Semantic classification of business images

Berna Erol and Jonathan J. Hull

Ricoh California Research Center
2882 Sand Hill Rd. Suite 115, Menlo Park, California, USA
{berna_erol, hull}@rii.ricoh.com

ABSTRACT

Digital cameras are becoming increasingly common for capturing information in business settings. In this paper, we describe a novel method for classifying images into the following semantic classes: document, whiteboard, business card, slide, and regular images. Our method is based on combining low-level image features, such as text color, layout, and handwriting features with high-level OCR output analysis. Several Support Vector Machine Classifiers are combined for multi-class classification of input images. The system yields 95% accuracy in classification.

Keywords: image classification, image retrieval, photo organizer

1. INTRODUCTION

As digital cameras, especially those integrated in cellular phones, become more ubiquitous, they play an important role in capture and sharing of visual information in work places. For example, an office worker may use a digital camera to capture the contents of a whiteboard, a document, information on a set of slides, a business card, or a scene with other people. Workplace studies have shown that people would use a digital camera in the office to capture these kinds of images if a camera were available [1]. Accordingly, digital cameras present a unique opportunity to increase workplace productivity. However, while many workers are apt to find uses for captured images, fundamental limitations remain. Office workers generally have little time for classifying, organizing, and applying the correct post-processing to the images they capture.

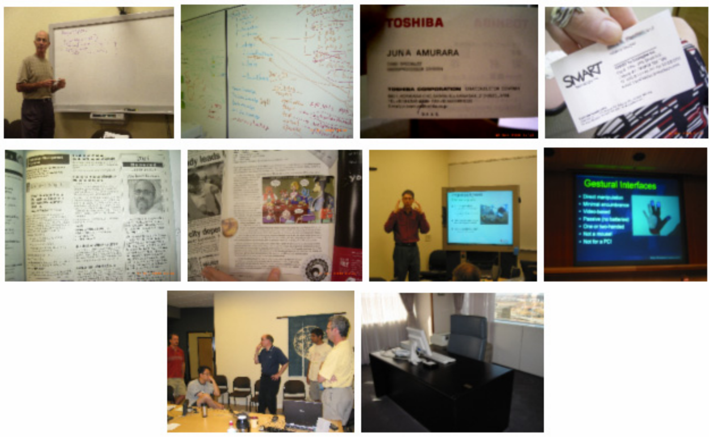


Figure 1. Examples of business images. In the raster scan order (2 in each group): whiteboard, business card, document, slide, and regular images.

In this paper we describe a novel method to automatically classify images captured with a digital camera into the following five groups: document images, whiteboard images, business card images, slide images, and regular images. Some examples of such images are shown in Figure 1. We employ image features that are based on color content, text layout, and handwriting analysis. For example, features that discriminate machine print and handwriting are found to be useful for separating whiteboard images from document images and layout analysis is successfully employed to differentiate slide images from business card images. We also use OCR output to compute some text features based on domain knowledge. For classification, we employ Support Vector Machines, where 10 SVM binary classifiers are trained for each semantic class pair to obtain a multi-class classifier. Experiments performed on a business image database show that the system achieves 95% accuracy on average in classifying digital camera images into the correct semantic class.

The classification results can be used for automatically organizing images in a photo browser application. It can also be very useful for automatically applying the right post processing or performing unconscious retrieval. For example, if the captured image is a whiteboard image, then a whiteboard clean-up algorithm can be automatically applied to make the image ready for viewing and printing. If the image is a slide, then it can be used for automatically retrieving a presentation recording with a method described in [2]. If the image is a document image, the OCR'ed text can be used to automatically retrieve the original version of the document in a local database or on the Internet.

2. RELATED WORK

Many researchers have addressed semantic classification of images based on low-level image features. Most of the work in this area [3][4][5] concentrates on semantic classes such as indoor/outdoor, people/non-people, building, city, landscape, sunset, forest, etc. These are more general, consumer-level, semantic classes compared to the business-use oriented semantic classes that we propose here.

Another relevant body of work concerns the separation of computer-generated graphic images, such as presentation slides, from photographic images [6][7][8]. Much of the prior research in this area utilizes low-level image features such as sharpness of edges or uniformity of colored regions to separate computer-generated content from photographic images. Many of these image features are not useful for our purpose, because when computer graphics is captured with a digital camera and compressed with JPEG, they become photo-like and most of these image features lose their discriminatory power. For example sharp edges in computer graphics become soft edges and uniform colors become non-uniform. Therefore, such image features cannot be employed in our work.

Layout analysis of document and business card images has been studied extensively in the literature [9][10]. These works mostly focus on analyzing the document and business card images captured by a scanner. Here the images we work with are captured by a digital camera not by a scanner. Therefore, it is likely that only part of a document is captured, the captured document is lower in resolution, and significantly more prone to distortions compared to scanned images, making robust layout analysis more difficult. Also, the prior art does not classify captured images into the semantic classes we propose here.

The novelty of our work includes the definition of a taxonomy of images captured by a digital camera in a business environment. We also propose an algorithm that can correctly classify an arbitrary digital camera image using a new combination of low-level image features and high-level analysis of text content based on domain knowledge of the business images.

3. SEMANTIC CLASSIFIER

The image features that we employ are based on analyzing structure and content of text regions, as well as color content analysis, which are explained in detail in the following sections.

3.1 Feature Extraction

The first step in feature extraction is identification of text regions and skew correction. Commercial OCR packages usually fail to identify text regions with handwriting; therefore we developed a text detection and skew correction algorithm as follows. We identify text-like regions by first re-sampling the image to 960x720, finding strong horizontal edges with the Canny edge detector, smearing the edges with a 64x2 smearing filter, thresholding, morphological closing, and then connected component analysis. Text-like regions without a certain height and width ratio are filtered out. Lines are fitted to each text region and a histogram of the tangent of the lines is computed. The histogram is then filtered with a 5-tab smoothing filter and the histogram bin with the maximum value is selected to be the skew angle. The text regions are rotated and binarized by adaptive thresholding.

3.1.1 Handwriting Features

Some features extracted from connected component height histograms is useful for separating machine print from hand writing, which is in turn useful for differentiating whiteboard images from document images. The histograms are computed only for connected components (which are individual letters or markings) in the regions that are identified as being text regions. Figure 2.a and Figure 2.b show connected component height histograms for document images and whiteboard images, respectively. Based on the component height histograms, we compute 2nd and 3rd order *X* moments, μx_2, and μx_3, respectively, and the spread of histogram bins,

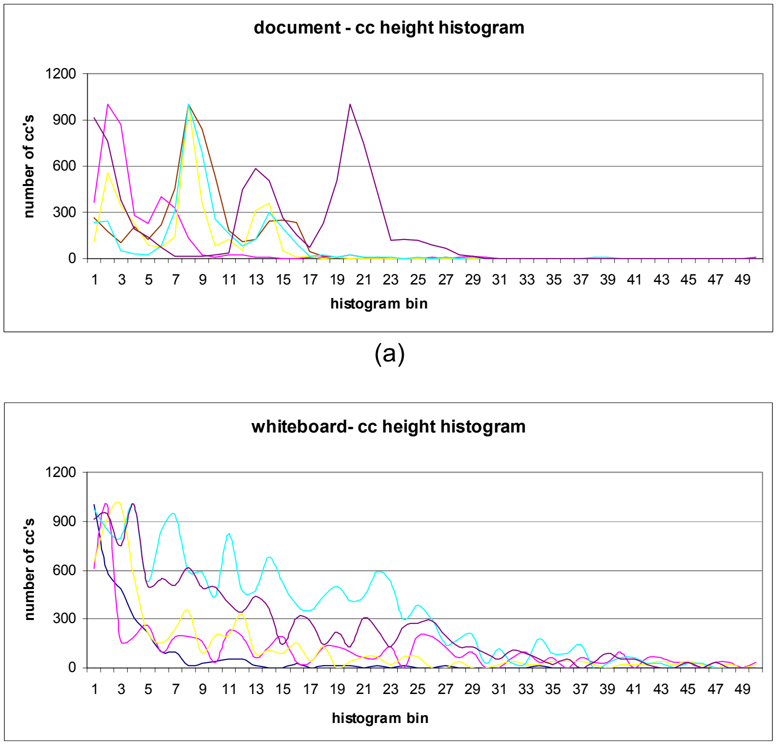
$$iS = \sum_{i} (i - \bar{i})\, x_i \ ,$$

where x_i is histogram value at bin *i*.

Because letters are connected in handwriting more so than that of machine print, the average height-to-width ratio of connected components in a handwriting text region is typically much smaller than that of the machine print. Based on this, we also compute

$$c_{av} = \frac{1}{N} \sum_{i=1}^{N} \frac{h_i}{w_i} \,,$$

where N is the number of connected components (individual binarized letters), h_i is the height, and w_i is the width of component *i*. Whiteboard images typically have a low connected component height-to-width ratio, c_{av}, in text regions, where as document, slide, and business card images have a high ratio.



(a)

(b)

Figure 2. Connected component height histograms for (a) document and (b) whiteboard images.

3.1.2 Text-based Features

A text confidence score, *tc*, is computed for an input image as follows:

$$tc = \sum_{m=1}^{M} \frac{T_w^m}{T_h^m},$$

where M is the number of text regions in the image, T_w^m is the width and T_h^m is the height of text region *m,* respectively. This is a good feature for discriminating document, slide, whiteboard, and business card images from regular images.

Text is extracted from the text regions using a commercial OCR package. The following text features are computed based on the OCR output:

> N_{words}: Number of words that have more than 3 characters. This feature is useful for separating document, slide, and business card images from regular images.
>
> $R_{capital}$: The ratio of words starting with capital letters to the number of words. This feature is useful for separating business card images from the document and whiteboard images.
>
> $R_{numerical}$: The ratio of words starting with numerical characters to the number of words. Business card images contain many words with numerical characters, such as phone number, fax number, zip codes, etc. So, this feature is again good for identifying business card images.
>
> R_{key}: The ratio of number of *business card keywords* to the number of words in an image. Business card keywords include the following: "phone, tel, fax, email, e-mail, web, www".
>
> B_{ratio}: The ratio of number of text lines starting with a bullet point to the total number of text lines. Although bullet points are not explicitly recognized during the OCR process, most of the time they are identified as one of the following ASCII characters: { * , - , . , o }. Therefore, if a line starts with one of these characters, it is counted as a bulleted line. If B_{ratio} has a high value, then the image is likely to be a slide image.

3.1.3 Layout features

It is particularly difficult to differentiate business card images from slide images with a white-background, since they contain similar amounts of machine-printed text and the same color background. The layout of slides or business cards may vary greatly. Nevertheless, in practice, most slides contain text lines that are aligned to the left side of the image and most business cards contain text lines that are either centered or have a two-column layout where most of the text resides at the right side of the image.

Because we use a wide smearing filter when identifying text boxes, spatially close words fall in the same text box. Therefore, projections of the number of text boxes give good information about the image text layout. We first segment out the image region that contains text boxes, divide the image into ten horizontal and ten vertical segments, compute the horizontal projection vector $\vec{P}_H = \{p_1, p_2, ..., p_{10}\}$, where p_n is the number of text boxes in the horizontal segment *i*. The vertical projection vector, $\vec{P}_V$, is also computed in a similar way. The maximum number of columns and lines are represented with p_x^h and p_x^v. Moreover, p_*^h and p_*^v, which are the median of the horizontal and vertical projection values, respectively, represent the median number of text columns and lines in an image. Median values are computed by ignoring p_n that are equal to zero. These features are useful for separating slide images from business card images and business card images from documents based on their text layout.

3.1.4 Color features

Some whiteboard images may contain very few foreground strokes. In those cases, it is difficult to differentiate them from regular images purely based on text region features. On the other hand, whiteboard images generally contain a large, relatively uniform background that is in a light color. We compute 2 color features that highlight these properties. First an 8-bin luminance histogram of the image is computed. The index of the dominant luminance pairs, I_d, is computed by

$$I_d = \underset{h_i \in HIST}{\arg\max} \{h_{i-1} + h_i\} ,$$

where h_i, is the value of the ith histogram bin. This is used instead of the dominant color as a feature in order to accurately represent the cases where the dominant value is divided somewhat equally between 2 neighboring bins. Then the percentage of the dominant luminance value pair, P_d, is computed by

$$P_d = \frac{\max\limits_{h_i \in HIST} \{h_{i-1} + h_i\}}{\sum\limits_i h_i}.$$

Both of these features are used for differentiating whiteboard images containing little text from regular images.

Text and background colors are useful for separating slide images from business card and document images − though exceptions exist where a slide image may have a white background with a black foreground. The average colors for text and background in text boxes are computed, and hue, $H_{f,}$, saturation, $S_{f,}$, and value, $V_{f,}$, of text (foreground), and those of $H_{b,}$, $S_{b,}$, and V_b, of background color, are used as features.

3.2 Classification with Support Vector Machines

All the image features are normalized to [0,1]. We employ Support Vector Machine based classification [11], as it was shown to offer excellent classification performance in many visual pattern recognition problems [12][13]. SVM is a binary classifier. In order to achieve a multi-class classifier, we train an SVM classifier for each semantic class pair. For example, business_card images vs. document images, regular images vs. document images, and so on. This results in 10 SVM classifiers. An input image is assigned to a semantic class if 4 of these SVM classifiers agree on the same decision.

The kernel function in SVM maps the feature vector to a higher dimensional space, where an optimal separating hyperplane can be found that maximizes the margin between two classes [11]. The kernel function we employ for training is a radial basis function because it offered the best performance compared to the polynomial and linear kernel functions in our experiments.

4. EXPERIMENTAL RESULTS

We evaluated the performance of our classifier on a database of 1025 business images. Our database contained 151 document, 132 business card, 278 slide, 80 whiteboard, and 384 regular images captured by our lab members either in the office or during conference or tradeshow trips. Regular images include both indoor images and outdoor/scenery images. Most of these images were captured with 2- to 4-Megapixel digital cameras.

Business Image Class	Number of Images in the Database	Precision	Recall
Document	151	%100	%94
Business Card	132	%87	%92
Slide	278	%100	%96
Whiteboard	80	%80	%95
Regular	384	%97	%99

Table 1. Business image classification results.

Ten SVM classifiers were trained with 30 images from each class using the SVM^{light} software [14] and queries were performed using all images in the database. The classification results are presented in Table 1. Here, *precision* is the ratio of correctly labeled images in a given category to all labeled images as being in that category, and recall is the ratio of correctly labeled images in a given category to all images in that category. As can be seen from the table, regular, slide, and whiteboard images were identified with 95% or higher accuracy, 94% of document images, and 92% of business card images were correctly labeled. Figure 3 shows a subset of images that were correctly classified.

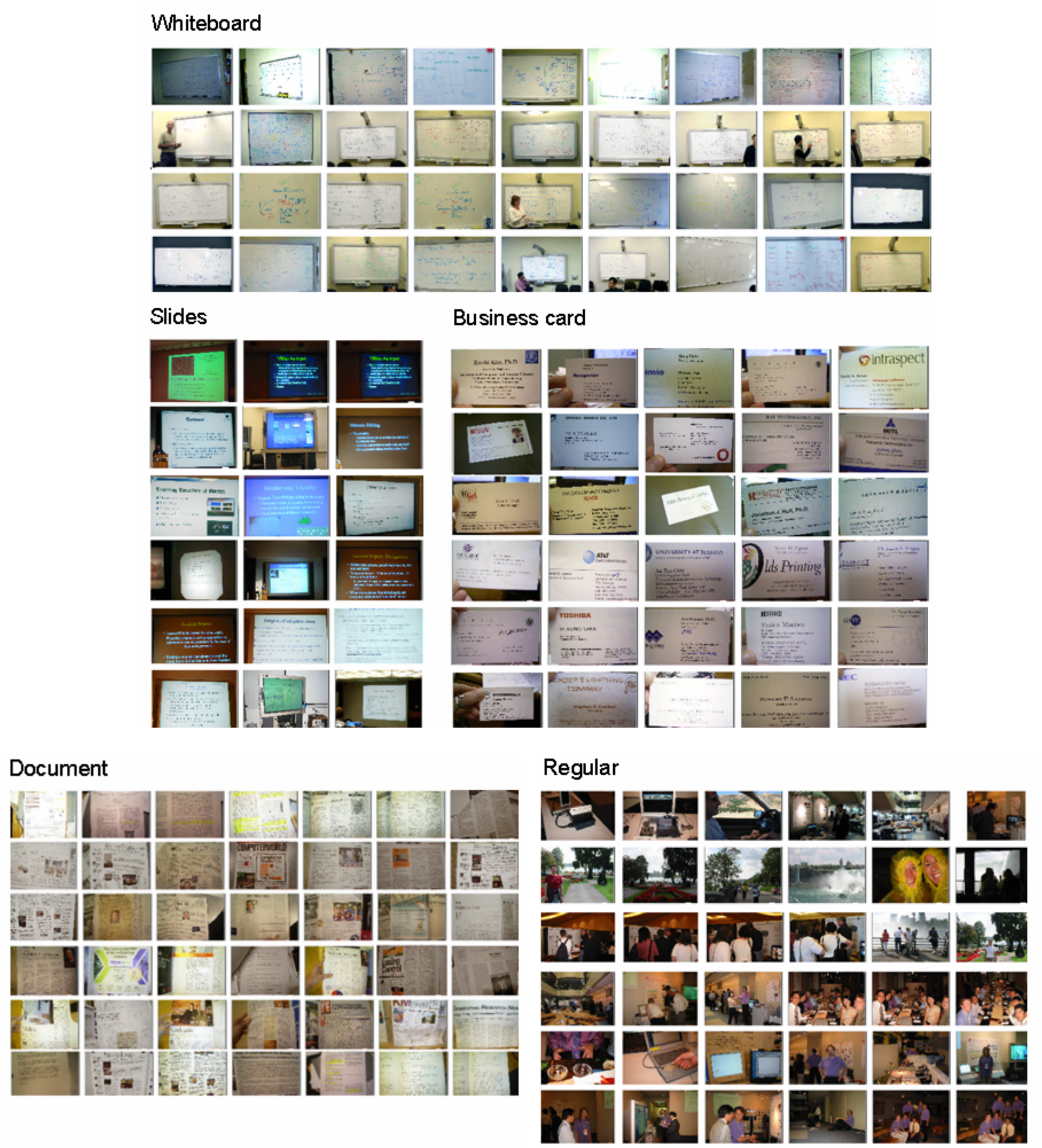


Figure 3. Examples of business images classified correctly into one of 5 image classes by our algorithm.

Some examples of mis-classification are presented in Figure 4. In classifying document images, our observation was that the black and white documents are very accurately classified and misclassification happened mostly on images of colored magazine pages that contain large fonts and many photos. An example of this is shown in Figure 4.b. Moreover, some of the regular images taken at conferences that have posters in the background, such as Figure 4.c, are mis-classified as whiteboard images, which yield to a lower *precision* score for the whiteboard category. In the future, a separate class may be considered for images with posters. Nevertheless, the overall correct classification rate of our business image classifier is above 95%.

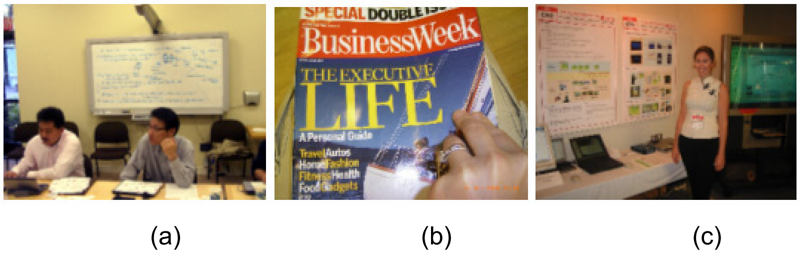


Figure 4. Examples of misclassified images: (a) Whiteboard image misclassified as a regular image (b) document image misclassified as a regular image (c) regular image misclassified as a whiteboard image.

5. CONCLUSIONS AND OUTLOOK

In this paper we presented a novel method for classifying digital camera images captured in a business environment that yields a good performance.

Our method is based purely on image analysis. It is possible to use other metadata about an image, such as time and location information of the picture, and the user's calendar to improve the classification results. For example, if a picture is taken during the time the user is scheduled to be attending a conference session, the picture is likely to be a slide or a regular image.

Currently we are integrating the business image classifier in an image browser that allows automatic post-processing, unconscious linking, and blogging of office events. Another interesting direction is using our algorithm for classifying images on the web.

6. REFERENCES

[1] B. Brown, S. Abigail, and K.P. O'Hara, "A Diary Study of Information Capture in Working Life." Proc. of ACM CHI Conference, p.438-445, 2000.

[2] B. Erol, J.J. Hull, and D.S. Lee, "Linking Multimedia Presentations with their Symbolic Source Documents: Algorithm and Applications," ACM Multimedia Conference, pp. 498-507, 2003.

[3] R. Zhao and W. I. Grosky, Bridging the Semantic Gap in Image Retrieval, Distributed Multimedia Databases: Techniques and Applications, T. K. Shih (Ed.), Idea Group Publishing, Hershey, Pennsylvania, pp. 14-36, 2001.

[4] J. Luo, and A. Savakis, "Indoor vs Outdoor Classification of Consumer Photographs using Low-level and Semantic Features," Proc. of ICIP, pp.745-748, 2001.

[5] A. Vailaya, A. K. Jain, and H.-J. Zhang, "On Image Classification: City Images vs. Landscapes," Pattern Recognition Journal, vol. 31, pp 1921-1936, December, 1998.

[6] J. Z. Wang, G. Li, and G. Wiederhold, "SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries," In IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 23, pages 947-963, 2001.

[7] S. Prabhakar, H. Cheng, J.C. Handley, Z. Fan Y.W. Lin, "Picture-graphics Color Image Classification," Proc. of ICIP, pp. 785-788, 2002.

[8] A. Hartmann and R. Lienhart,"Automatic Classification of Images on the Web," In Proc of SPIE Storage and Retrieval for Media Databases, pp. 31-40, 2002.

[9] M. Aiello, C. Monz, L. Todoran, and M.Worring, "Document Understanding for a Broad Class of Documents," Proc. of IJDAR, pp. 1-16 , 2002.

[10] T. Watanabe, X. Huang, "Automatic Acquisition of Layout Knowledge for Understanding Business Cards," Proc. of ICDAR, pp. 216-220, 1997.

[11] N. Cristianini and J. Shawe-Taylor, "Support Vector Machines and other Kernel-based Learning Methods," Cambridge University Press, ISBN 0-521-78019-5, 2000.

[12] P. Wang and Q. Ji, "Multi-View Face Detection under Complex Scene based on Combined SVMs," Proc. of ICPR, pp. 179-182, 2004.

[13] H. Miyao and M. Maruyama, "An Online Handwritten Music Score Recognition System," Proc. of ICPR, pp. 461-464, 2004.

[14] SVM Light software, http://www.cs.cornell.edu/People/tj/svm_light/