda=\exp(\lambda\boldsymbol{S}_0)$, where $\boldsymbol{S}_0 = \boldsymbol{F}_{\mathrm{L}}\boldsymbol{F}_{\mathrm{L}}'$ is assumed to capture the initial semantic relationships. The SDK kernel can be expressed as

$$\boldsymbol{K}_{\mathrm{L}}^{(\mathrm{D})} = \boldsymbol{F}_{\mathrm{L}}'\boldsymbol{S}_\lambda\boldsymbol{F}_{\mathrm{L}} = \boldsymbol{F}_{\mathrm{L}}'\left[\exp(\lambda\boldsymbol{S}_0)\right]\boldsymbol{F}_{\mathrm{L}} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}', \quad (12)$$

where the bandwidth factor λ ($0\leq\lambda\leq1$) ensures that the longer range effects decay exponentially, $\boldsymbol{F}_{\mathrm{L}}=\boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{V}'$ is the SVD decomposition of $\boldsymbol{F}_{\mathrm{L}}$, and $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^2\exp(\lambda\boldsymbol{\Sigma}^2)$. Similar to the latent semantic kernels in text categorization (Cristianini *et al.*, 2002), we can also perform the LSI analysis on $\boldsymbol{K}_{\mathrm{L}}^{(\mathrm{C})}$ or $\boldsymbol{K}_{\mathrm{L}}^{(\mathrm{D})}$ to obtain the latent linkage semantic kernel corresponding to the latent semantic space.

On the other hand, the CLRV representation of images captures the semantic information among linked images by directly exploiting class-based aggregated link features. So to reveal the deep correlation underlying link structure, we can directly use the co-citation matrix

$$\ddot{\boldsymbol{K}}_{\mathrm{L}}^{(\mathrm{C})} = \ddot{\boldsymbol{F}}_{\mathrm{L}}'\ddot{\boldsymbol{F}}_{\mathrm{L}}, \quad (13)$$

where $\ddot{\boldsymbol{F}}_{\mathrm{L}}$ is the CLRV link feature matrix for $I_{(\mathrm{T})}$ with dimension $5|C|\times N_{(\mathrm{T})}$. We refer to it as class-based co-citation kernel (CCK). However, to calculate the CLRV feature vector for image $I_i\in I_{(\mathrm{T})}$, the label attributes of its neighboring images (including the neighboring target images) must be known. On the other hand, the prediction of the label attributes of all target images is exactly one of the main goals in calculating the link features; so this creates a circular argument. A possible solution is to use a bootstrapping step. That is, we first assign an initial category to each unlabelled image based solely on its visual and textual features, then calculate the CLRV link features, and finally exploit the CCK kernel based models for re-classifying these target images.

In summary, instead of being used for exploiting slightly different kernel construction methods, the two kinds of linkage kernels can be used to reveal the semantic relationships underlying link structure. In practice, they can be used by any "kernelized" algorithm (or model) for link data if the algorithm can be stated so that each vector of input data only appears within a dot product operation.

RELATIONAL SUPPORT VECTOR CLASSIFIERS (RSVCs)

For image classification, the ConWic system uses a support vector machine (SVM). SVMs have strong theoretical foundations and excellent empirical successes, and have been applied to tasks such as handwritten digit recognition, object recognition, and text classification. When each image is represented by visual and textual features, the content kernel $\boldsymbol{K}_{\mathrm{C}}$ can be exploited by an SVM classifier (SVC) for classification of Web images. From our point of view, SVM has the advantage of being especially well-suited for incorporating a priori knowledge by proper choice of an appropriate metric consequently leading to higher generalization capacity of the classifier (Siolas and d'Alché-Buc, 2000). As mentioned before, the metric is built from CMC analysis data.

We can also use a SVC for link-based image classification, by combining the content kernels with linkage kernels (as in Joachims *et al.*, 2001). In that case, however, we would classify the target images in $I_{(\mathrm{T})}$ separately, and consequently ignore the correlation among the unlabelled target images—the correlation endowed by these links is one of our main goals in defining the LCM model. In this work, we propose a relational support vector classifier (RSVC) model, which allows the collective classification of all the target objects together so as to take special advantage of the correlations between the labels of related entities. According to (Jensen *et al.*, 2004), collective inference can effectively improve relational classification, while relational models that do not exploit collective inference generally have much larger parameter spaces and require much larger data samples to learn relational models reliably.

Kernel combination

At this point, we have two sets of kernels: content kernel $\boldsymbol{K}_{\mathrm{C}}$ that is calculated by using visual and textual features, and linkage kernel $\boldsymbol{K}_{\mathrm{L}}$ (i.e., $\boldsymbol{K}_{\mathrm{L}}^{(\mathrm{D})}$ or $\ddot{\boldsymbol{K}}_{\mathrm{L}}^{(\mathrm{C})}$). We can combine the two kernels to obtain a valid kernel that can perform better than the other two considered separately. Joachims *et al.*(2001) validated that the combination of kernels is beneficial as long as both kernels are independent in that they do not extract the same features.

The simplest method is the convex combination of the content kernel $\boldsymbol{K}_{\mathrm{C}}$ and the linkage kernel $\boldsymbol{K}_{\mathrm{L}}$ (Joachims *et al.*, 2001), i.e.,

$$\boldsymbol{K} = (1-\beta)\boldsymbol{K}_{\mathrm{C}} + \beta\boldsymbol{K}_{\mathrm{L}}\,, \tag{14}$$

where β is a weight, $0\leq\beta\leq 1$. Alternatively, Kandola *et al.*(2002) proposed a von Neumann kernel, based on the mutual reinforcement assumption. The von Neumann kernel can be treated as a non-linear combination of two kernels. However, this method also suffers from much higher complexity due to the iterative computation. We thus do not intend to apply it in this paper.

RSVCs for collective classification

Similarly to the topographic SVM in (Mohr and Obermayer, 2005), we propose a RSVC model using the composite kernel for collective classification. If the vector $\boldsymbol{\alpha}$ and the scalar b are the parameters of the hyperplane learned from the training images, then the RSVC decision rule is defined as:

$$\begin{aligned} y_i &= \operatorname{sgn}\left(\sum_{j=1}^{l}\alpha_j y_j\Big[(1-\beta)\boldsymbol{K}_{\mathrm{C}}(\boldsymbol{f}_i^{(\mathrm{C})},\boldsymbol{f}_j^{(\mathrm{C})}) \right. \\ &\qquad \left. +\,\beta\boldsymbol{K}_{\mathrm{L}}(\boldsymbol{f}_i^{(\mathrm{L})},\boldsymbol{f}_j^{(\mathrm{L})})\Big]+b\right) \\ &= \operatorname{sgn}\left(\beta\sum_{j=1}^{l}\alpha_j y_j\boldsymbol{K}_{\mathrm{L}}(\boldsymbol{f}_i^{(\mathrm{L})},\boldsymbol{f}_j^{(\mathrm{L})})+\theta_i\right) \end{aligned} \tag{15}$$

where $\theta_i = (1-\beta)\sum_{j=1}^{l}\alpha_j y_j\boldsymbol{K}_{\mathrm{C}}(\boldsymbol{f}_i^{(\mathrm{C})},\boldsymbol{f}_j^{(\mathrm{C})})+b$ which is the decision function of a conventional SVM, and l denotes the number of support vectors (SVs). Here each image is represented by $\left(<\boldsymbol{f}_i^{(\mathrm{C})},\boldsymbol{f}_i^{(\mathrm{L})}>, y_i\right)$, where $\boldsymbol{f}_i^{(\mathrm{C})}$ and $\boldsymbol{f}_i^{(\mathrm{L})}$ are its content feature vector and link feature vector respectively, y_i is its class label. Note that when a RSVC is trained using the composite kernel, the resulting SVs will still contain the relevant information about the content features, and the link feature information required good distinction of the classes.

However, the situation is different when we use $\ddot{\boldsymbol{K}}_{\mathrm{L}}^{(\mathrm{C})}$ as the linkage kernels. As mentioned before, the calculation of the CLRV link features needs the neighboring kernel label attributes of $I_i\in I_{(\mathrm{T})}$. Thus to collectively classify the images in $I_{(\mathrm{T})}$, an iterative approach is used to achieve a self-consistent solution to the classification problem. We denote the label at step τ as $y_i\big|_\tau$, and use $\ddot{\boldsymbol{f}}_i^{(\mathrm{L})}\Big|_\tau$ to denote the CLRV link feature at step τ. Then at each step τ new labels are assigned according to

$$y_i\big|_\tau = \operatorname{sgn}\left(\beta\sum_{j=1}^{l}\alpha_j y_j\ddot{\boldsymbol{K}}_{\mathrm{L}}^{(\mathrm{C})}\left(\ddot{\boldsymbol{f}}_i^{(\mathrm{L})}\Big|_{\tau-1},\ddot{\boldsymbol{f}}_j^{(\mathrm{L})}\Big|_{\tau-1}\right)+\theta_i\right) \tag{16}$$

where $y_i\big|_{\tau=0}=\operatorname{sgn}(\theta_i)$ and θ_i do not change with τ. This leads to an iterative assignment of new labels: at step τ=0. The results from a conventional SVC (β=0) are used to initialize the labels; at the following steps, the estimates of the neighboring labels are available from the previous iteration. And a criterion may be used to determine whether the iteration process will be terminated, i.e., $y_i|_\tau = y_i|_{\tau-1}, \forall I_i \in I_{(\mathrm{T})}$ with a minimal τ.

EXPERIMENTS

Overview

Several sets of experiments were designed to evaluate the classification performance of the ConWic system. When images are represented by visual and/or textual features, we use SVMs for classification; while when images are represented by the three representations, we use RSVCs for classification. In the experiments, our main goal was to demonstrate the utility of linkage semantic kernels in Web image classification, and show the superior performance of RSVC models over the SVM models using only visual and/or textual features. For all experiments, we evaluate the classification performance on the basis of accuracy. Results reported are based on averaging at least four independent tests.

All the data used in our experiments are crawled from Yahoo! sports site (http://sports.yahoo.com/). The dataset contains approximately 4419 images grouped into three types (i.e., basketball, football and baseball). As in (Cai *et al*., 2004), we filtered those images whose width and height are both smaller than 60 pixels, and those images whose ratios between width and height are greater than 5 or smaller than 1/5. The navigational hyperlinks and advertisement hyperlinks were also removed using several heuristic rules.

Obviously, the different types of images are visually similar (e.g., in the players' wear, or in the playfield), particularly for the football and baseball images. Therefore, it is necessary to exploit textual and link information to aid classification.

Results

1. Classification using visual feature

The first set of experiments was performed by using visual features only. In general, different visual features capture different aspects of images, and it is difficult to find one kind of visual features that are robust for all types of images. Therefore, we evaluate the performance of image classifiers that use 15 different combinations of color features, texture features and shape features. Note that all these feature combinations include the 64-bins color histogram.

Table 1 shows the experimental results indicating that when using color histogram and color coherence vector as visual features (i.e., CH+CCV) the image classifier performs best, followed by the case

Table 1 The classification performance using different visual feature combinations

Features	SVs	Fea. Dim.	Accuracy for each category			
			Football	Basketball	Baseball	Avg.
CH	227/240/362	64	59.78	60.86	72.87	64.50
CH+CM	220/168/ 327	73	64.00	66.76	74.96	68.57
CH+CM +Auto	207/149/300	98	67.11	69.41	67.90	68.14
CH+CM+EH	180/137/264	223	45.78	52.17	81.06	59.67
CH+CM+Auto+EH	169/136/255	183	36.89	66.53	77.45	60.29
CH+CM+Auto+EH+MI	166/140/249	255	32.67	64.54	79.78	58.99
CH+Auto	219/183/307	89	54.44	82.53	53.93	63.64
CH+EH	195/185/291	214	44.00	54.06	85.15	61.07
CH+EH+Auto	181/158/270	239	44.22	70.77	68.86	61.29
CH+ZM	207/216/327	100	48.67	61.19	69.10	59.65
CH+MI	217/237/352	71	60.67	61.76	71.83	64.75
CH+CM+MI	214/160/325	80	67.78	67.85	66.13	67.25
CH+CCV	193/161/265	320	73.33	66.24	76.65	72.07
CH+CCG	198/180/227	1344	59.56	60.62	76.81	65.66
CH+Gabor	227/240/362	88	59.78	60.86	72.87	64.50

CH: color histogram; CM: color moment; CCV: color coherence vector; CCG: color correlogram; Gabor: Gabor wavelet; EH: edge histogram; ZM: Zernike moment; MI: moment invariant; Auto: autocorrelation using color histogram and color moment as visual features (i.e., CH+CM), and the case using color histogram, color moment and autocorrelation as visual features (i.e., CH+CM+Auto). However, the feature dimensionality of the CH+CCV group is much larger than that of the other two cases (up to 320). It is known that the higher the feature dimensionality is, the longer the training time of the SVM classifier is. Thus to trade-off accuracy and training time, we use the CH+CM group as visual features in the following experiments.

It should be noted that we cannot assure that the CH+CM feature group is the optimal set of visual features on the dataset. And the main goal of our experiments here is not to find such an optimal set of visual features. However, it is safely concluded that there is plenty of room for improvement for Web image classification using visual features only.

2. Classification using visual and textual features

The second set of experiments is aimed at evaluating the classification performance by using textual features. As mentioned before, the textual features of images can be represented by TF model and TFIIF model. Fig.5a compares classification accuracies using TF and TFIIF textual features. Interestingly, the difference in classification accuracy between the two representation models of textual features is not significant. Moreover, the classification accuracy using textual features only is rather low (about 50%), which is much beyond our expectation. An important reason is that there are about 60% images which have few or even no surrounding texts, so that most elements in each textual feature vector are zero. We have also performed experiments by including textual features extracted from other sources such as image filename, ALT, but the results being quite similar seems to indicate that Web image classification using textual features only often performs poorly.

Fig.5b compares the classification accuracies using visual, textual and visual-textual features respectively. Note that here the textual features are represented by the TFIIF model. Averagely, Web image classification using visual-textual features outperforms that using visual features by about 7.6%. Clearly, combining visual and textual features can indeed improve the performance of Web image classification, but the improvement is still not significant.

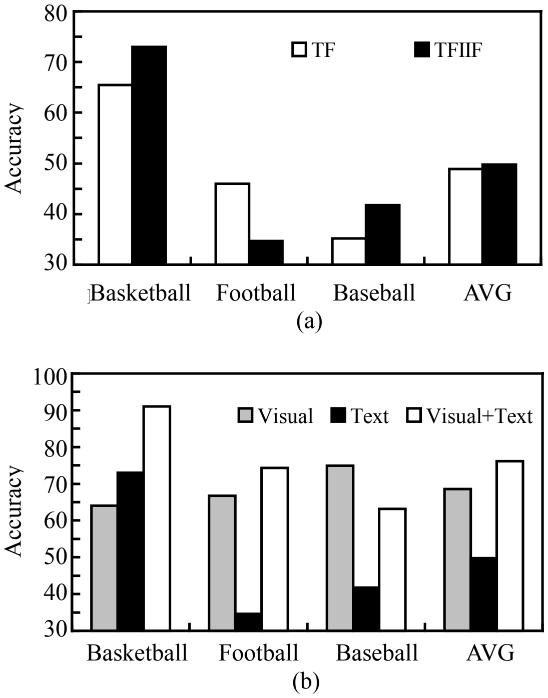


Fig.5 The classification performance using textual feature (a) Using TF and TFIIF textual features; (b) Using textual, visual and textual-visual features

For example, the classification accuracy for the baseball images using visual-textual features is even lower than that using visual features only. We can also see that the visual features are the most important determinant of Web image classification performance when using both visual and textual features.

3. Classification using cross-modal correlation model

We also performed experiments to evaluate the effects of the LSI and CMC analysis on Web image classification. When images are represented by textual features, we can exploit the LSI analysis to reduce the feature dimensionality and remove the noisy information; when images are represented by both visual and textual features, both LSI and CMC can be used.

Table 2 shows the average dimension-reduction ratios. We can see that both LSI and CMC can effectively reduce the dimensionality of the feature space used. Comparatively, the CMC analysis has larger dimension-reduction ratio (DR). For example, when using visual-textual features, the average dimension-reduction ratios are 3.4 for the LSI analysis, and 5.7 for the CMC analysis. Note that here we use the same eigengap based method to determine an appropriate *k* value for LSI and CMC.

On the other hand, we found that the two techniques surprisingly yielded no improvement in classification accuracy when using visual and textual features (Fig.6). Obviously, the two techniques cannot capture latent semantic relationships between features of different modalities. A possible reason is the highly uneven distribution of textual features, which is the main difficulty for correlation analysis of different modalities. Therefore, how to develop more robust and effective CMC analysis technique will be a future research topic.

Table 2 The comparison of average dimension-reduction ratios

	LSI (textual)	LSI (visual-textual)	CMC (visual-textual)
Before	730	827	827
After	216	246	144
DR	3.4	3.4	5.7

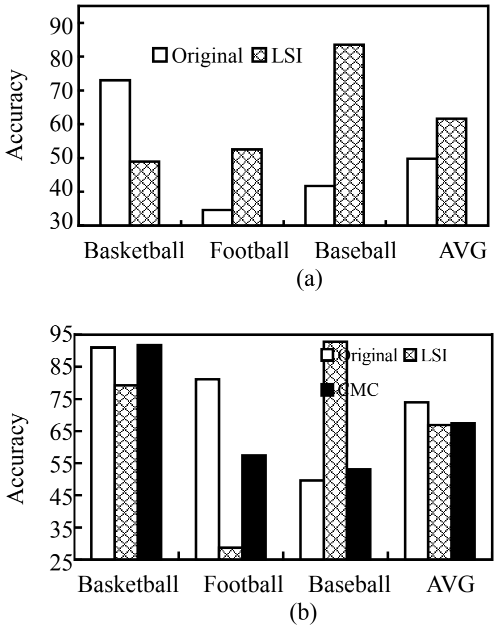


Fig.6 The classification performance using LSI and CMC analysis
(a) Using textual features only; (b) Using visual and textual features

4. Classification using multi-context models

The last set of experiments was aimed at evaluating the classification performance by using linkage semantic kernels. Two kinds of linkage semantic kernels (i.e., SDKs, CCKs) are utilized in RSVC classifiers for Web image classification. They are denoted respectively by RSVC$_{SDK}$ and RSVC$_{CCK}$. The SVM classifier using visual and textual features (denoted by SVM$_{NoLink}$) is used as the baseline model. Note that in RSVC$_{SDK}$ and RSVC$_{CCK}$, the kernel combination weight β is set to 0.5.

Fig.7 shows the average classification results of SVM$_{NoLink}$, RSVC$_{SDK}$ and RSVC$_{CCK}$. We can see that the two RSVC classifiers using linkage semantic kernels yield significant improvement in accuracy over the SVM classifier using visual and textual features. Among them, the RSVC$_{CCK}$ model yields the best performance, and outperforms the SVM$_{NoLink}$ model by about 25% of classification accuracy; and the RSVC$_{SDK}$ model also outperforms the SVM$_{NoLink}$ model by about 17% of classification accuracy. Clearly, these results indicate that linkage semantic kernels are very helpful for Web image classification.

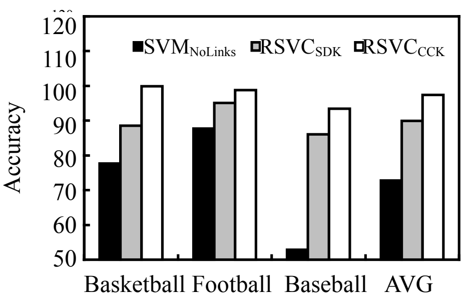


Fig.7 The average classification accuracy using linkage semantic kernels.

We note that the RSVC$_{CCK}$ model performs much better than the RSVC$_{SDK}$ model. This means that the CCK kernel is more robust in link-based Web image classification, although the aggregation operation in the calculation of CLRV link features may lose some useful link information among images. This also coincides with the SV numbers of the used SVM classifiers. In general, the less the SV number is, the better generalization the SVM classifier used has. In the experiments, the SV number of the RSVC$_{CCK}$ model is much less than that of the RSVC$_{SDK}$ model. Surprisingly, the SV number of the RSVC$_{SDK}$ model is even more than that of the SVM$_{NoLink}$ model. Therefore, how to improve the generalization of the RSVC$_{SDK}$ mode will be another research topic in the ongoing work.

In our approach, the training images are used as background entities. So we also perform several experiments to investigate the effects of the training sample distributions on the classification performance. Table 3 shows the classification performances of RSVC_{SDK} and RSVC_{CCK} under four different training sample distributions. We can see that the standard deviation in classification accuracy is 0.76% for RSVC_{CCK}, and 3.32% for RSVC_{SDK}. Comparatively, the training sample distribution has higher influence on RSVC_{SDK} than on RSVC_{CCK}.

Fig.8 shows the effect of parameter β on the classification performance. Clearly, when β=0, both RSVC_{SDK} and RSVC_{CCK} are degraded into the $\text{SVM}_{\text{NoLink}}$ model, and only visual and textual information are exploited for image classification; while when β=1, the RSVC_{SDK} and RSVC_{CCK} models only exploit link information for classification. We can see that the RSVC_{SDK} model yields the best performance when β=0.5, while the RSVC_{CCK} model shows surprisingly low dependency on β over much of its range. So for the RSVC_{CCK} model, the utilization of the CCK kernel helps for almost any value of β. While for the RSVC_{SDK} model, we should carefully select an appropriate β value for best prediction.

Discussion

In this section, we have discussed several Web image classification experiments on a collection of Web sports images by using visual, textual and link information. Two significant conclusions can be obtained:

(1) The performances of Web image classification by using a single modal of features (i.e. visual features, or textual features) are constantly at very low levels. Combining visual and textual features is helpful for Web image classification, but the improvement is not significant in some cases.

(2) Links among Web images are very useful for better classification. Linkage semantic kernels can effectively reveal the semantic relationships underlying link structure, and can be naturally embedded into kernelized relational classification algorithms such as RSVCs. Compared with the SDK kernel, the

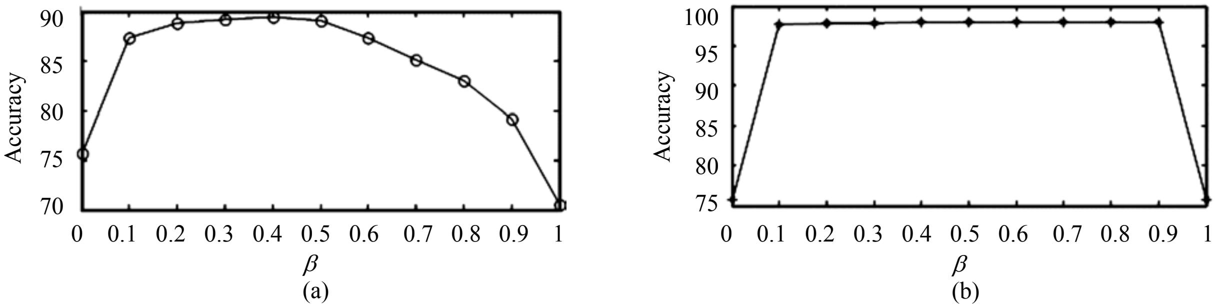


Fig.8 Effects of parameter β for the RSVC_{SDK} (a) and RSVC_{CCK} (b) models

Table 3 The classification performance of link-based models under different training sample distributions

Train samples	Model	SVs	Accuracy for each category			
			Football	Basketball	Baseball	Avg.
133/466/312	RSVC_{SDK}	230/258/262	80.13	97.64	77.88	85.22
	RSVC_{CCK}	68/57/66	100.0	99.56	93.95	97.84
135/165/214	RSVC_{SDK}	229/209/207	96.01	87.81	91.57	91.80
	RSVC_{CCK}	63/38/57	99.79	98.64	94.25	97.56
129/146/146	RSVC_{SDK}	211/190/194	78.42	96.04	79.73	84.73
	RSVC_{CCK}	55/41/64	100.0	98.13	90.65	96.26
134/168/189	RSVC_{SDK}	201/203/214	92.03	83.60	90.75	88.80
	RSVC_{CCK}	55/ 47/58	99.79	98.82	95.04	97.88
Deviation	RSVC_{SDK}					3.32
	RSVC_{CCK}					0.76

CCK kernel is more robust in Web image classification tasks.

In the ongoing experiments, we need to evaluate the performance of the RSVC on different image classification tasks such as the identification of sensitive images and artistic images.

CONCLUSION

In this paper, we present a context-based Web image classification system, ConWic. The main objective of ConWic system is to exploit the visual, textual and link information to aid the classification of Web images. Our main contributions are summarized as follows:

First, a multi-context analysis method is introduced into Web image classification tasks. For a given image, we model the relevant textual information as its multi-modal context, and regard the related images connected by hyperlinks as its link context. Two kinds of context analysis models, i.e., CMC analysis and LCM model are exploited to capture the correlation among different modals of features and the topical dependency among images that is induced by the link structure.

Second, we propose a new collective classification model called RSVC, which is based on the linkage semantic kernels and a self-consistent solution to the label assignment. A significant advantage of the RSVC is that the relational information can be utilized for classification in an SVM-like manner but the correlations between the labels of related entities can also be explicitly exploited.

On a sports Web image collection crawled from Yahoo!, the RSVC classifiers using the two linkage semantic kernels yield significant improvement in accuracy over the SVM classifier using visual and/or textual features. The experimental results demonstrate the general applicability of the RSVC classifiers using the linkage semantic kernels.

References

Cai, D., Yu, S., Wen, J.R., Ma, W.Y., 2003. VIPS: A Vision-base Page Segmentation Algorithm. Technical Report, MSR-TR-2003-79, Microsoft Research Asia.

Cai, D., He, X.F., Li, Z.W., Ma, W.Y., Wen, J.R., 2004. Hierarchical Clustering of WWW Image Search Results Using Visual, Textual and Link Analysis. Proceedings of 12th ACM International Conference on Multimedia, New York, USA, p.952-959.

Chen, Z., Liu, W.Y., Zhang, F., Li, M.J., Zhang, H.J., 2001. Web mining for Web image retrieval. *Journal of American Society of Infomation Science and Technology*, **52**(10):831-839.

Cristianini, N., Shawe-Talyor, J., Lodhi, H., 2002. Latent semantic kernels. *Journal of Intelligent Information Systems*, **18**(2/3):127-152.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. *Journal of American Society of Infomation Science and Technology*, **41**(6):389-401.

Gärtner, T., 2003. A survey of kernels for structured data. *SIGKDD Explorations*, **5**(1):49-58.

Jensen, D., Neville, J., Gallagher, B., 2004. Why Collective Inference Improves Relational Classification. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, p.593-598.

Joachims, T., Cristianini, N., Shawe-Talyor, J., 2001. Composite Kernels for Hypertext Categorization. Proceedings of the International Conference on Machine Learning (ICML-2001), Morgan Kaunfmann Publishers, San Francisco, US. p.250-257.

Kandola, J., Shawe-Talyor, J., Cristianini, N., 2002. Learning Semantic Similarity. Proceedings of International Conference on Advances in Information Processing System (NIPS).

Kondor, R.I., Lafferty, J., 2002. Diffusion Kernels on Graphs and Other Discrete Structures. Proceedings of the 19th International Conference on Machine Learning (ICML02), p.315-322.

Lempel, R., Soffer, A., 2002. PicASHOW: pictorial authority search by hyperlinks on the Web. *ACM Transactions on Information Systems*, **20**(1):1-24.

Li, D.G., Dimitrova, N., Li, M.K., Sethi, I.K., 2003. Multimedia Content Processing through Cross-modal Association. Proceedings of 11th ACM International Conference on Multimedia, Berkeley, California, USA, p.604-611.

Ma, W.Y., Zhang, H.J., 1998. Content-based Image Indexing and Retrieval. *In*: Furht, B.(Ed.), Handbook of Multimedia Computing. CRC Press, Boca Raton, FL.

Mohr, J., Obermayer, K., 2005. A Topographic Support Vector Machine: Classification Using Local Label Configurations. Advances in Neural Information Processing Systems 17, MIT Press, Cambridge, MA, p.929-936.

Neville, J., Jensen, D., 2003. Collective Classification with Relational Dependency Networks. Proceedings of 2nd Multi-Relational Data Mining Workshop, 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, p.77-91.

Paek, S., Sable, C.L., Hatzivassiloglou, V., Jaimes, A., Schiffman, B.H., Chang, S.F., McKeown, K.R., 1999.

Integration of Visual and Text-Based Approaches for the Content Labeling and Classification of Photographs. Proceedings of ACM SIGIR'99, Workshop on Multimedia Indexing and Retrieval.

Schölkopf, B., 2000. Statistical Learning and Kernel Methods. Technical Report, MSR-TR-2000-23, Microsoft Research.

Siolas, G., d'Alché-Buc, F., 2000. Support Machine Learning Based on Semantic Kernel for Text Categorization. Proceedings of the International Joint Conference on Neural Network (IJCNN).

Tian, Y.H., Huang, T.J., Gao, W., 2004. Two-phase Web site classification based on Hidden Markov Tree models. *International Journal: Web Intelligence and Agent System*, **4**(2):249-264.

Tian, Y.H., Huang, T.J., Gao, W., 2005. Latent linkage semantic kernels for collective classification of link data. *Journal of Intelligent Information Systems* (in Press).

Vailaya, A., Figueiredo, M., A.T., Jain, A.K., Zhang, H.J., 2001. Image classification for content-based indexing. *IEEE Trans. Image Processing*, **10**(1):117-129.

Wang, X.J., Ma, W.Y., Xue, G.R., Li, X., 2004. Multi-Model Similarity Propagation and its Application for Web Image Retrieval. Proceedings of 12th ACM International Conference on Multimedia, New York, USA, p.944-951.

Yang, Y., Slattery, S., Ghani, R., 2002. A study of approaches to hypertext categorization. *Journal of Intelligent Information System*, **18**(2/3):219-241.

Yu, H., Li, M., Zhang, H.J., Feng, J., 2002. Color Exture Moments for Content-based Image Retrieval. International Conference on Image Processing, p.28.

Zhao, R., Grosky, W.I., 2002. Narrowing the emantic gap—improved text-based Web document retrieval using visual features. *IEEE Trans. Multimedia*, **4**(2):189-200.