Influence of Fusion Strategies on Feature-based Identification of Low-resolution Documents

Ardhendu Behera
Department of Informatics
University of Fribourg, Chemin du Musee 3, CH-1700, Switzerland
+41 26 429 6678
Ardhendu.Behera@unifr.ch

Denis Lalanne
Department of Informatics
University of Fribourg, Chemin du Musee 3, CH-1700, Switzerland
+41 26 429 6596
Denis.Lalanne@unifr.ch

Rolf Ingold
Department of Informatics
University of Fribourg, Chemin du Musee 3, CH-1700, Switzerland
+41 26 300 8466
Rolf.Ingold@unifr.ch

ABSTRACT
The paper describes a method by which one could use the documents captured from low-resolution handheld devices to retrieve the originals of those documents from a document store. The method considers conjunctively two complementary feature sets. First, the geometrical distribution of the color in the document's 2D image plane is preferred. Secondly, the shallow layout features is considered due to the poor resolution of the captured documents. We propose in this article to fuse those two complementary feature sets in order to improve document identification performance. Finally, in order to test the influence of merging strategies on document identification performance, a synergic method is proposed and evaluated relative to a similar method in which feature sets are simply considered sequentially.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models, search process*.

H.3.4 [**Information Storage and Retrieval**]: System and Software – *performance evaluation*.

General Terms
Algorithms, Performance, Experimentation.

Keywords
Document retrieval, document signature, shallow layout features, geometrical color distribution.

1. INTRODUCTION
The need for document image retrieval system has become increasingly apparent, as more and more heterogeneous documents are archived in digital libraries, office automation, etc. These systems operate on document databases to recover relevant documents in response to a query and the matching of document image is used as the kernel technology. The concept of relevance is most likely to be associated with the contents. The content could be expressed using two types of features sets: character or local features and layout or global features. The extraction of local features such as textures, shapes, etc. is mostly dependent on the distortion, noise and the quality of the captured documents. The global features helps to retrieve the set of similar documents but it under performs for finding the exact document as compared to the local features. However, in many practical situations local features are correlated with the global features. Therefore, a successful document image retrieval algorithm should combine both the local and global feature to achieve more outstanding performance.

The classical approach for retrieving document image is based on character features by matching of textual content of the document using OCR technology [6]. The drawback of such approaches is that it is time-consuming and requires different OCR systems to deal with different languages. The other approach is based on the image information directly such as texture [3], layout [5], etc. The input of most of the above-mentioned systems is either a scanned document of 300 dpi or higher and of uniform background. Such systems are difficult to use if the perceived document image is captured from a low-resolution handheld device and compressed with quality-losing format such as JPEG.

In this paper, we propose an image-based retrieval method for documents captured from low-resolution handheld devices targeting documents with a limited textual content and a variable layout such as projected slides during presentations. Such captured documents could be queried to the system to retrieve the original slides, which are linked with audio/video recordings of conferences, seminars, meetings, lectures, etc. Slide identification has already been tackled using OCR and text layout [4]. However, such systems need at least some textual content and a uniform background which is not always the case in a slide. On the other hand, color is a low-level feature that has been rarely incorporated along with the layout features for document identification. The proposed method uses two types of feature to retrieve the original electronic documents. One is the geometrical distribution of colors in the document image plane and is considered as a set of global feature. The other one is the shallow layout feature and is considered as a set of local features.

2. FEATURE EXTRACTION
The projected slides captured from a handheld device not only contain the projected part but also the surrounding background. It is thus necessary to remove the background and to rectify the skewing of the remaining document image [1]. Each rectified captured document as well as each original electronic document is

DocEng'05, November 2–4, 2005, Bristol, United Kingdom.

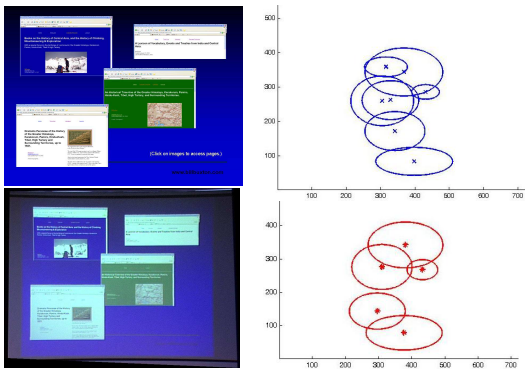

Figure 1. The geometrical features are represented with the ellipses a) original documents (top row) and b) captured document (bottom row).

processed for the extraction of the features set, which facilitates the identification of the captured low-resolution documents by matching the corresponding features set between the captured and original electronic documents. The global feature set is first computed and followed by its local feature set.

2.1 Geometrical Color Distribution Features

This feature set is extracted by considering the geometrical distribution of the similar pixels in the document image plane. Often, the values of pixels in the captured images and corresponding pixels in the original image are not the same due to the presence of color cast, which is the predominant superimposed color. This is due to changes in the lighting environment, surface properties of the target object and even the characteristics of the capture devices. However, the geometrical distributions of the pixels in the image plane remain preserved. Therefore, we believe that the geometrical distribution of similar pixels is more powerful than the spatial distributions in any one of the color space in the case of image captured from the low-resolution handheld devices. The distribution is computed after grouping of the pixels of similar color and is done by using the K-Means clustering. The value of K is derived from the number of predominant peaks in the reduced RGB color histograms. The geometrical distributions of pixels in each of the cluster are computed and represented with an ellipse with *five* parameters. These are *center*, *variance*, and *density* of the cluster and represent the geometrical distribution of each cluster in the 2-D image plane [1]. In Figure 1, the number of ellipse corresponds to the number of clusters, the *center* and *axes* of each ellipse represents the respective geometrical center and variance in the 2-D image plane. It is observed that the number of clusters in the captured image is often less than that of the original and is due to the presence of color cast.

2.2 Shallow Layout Features

This feature set is mainly based on the layout information of the document. The resolution of the captured documents is very low (450 × 560 pixels, < 75 *dpi*) for the extraction of the complete layout structure. For this reason, the shallow layout feature is extracted, which is based on the layout structure and close to the perception of human vision. The extraction process is a top-down approach i.e. first of all, the global information of the document is considered and then partition the document into blocks before classifying them into texts, images, solid bars, bars with text.

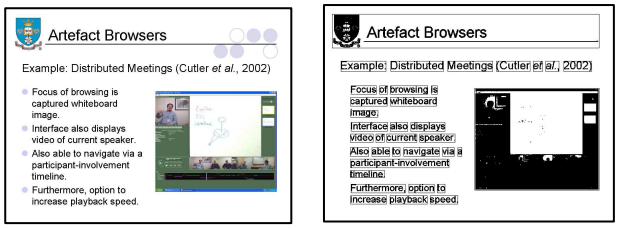

Figure 2. Bounding box containing layout features of the original slide image

Moreover, the text blocks are separated with individual text lines and further processed to the word level. Other features like bullet and vertical text lines are also extracted. Each feature in the features set is structured according to its priority and has a label tag of one of the following: *horizontal text line*, *vertical text line*, *image*, *bullet* and *solid bar*. The geometrical information about each feature like, location, width, height and the bounding box density of each feature, which is stored in a symbolic file called *Visual Signature*. The detailed extraction procedures and structuring of the above-mentioned feature set has been explained by *Behera et al.* [2]. Figure 2 illustrates the bounding boxes of each of the feature, such as text lines, solid lines, words, graphics, bullets, etc. of an original document.

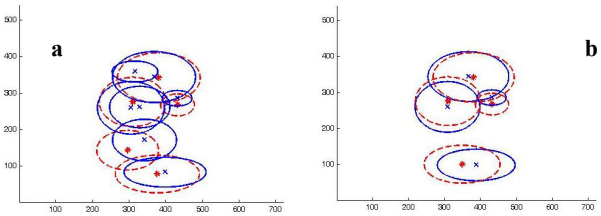

Figure 3. Merging of clusters during matching a) original clusters in the original (solid) and captured documents, b) merged clusters to bring the center closer.

3. FEATURE MATCHING STRATEGIES

For the matching of *geometrical color distribution features*, the algorithm takes into account the number of clusters and the properties of each cluster in both the original and queried documents. Often, the number of clusters in the captured and the original image is different and is due to presence of color cast in captured documents. The idea is to bring the number of clusters and their geometrical distribution of both the captured and original image as close as possible by merging and comparing. Two or more clusters are to be merged if and only if the following conditions are satisfied: a) the sum of the cluster densities of the merged clusters and b) the resulting geometrical centroids are close to the respective cluster density and geometrical centroid of the one to be compared with. Figure 3 is an example of such merging and comparing of clusters in the original and captured image of Figure 1. Before merging, there are 7 clusters in the original image (solid ellipse) and only 5 in the captured image (dotted ellipse, Figure 3a). The clusters are merged and compared using the above-mentioned method and the final number of clusters is brought down to 4 and is the closest one. Figure 3b shows the resulting clusters whose corresponding center and axes are much closer with comparison to Figure 3a. The clusters are compared in an ascending order of their densities. The similarity distance between each cluster's geometrical properties, in the original and in the captured image, is then computed [1].

In case of *layout features*, the feature score is computed at each of the feature level (text, image, bullets, bars, etc.) of the layout signature by comparing the number of elements and their geometrical properties. A predefined weight is assigned to each feature level according to their priority level. The weighted sums of the features' score are computed (f_l) and compared. The signature having the highest score is picked up and corresponds to the identified document [2].

In case of sequential matching, the matching procedure considers a single feature set; followed by another till the desired solution is reached [1]. In this scenario, some solutions are sometimes removed too early due to the fact that the first feature tested is weaker as compared to all other existing features. Therefore, a better strategy to overcome this drawback consists in fusing features sets. In this case, first the distributed color feature score (f_{dc}) is used for pre-filtering the solution space and then the final score is computed by summing up the scores of both features. Let $D = \{s_1, s_2, ..., s_n\}$ is the set of signatures of the documents in the repository and s_q is the signature of the queried image. The set $S = \{s_1, s_2, \ldots, s_m\}$, $m \leq n$ is considered from D if $f_{dc}(s_q, s_i) < T_c$, where i = 1…n. The fused score, $f_{c,j} = W_l f_l(s_q, s_j) + W_{dc} f_{dc}(s_q, s_j)$, $1 \leq j \leq m$ is computed and the signature having the highest $f_{c,j}$ is picked up as the required solution. The weight W_l and W_{dc} corresponding to the layout feature score and the distributed color feature score are assigned adaptively by considering the content of the document. If the text feature score of the layout feature set is high then $W_l > W_{dc}$ and if the clusters in the color signature are not merged during the matching then $W_l < W_{dc}$.

4. EVALUATION, RESULTS & ANALYSIS

The evaluation of the proposed method has been performed by querying 355 slides from 16 different slideshows captured using a DV camera (Sony, DCR-TRV27E, PAL, 1 mega pixels). There are a total of 2000 slides in the repository from 60 different slideshows. The identification rate, I = # *correct documents retrieved* / # *total documents queried*, is used for the evaluation.

In Table 1, the average identification rate for each slideshow is displayed for a) sequential matching of features, first layout and then distributed color features and b) the fused features matching. The procedure compares the performances for identifying the exact match and further to find the correct match among the top five and top ten solutions. The last row of the table represents the average values for all 16 slideshows. In case of sequential matching of the features sets, the performance is 89%, whereas it is 92% and 93% of the time within the top five and top ten, respectively. Often, we observed that several solutions have the same final matching score and in this case the system picks up the first one. This is due to the fact that some slides such as the title slide, the last slide (Thank You) and result slide with similarly structured figures and legends from different presentations result in similar feature sets. The performance of the matching using the above-mentioned fused feature sets surpasses (91%, 94% and 95%) over than that of the sequential matching often due to the poor quality of the low-resolution captured images, which results in the introducing error in the layout features (local), which is used for the final identification in the sequential matching. This is overcome by the matching using fused features as the distributed color feature which is considered as the global feature is comparatively less sensitive to the noisy low-resolution captured images. The fusion strategy is simply the linear combination of the geometrical distribution of color and layout features. In near future, we plan to improve the fusion strategy by using heuristics to combine the different features from the various feature sets.

5. CONCLUSION AND FUTURE WORK

The method proposed here, identifies documents captured from low-resolution handheld devices based on (a) the geometrical distribution of the color, and (b) the shallow layout structure of the documents. Finally, an evaluation of the proposed method has been presented. In the near future, we plan to develop a method in order to calibrate colors automatically and an identification method which considers only the color features for improvement in performance.

Table 1. Comparative identification rates

# Slides	Sequential features			Fused features		
	Best one	Top five	Top ten	Best one	Top five	Top ten
34	0.92	0.92	0.92	0.96	1.00	1.00
10	0.90	1.00	1.00	0.94	1.00	1.00
15	0.75	0.94	1.00	0.90	1.00	1.00
28	1.00	1.00	1.00	1.00	1.00	1.00
30	0.96	0.96	0.96	1.00	1.00	1.00
24	0.86	0.86	0.86	0.86	0.90	0.90
19	1.00	1.00	1.00	1.00	1.00	1.00
28	0.96	0.96	0.96	0.96	0.96	0.96
25	0.80	0.84	0.84	0.80	0.90	0.90
20	0.94	0.94	0.94	0.94	0.94	0.94
29	0.98	1.00	1.00	1.00	1.00	1.00
17	1.00	1.00	1.00	1.00	1.00	1.00
15	1.00	1.00	1.00	1.00	1.00	1.00
16	0.75	0.86	0.86	0.86	0.86	0.86
20	0.72	0.72	0.85	0.75	0.75	0.90
25	0.67	0.67	0.67	0.68	0.74	0.74
355	**0.89**	**0.92**	**0.93**	**0.91**	**0.94**	**0.95**

6. REFERENCES

[1] Behera A., Lalanne D., and Ingold R. Enhancement of Layout-based Identification of Low-resolution Documents using Geometrical Color Distribution, In *Proc. 8th ICDAR*, Seoul, Korea, Vol. 1, pp. 468-472, 2005.

[2] Behera A., Lalanne D., and Ingold R. Visual Signature based Identification of Low-resolution Document Images, *ACM Symposium on Document Engineering*, Milwaukee, USA, October 28-30, pp. 178-187, 2004.

[3] Cullen J., Hull J., and Hart P. Document image database retrieval and browsing using texture analysis, In *Proc. 4th ICDAR*, Ulm, Germany, pp. 718-721, 1997.

[4] Erol B., and Hull J. Linking Presentation Documents Using Image Analysis, In *Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 9-12, 2003.

[5] Herrmann P., and Schlageter G. Retrieval of document images using layout knowledge, In *Proc. 2nd ICDAR*, Tsukuba City, Japan, pp. 537-540, 1993.

[6] Marinai S., Marino E., and Soda G. Indexing and retrieval of words in old documents, In *Proc. of 7th ICDAR*, Edinburgh, Scotland, pp. 223-227, 2003.