Identifying Story and Preview Images in News Web Pages

Jianying Hu Amit Bagga
Avaya Labs Research
233 Mount Airy Road
Basking Ridge, NJ 07920
{jianhu, bagga}@avaya.com

Abstract

The World Wide Web provides an increasingly powerful and popular publication mechanism. Web documents often contain a large number of images serving various different purposes. This paper focuses on images that are associated with a story or preview to a story. Such images often accompany the key content on a web page, thus their identification is important for applications such as web page summarization and mobile access. We present a novel algorithm for automatic identification of story/preview images which combines features extracted from both the image itself and the surrounding text. The effectiveness of this algorithm is demonstrated by experimental results on over 1500 images collected from 25 news web sites.

1. Introduction

The World Wide Web as an on-line publication mechanism has become increasingly multimedia. Many web documents contain a large number of images, and these images tend to be highly heterogeneous in terms of their functionalities. For example, a news web page may contain images corresponding to specific news stories, icons (*e.g.*, an image representing a sunny forecast), logos, ads, and images containing mostly text serving as section headings, etc.

Among the various images in a web documents, those associated with a story or the preview to a story are of particular interest. This is because these images are often related to the key content of the page. Identifying these images enables a web summarization/re-authoring system [2, 14, 4] to give them higher priority for transmission to a mobile device.

While there have been some research activities on the analysis of web images, they have been largely focused on two particular aspects. One is the extraction and recognition of text contained in web images [12, 1]. The other is image search and retrieval on the web [6, 17]. There has been no previous study on functionality based image categorization.

Table 1. Image categories in news web pages.

Category	# of Images	Percentage
Story (S)	91	10.1
Preview (P)	16	1.8
Host (A)	9	1.0
Commercial (C)	110	12.2
Icons and Logos (I)	293	32.6
Headings (H)	198	22.0
Formattings (F)	182	20.3

In this paper we present an algorithm for automatic identification of story and preview images. The algorithm makes use of both visual image features and information contained in the surrounding text to separate story and preview images from ads, icons, formatting images, etc. We report experimental results which demonstrate the effectiveness of this algorithm.

2. Algorithm Overview

In a previous study, we manually analyzed 899 web images found in the front pages of 25 randomly selected news web sites and identified seven functional categories [7]. The categories and their distributions are listed in Table 1. Figure 1 shows some example images in their respective surrounding contexts. As seen from Table 1, story and preview images only make up less than 12% of the collection.

A quick study of the image categories reveals that the seven categories can be grouped into two *super classes*. The fist super class, denoted SPA, includes categories Story, Preview and Host (images of hosts of regular columns or programs). The second one, denoted CIHF, contains the rest of the categories: Commercial, Icons and Logos, Heading and Formmatings. The SPA class is more likely to contain photographic images of "regular" aspect ratios, and they are often associated with some story. On the other hand, images in the CIHF class are more likely to be graphic, have "irregular" aspect ratios (e.g, extremely long or wide), and are

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE

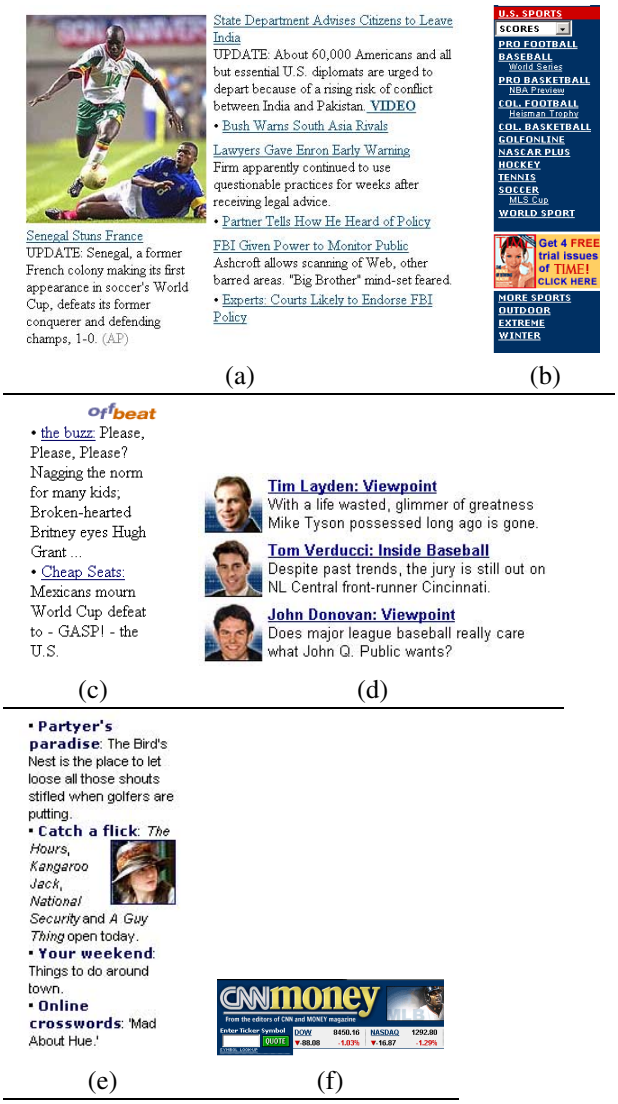

Figure 1. Examples of images found on the web with their contexts: (a) Story; (b) Commercial; (c) Heading; (d) Host; (e) Preview; (f) Icon/Logo.

often not associated with a story.

Based on this observation, we designed our classification procedure as follows. First, a simple size-based screening process is applied to remove very small images and images of "irregular" aspect ratios. To be more specific, an image is removed if its height is less than 20 pixels or the ratio between the larger and the smaller dimensions is greater than 2.5. Our experiments showed that this simple procedure reduces the total number of images to be considered by about half without removing any images from the desired SPA class.

For the remaining images, the main classifier to separate the SPA and CIHF classes is built using both image features and features extracted from the associated text. Then a secondary classifier using only text features is used to further separate out Host images from the SPA class. The remaining images are considered Story and Preview images.

In the following sections we describe the image features, text features, and the combined classifier in detail.

3. Photographic/Graphic Image Classification

The image characteristic that stands out most at the first glance of a web page is whether an image is photographic (also referred to as "natural"), or graphic (also referred to as "synthetic"). The first question one might ask is whether the format of the image can be used for this classification. The two most common formats used for images in web documents are GIF and JPEG. While GIF in general is more suitable for graphics and JPEG is more suitable for photographic images, such separation is not always followed in practice. In the database described in Table 1, 14 of the 91 images in category S are photographic images in GIF format, while 13 of the 127 JPEG images are graphic images. Other researchers have also observed similar mixture of image classes within each single format [6]. Thus, image format cannot be used as the definitive indication of photographic *vs.* graphic images.

Much of previous research on image based classification in the document analysis community has been focused on the classification of text *vs.* non-text regions within an image [11, 10], predominantly using frequency domain analysis of image intensity. Some researchers have also explored the color characteristics of different classes of images for classification. Swain *et al.* proposed using features such as degree of color saturation and number of dominant colors to separate photographic and graphic images [6]. Lopresti and Zhou [12] used similar features to identify text regions in web images.

After investigating the two different approaches described above, it became clear to us that they are complimentary to each other and a combination of the two would likely lead to improved performance. We thus designed a new algorithm incorporating features from both the frequency domain and the color domain.

3.1. Frequency Domain Features

Much of the characteristics separating photographic and graphic images are reflected in spatial features of image intensity. To exploit such characteristics, we derive features from the DCT (Digital Cosine Transform) coefficients of 8×8 subregions (blocks) of an image. Such features have been used successfully before for text *vs.* non-text image classification [15, 11, 10]. The main innovation in our algorithm is that a clustering procedure is applied first to handle the fact that graphic images in general are much less uniform compared to text images.

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE

The 8×8 DCT results in 64 coefficients. A subset of these are selected using a discriminative analysis carried out on data extracted from a set of training images. First the absolute values of the coefficients are taken, which we will refer to as *absolute coefficients* hereafter. The values corresponding to each absolute coefficient are then normalized by the standard deviation. To estimate the class discriminative power of each coefficient, we compute the within class and between class variances as following. Suppose there are a total of n_p photographic image blocks and n_g graphic image blocks. Let $P = \{p_1, p_2, ..., p_{n_p}\}$ and $G = \{g_1, g_2, ..., g_{n_g}\}$ represent the indexes of photographic and graphics image blocks, respectively. Let α_k refer to the absolute value of the kth DCT coefficient. Let $(\bar{\alpha_k})_P$ and $(\bar{\alpha_k})_G$ represent the means of α_k over set P and G respectively. The within class variance is defined as:

$$\sigma_k{}^2 = \frac{1}{n_p + n_g}(\sum_{j \in P}((\alpha_k)_j - (\bar{\alpha_k})_P)^2 + \sum_{j \in G}((\alpha_k)_j - (\bar{\alpha_k})_G)^2).$$

The between class variance is defined as:

$$\tau_k{}^2 = \frac{1}{n_p + n_g}(\sum_{j \in P}((\alpha_k)_j - (\bar{\alpha_k})_G)^2 + \sum_{j \in G}((\alpha_k)_j - (\bar{\alpha_k})_P)^2).$$

And the discriminative power of the kth coefficient is measured by: $\delta_k = \tau_k / \sigma_k$. The top $M < N$ coefficients with largest δ_k are then selected as the DCT features.

While DCT features similar to that described above were used directly with success in past efforts to classify an image block as text or non-text, our experiments showed that such a strategy does not work well for photographic and graphic image classification. This is because both categories contain a large range of different image blocks. For example, while graphic images tend to contain sharper edges, they often contain uniform blocks as well. On the other hand, photographic images sometimes contain regions of high frequency variation such as scene text and fences, as well as the more typical smooth-transition areas.

To accommodate the large variation within each class, we apply unsupervised clustering on the training image blocks using the M selected DCT coefficients. To be specific, the K-means clustering algorithm [8] is used to group the training image blocks into a predetermined number of K clusters. Each training image block is then labeled by its cluster index. Finally a normalized cluster histogram is computed for each image, yielding a K dimensional feature. Parameters M and K are chosen empirically and we settled on $M = 18$ and $K = 15$ in our experiments. For classification, each image block is assigned to the cluster with nearest cluster center and the same K dimensional cluster histogram is computed and used as the feature representing the whole image.

3.2. Color Features

Swain *et al.* proposed 8 color related features to distinguish graphic and photographic images [6]. A study of those features revealed that many of them are various heuristic ways of implementing aspects of the frequency domain characteristics that are better captured by the frequency domain features described above. We thus selected 2 of the color features that are completely independent from the frequency domain features and thus add most discriminative power. These two features are summarized below for completeness.

The first feature is called the "band different feature". A threshold T between 0 and 255 is selected and a counter C is initialized to 0. For each pixel in the image, if the difference between the largest and the smallest RGB components is greater than T, then the counter C is increased by one. After all pixels in the images have been examined, the band difference feature is calculated as C/S where S is the total number of pixels in the images. This feature has a range of $[0, 1]$ and is a rough measure of the degree of color saturation in the image. Graphic images tend to get higher values since they tend to contain purer colors. We chose $T = 50$ as suggested in the original paper.

The second feature is called "the most common colors feature". Given a predetermined number N, the N most common colors in the images are found. Then the feature is simply defined as the fraction of pixels in the images that have one of those colors. This feature again as a range of $[0, 1]$ and is a rough measure of the degree of color concentration. Again, graphic images tend to get higher values since they are often dominated by a few colors. We chose $N = 10$ in our experiments.

3.3. Combining the Image Features

We adopted a two stage approach to combine the 18 frequency features with the 2 color features. First a frequency domain classifier is trained using the 18 DCT features. The same classifier is then applied to both training and testing images, giving a classification score for each image. This single score is then used as the frequency domain feature, which is concatenated with the 2 color features to form a 3 dimensional image feature. The photographic/graphic image classifier is then trained using these 3 features.

This approach is tested on 462 images collected from 25 news front pages [7], which includes 232 photographic images and 230 graphic images. The procedure which produced this dataset is explained in detail in Section 5. The data set is divided into five roughly equal parts, each containing roughly equal numbers of graphic and photographic images. A Support Vector Machine (SVM) classifier [16, 5] was then trained on each four of the five parts and tested on

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE

the remaining part. The process is rotated and the combined five part results were then pooled together to arrive at the overall accuracy of the classification algorithm. For SVM classifier training and testing, we used the SVM^{light} system implemented by Thorsten Joachims [9] and tested both the linear kernel and the Radial Basis Function (RBF) kernels.

Our experiments indicated that the RBF kernels performs better than the linear kernel for both the intermediate frequency domain classifier and the final image classifier. The final graphic/photographic image classifier achieved an accuracy of 92.5% on the dataset described above.

4. Identifying Story and Preview Images

As described earlier, our strategy is to first build a main classifier to separate the SPA images from the others and then to use a secondary classifier to remove the A images. This section describes the main classifier which uses both image and text features. For the image feature, the classifier uses the output of the SVM classifier used to classify photographic and graphic images (as described in Section 3).

A description of the text features follows.

4.1. Text Features

Images on the web are almost always accompanied by text and such text often contains useful information about the nature and content of the images. Much research has been carried out in the past on using the associated text for image searching and indexing on the web [6]. For that particular task, it was found that the most relevant text fields are: image file names, image captions and alternate text (defined by the `<alt>` tag in HTML). The functional classification of images is a different problem requiring a different set of features as well as techniques. Since in this case the goal is not to search for a particular image, but rather to classify any given image into one of several broad functional categories, the text fields mentioned above are too specific. Instead, as can be seen from the example images given in Figure 1, the surrounding text of an image (text found in the immediate neighborhood of the image) plays a much more important role in identifying its functionality.

The extraction of the surrounding text of an image is a non-trivial task by itself. Ideally, one should use spatial proximity to judge what text is near a particular image. Unfortunately, while tools for querying spatial information of nodes in an HTML DOM tree are being developed, they are not yet widely available. To get around this problem we used an approximation in our experiments. For each image, we extracted text nodes in the neighborhood of the image node in the DOM tree, within a maximum of 2 levels. A maximum of 20 words each are extracted for "before text" (from text nodes to the left of the image node) and "after text" (from text nodes to the right of the image node). Structural features such as node boundary and whether each node is a hyperlink are preserved during extraction.

For each image, the classifier analyzes the set of extracted text nodes from the neighborhood of the image. The following feature values are computed over the set of text nodes:

Hyperlink Count. This is simply a count of the number of nodes that are hyperlinks. Images in class H (Heading) are likely to have larger values for this feature as compared to images in either the S, P, or A classes.

Number Count. This is a count of the number of all numeric words in the nodes. Images in class C are likely to have larger values for this feature.

Caps Count. This is a count of the number of capitalized words present in the text nodes. If the first word of a node is capitalized, then it is not included in the count as we assume that it is the beginning of a sentence. Images in the SPA superclass are likely to have higher values for this feature as their contexts usually contain proper names.

Non-dictionary Word Count. This feature computes the number of words in the text nodes that do not belong to a dictionary. It is complementary to Caps count feature since most proper names are not found in dictionaries. The dictionary used is WordNet [13], an on-line lexical database developed at Princeton University.

Maximum Words Count. This feature computes the maximum number of words in any of the text nodes. Since SPA superclass images are likely to be accompanied by descriptions, the value of this feature for the superclass will likely be high.

4.2. Combining Image and Text Features

The image and text features were combined and a Support Vector Machine classifier using SVM^{light} [9] was trained and tested on the same training and testing sets used in Section 3.3 and described in Section 5. The best results were obtained using the linear kernel of SVM^{light}. The final results are given in Section 5.

Once the SPA superclass is identified, we used a secondary rule-based classifier to separate out the host images (A). The classifier consisted of the following two rules:

1. For each text node corresponding to an image, identify the proper names[1] in the node (if any) and then compute the percentage of the words in the node that belong to the person proper names (for example, in "Larry King Live," 66.67% of the words belong to the name "Larry King"). Once this computation is performed for all text nodes corresponding to an image,

[1] We used BBN's IdentiFinder[3] to identify and classify the person proper names.

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE

take the maximum value. If the maximum value is greater than 50%, then proceed to the second rule. Otherwise, the image is not a host image.

2. For the text node which contains the maximum percentage value (determined in the first rule), check if the node is also a hyperlink. If so, then identify the current image as a host; otherwise, it is not.

The choice of the rules was made with emphasis on the precision of identifying the host class in mind. We wanted to reduce the number of false positives for the host class as each such instance directly reduces the recall of the stories and previews class.

5. Experiments

The data collected from the 25 web sites consists of 899 images [7]. To increase the sizes of the training and testing sets, we collected a second set of front pages from the same sites but a different date. The new set consists of 960 images. The resulting set of 1859 images was subjected to the simple size-based screening test (as described in Section 3). After the screening, the resulting set consisted of 462 images.

The set of 462 images was then divided into 5 roughly equal parts containing an approximately equal number of graphic and photo images. Four of these parts are used for training while one is used for testing. The five-fold validation method was employed. In other words, the experiments were run five times where, in each run, one of the five parts was designated as a test set with the remaining four acting as training sets.

The precision and recall numbers achieved by the SVM classifier for the SPA super class are 90.5% and 95.4% respectively. After the rule based host image identification and removal, the final precision and recall numbers for the story and preview images are 82.6% precision and 95.3% recall.

6. Summary

As the popularity of the World Wide Web soars, it is increasingly being used as a publication medium that has the ability to instantaneously reach millions of people globally. As a result, web pages now contain an increasing number of images that serve different purposes. In this paper we described an image categorization system that attempts to identify the most important images: story and preview images. The system uses both image and text features and a hierarchical classification algorithm to achieve precision and recall numbers of 82.6% and 95.3% respectively.

References

[1] A. Antonacopoulos and D. Karatzas. An anthropocentric approach to text extraction from www images. In *Proc. DAS2000*, pages 515–526, Rio de Janeiro, Brazil, December 2000.

[2] T. Bickmore, A. Girgensohn, and J. Sullivan. Web page filtering and re-authoring for mobile users. *The Computer Journal*, 42(6):334–346, 1999.

[3] D. Bikel, R. Schwartz, and R. Weischedel. An Algorithm that Learns What's in a Name. *Machine Learning*, 34:1–3, 1999.

[4] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Seeing the whole in parts: text summarization for web browsing on handheld devices. In *Proc. WWW2001*, Hong Kong, China, May 2001.

[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–296, 1995.

[6] C. Frankel, M. Swain, and V. Athitsos. Webseer: an image search engine for the world wide web. *University of Chicago Technical Report TR96-14*, 1996.

[7] J. Hu and A. Bagga. Categorizing images in web documents. In *Proc. SPIE Conference on Document Recognition and Retrieval X*, Santa Clara, US, January 2003.

[8] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[9] T. Joachims. Making large-scale svm learning practical. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

[10] I. Keslassy, I. Keslassy, D. Wang, and B. Girod. Classification of compound images based on transform coefficient likelihood. In *Proc. of ICIP'2001*, Thessaloniki, Greece, October 2001.

[11] J. Li and R. Gray. Text and picture segmentation by the distribution analysis of wavelet coefficients. In *Proc. ICIP'98*, pages 566–570, Chicago, October 1998.

[12] D. Lopresti and J. Zhou. Locating and recognizing text. *Information Retrieval*, 2:177–206, 2000.

[13] G. A. Miller. Five Papers on WordNet. Technical Report 43, Cognitive Science Laboratory, Princeton University, July 1993.

[14] G. Penn, J. Hu, H. Luo, and R. McDonald. Flexible web document analysis for delivery to narrow-bandwidth devices. In *Proc. ICDAR01*, pages 1074–1078, Seattle, WA, USA, September 2001.

[15] K. Perlmutter, N. Chaddha, J. Buckheit, R. Gray, and R. Olshen. Text segmentation in mixed-mode images using classificaiton trees and transform tree-structured vector quantization. In *Proc. ICASSP'96*, volume 4, pages 2231–2234, 1996.

[16] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[17] J. Yang, Q. Li, and Y. Zhuang. Octopus: aggressive search of multi-modality data using multifaceted knowledge base. In *Proc. WWW2002*, Hawaii, May 2002.

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE