Color and Layout-based Identification of Documents Captured from Handheld Devices

Ardhendu Behera, Denis Lalanne, and Rolf Ingold

***Abstract*—**This paper proposes a method, combining color and layout features, for identifying documents captured from low-resolution handheld devices. On one hand, the document image color density surface is estimated and represented with an equivalent ellipse and on the other hand, the document shallow layout structure is computed and hierarchically represented. Our identification method first uses the color information in the documents in order to focus the search space on documents having a similar color distribution, and finally selects the document having the most similar layout structure in the remaining of the search space.

***Keywords*—**Document color modeling, document visual signature, kernel density estimation, document identification.

I. Introduction

CAPTURING and identifying images of documents using low-resolution handheld devices, webcams or digital cameras have a variety of applications in academics, research and knowledge management. Most of the document identification system, capturing document images from such devices and identifying them, by either global image matching (binary image) or text string matching using OCR [1]. However, these methods are generally time consuming and inadequate for low-resolution images. We propose in this paper an identification method that benefits from both the color and layout features of documents, and that is robust not only for low-resolution images but also to color deformations due to the various handheld capture devices properties and to the varying capture lighting conditions.

The application currently targeted by our method is the identification of documents captured during meetings, presentations, lectures, etc. In such environments, documents play an important role and are either displayed on the screen (e.g. slides) or simply laid on the table of the conference room. In our smart meeting application [2], such documents are captured using handheld devices and identified by comparing them with their corresponding electronic documents (e.g. PDF, PowerPoint). After identification, the relevant portion of meeting/lecture/conference can then be retrieved by querying captured document images from the handheld devices on the multimedia repository. The current focus is on the identification of the captured projected slides.

A. Behera, D. Lalanne, and R. Ingold are with the Computer Science Department, University of Fribourg, CH 1700, Switzerland (phone: +41-26-429-6678; fax: +41-26-300-9731; e-mail: {ardhendu.behera, denis.lalanne, rolf.ingold}@unifr.ch)

II. Color based Retrieval

Since most of the slide images in a slideshow have a similar color, texture and shape, our slide identification system should consider not only the layout structure of the slide images but also the color feature.

Color, as well as texture and shape [3], are low-level visual features extensively used in many systems in order to retrieve images having similar content as the queried ones. Retrieval systems based on such visual features work efficiently when queried on similar images, but do not when the captured image is taken from a different angle or has a different scale [4]. Furthermore, such features are very dependent on illumination conditions, shading and compression and for this reason we believe that a distribution of features is a better visual representation i.e. more robust to all the cited effects, than an individual feature vector.

The color histogram method is commonly used for the color-based image retrieval. It describes the color distribution of an image in a specific color space. Often, the RGB space is considered for the color feature extraction. A standard way of generating the RGB color histogram of an image is to consider the m higher order bits of the Red, Green and Blue channels [5]. The histogram consists of 2^{3m} bins, which accumulate the number of pixels having similar color values. In our approach, the generation of the color histogram has been reduced to two-dimensional chromatic space r = R/I and g = G/I (2^{2m} bins), where I = R + G + B is the brightness, $0 \le$ R, G, B $\le 2^m$-1 and b = B/I could be represented as 1 - r - g. The chromatic values r, g from RGB or a, b from the *Lab* are invariant to the illumination geometry. Let us consider a color image P of size $n_1 \times n_2$. Then $P = \{r_{i,j}, g_{i,j}\}$ could be represented with the chromatic values, where $i = 1 \ldots n_1$ and $j = 1 \ldots n_2$. The reduced color histogram $h(r, g)$ in rg- space is obtained as:

$$r = \text{int}(Mr_{i,j}),\ g = \text{int}(Mg_{i,j}),\ M = 2^m - 1$$
$$h(r,g) = \frac{\#\ pixels\ fall\ in\ bin\ r,g}{n_1 \times n_2},\ 0 \le r, g \le M \quad (1)$$

Finally, the similarity between two images is very often measured by computing the similarity distance between the respective histograms [6]. In the histogram representation the drawback is that the shape of the histogram strongly depends on the number of pixels and of the method used for the image representation. If the image size is small then there are very few points available for the histogram, which thus gives back the erroneous results for the histogram-based comparison. To

overcome the above-mentioned problems, we propose in the following section a smooth nonparametric estimation of the color distribution, instead of a discrete histogram representation, based on the concept of nonparametric density estimation [7].

III. Color Density Estimation

Density estimation describes the process of obtaining the probability density function (*pdf*) $f(x)$ from an observed random quantity. In general, the density functions of the random samples are unknown. The simplest and oldest form of the density estimation is histogram. In this case, the sample space is first divided into a grid of width h. Then the density at the center of the grid is estimated by $f(x)$ = *#samples in one bin / h*. In such estimation, the drawbacks are 1) the offset dependence 2) the lack of differentiability 3) sensitive to the rotation of coordinate axis and 4) in higher dimensions it causes sparse occupancy.

The drawbacks above are overcome by the Kernel Density Estimation (KDE) procedures. However, most nonparametric methods require either all samples or extensive knowledge of the problem. In this technique, the underlying probability density function is estimated by placing a kernel function on every sample in the sample space and then summing up all the functions for each sample. Given a d-dimensional sample space $X = \{x_i\}$, where $i = 1 \ldots N$, the multivariate kernel density at any point x is estimated as:

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h_1 \ldots h_d}\kappa\left(\frac{x_{i_1}-x_1}{h_1},\ldots,\frac{x_{i_d}-x_d}{h_d}\right) \qquad (2)$$

Where κ is the d-dimensional kernel function, which determines the shape of the ''bumps'' placed around the data points in the sample space and $h_1 \ldots h_d$ is the bandwidths for each dimension. The d-dimensional kernel functions are commonly represented as the product of the one-dimensional kernel functions i.e. $\kappa(u_1, u_2, \ldots, u_d) = K(u_1)K(u_2) \ldots K(u_d)$. In our approach, the two-dimensional chromaticity rg-space is used with the same bandwidth in both dimensions ($h_1 = h_2 = h$, i.e. radial-symmetric kernel function). The resulting kernel density estimation in two-dimensional space is:

$$\hat{f}(x) = \frac{1}{Nh^2}\sum_{i=1}^{N}\left\{\prod_{j=1}^{2}K\left(\frac{x_{i_j}-x_j}{h}\right)\right\} \qquad (3)$$

The estimation of the kernel density depends on the kernel function and the bandwidth h. We consider the *Epanechnikov* kernel, which has been shown to be robust to outliers and optimum in the sense of having minimum *mean integrated square error* (MISE) in comparison with other kernels [8].

$$K(u) = \begin{cases}\frac{1}{2}c_d^{-1}(d+2)(1-u^Tu) & \text{if } u^Tu \prec 1\\ 0 & \text{otherwise}\end{cases} \qquad (4)$$

Where c_d is the volume of the unit d-dimensional sphere and u is the d-dimensional data point. Fig. 1 illustrates the *KDE* of a sample slide document.

Jones and Rehag [9] reported that 77% of the possible 24-bit RGB colors were never encountered on images collected from the web. Furthermore, we observed no perceptive degradation of the *KDE* for 7-bits compared to 8-bits per RGB channels (Fig. 1), which tends to prove that reducing the color space do not affect much the color density estimation. For this reason and since the color feature is not used in our method to identify the original matching slide but in order to identify the slideshows or groups of slides having similar background pattern and color, it is judged reasonable to consider for the *KDE* the 7 most significant bits (*msb*) of each of the RGB channels, which reduces the sample space to its ¼, and thus heavily speeds-up the computation time of the *KDE*.

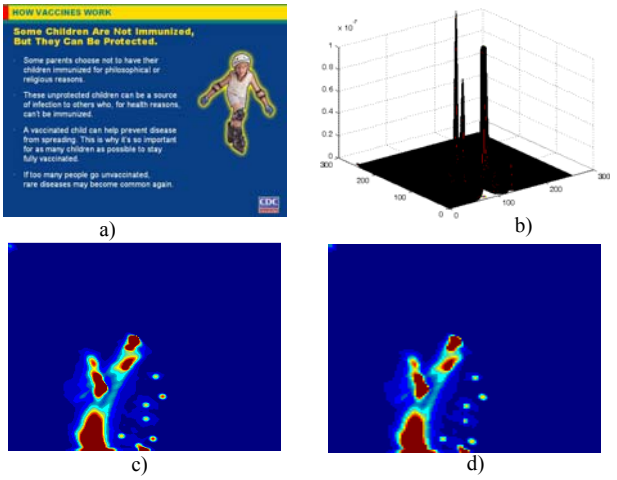

Fig. 1 a) Original image; b) *KDE* of the color distribution in the *rg* color space, c) its pseudo-color representation for the true color (24-bits) and d) reduced color (21-bits).

IV. Document's Signature

In our identification method, each of the captured and original electronic documents is represented with a signature containing mainly two parts: a) The documents' color distributions and b) the documents' shallow layout structure with the respective labeling.

A. Color Features Extraction

Once the *KDE* is done, the density distribution in the *rg*-plane of image colors is then analyzed by looking at its kernel density distribution $K_d(r, g)$. The mean (μ_r, μ_g) and variance (σ_r, σ_g) of the density surface in the *rg*-plane is computed as:

$$\mu_r = \int_r rK_d(r,g)dr \;,\; \mu_g = \int_g gK_d(r,g)dg$$
$$\sigma_r^2 = \int_r (r-\mu_r)^2 K_d(r,g)dr \;,\; \sigma_g^2 = \int_g (g-\mu_g)^2 K_d(r,g)dg \qquad (5)$$

Then the density distribution of each surface is associated to an *Equivalent Ellipse (EE)* with its center $C = (\mu_r, \mu_g)$, semi major axis $a = max\ (\sigma_r, \sigma_g)$, semi minor axis $b = min\ (\sigma_r, \sigma_g)$ and an orientation angle of θ. We could have considered the estimated density surface for matching rather than the equivalent ellipse but in this case, the position(s) of the peak(s) and valleys in the density surface would not have been the same in both the original (Fig. 1b) and captured images (Fig. 2b) due to the presence of superimposed dominant color

(color cast, Fig. 2a), which is usually due to the changes in lighting conditions and or capture devices. On the other hand, it is observed that most of the properties (eccentricity, orientation, etc.) of the *EE* of both the captured and original images are preserved and that only the *EE* location is shifted (Fig. 2d).

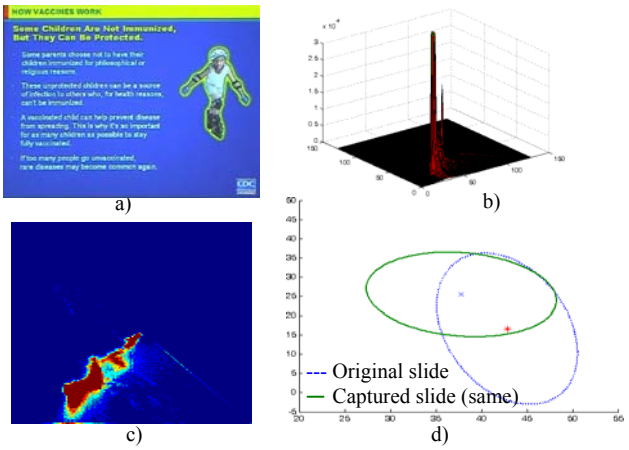

Fig. 2 a) Captured image of Fig. 1; b) its *KDE* of the color distribution, and c) pseudo-color representation for 21-bits in the *rg*-space, and d) equivalent ellipses of the density surfaces of both the original and the captured slide.

The feature vector for the color is finally $c_f = \{\mu_r, \mu_g, \sigma_r, \sigma_g, \theta, d\}$, where d is the density of the estimated kernel density distribution over the elliptical surface area. Fig. 3 shows the *EE* of 50 slides randomly picked up from 5 different slideshows (10 each) and it is possible to observe most of the slides within a slideshow have similar color since the properties of EE are close. In some cases only the centers of *EE* are close but the orientation and axes are dissimilar, which help to differentiate slides having different colors.

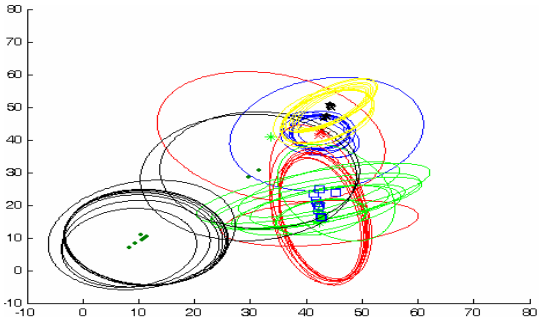

Fig. 3 Equivalent ellipse representation of the estimated color densities in the reduced *rg*-space of slides randomly picked from 5 different slideshows.

B. Layout Features Extraction

Document images are different from natural images and they contain mainly text, with few graphics and images. Due to the very low-resolution of images (the average size of the projected part is 450 × 560 and $dpi \leq 75$), captured with handheld devices, it is hard to extract the complete layout structure (logical or physical) of the documents. For this reason, we targeted a shallow representation, close to the perception of human vision, that we call a *visual signature*. This signature is hierarchically structured according to document's shallow physical layout structure with its respective labeling (text, graphics, solid bars, etc.). The motivation for slide documents with such signatures is that often the slides' content is limited and its layout varies a lot as compare to other type of documents (e.g. newspaper, articles, etc.). The detailed extraction procedure for the signature of each original electronic slide documents and captured slide image is explained in [1]. The signature of each slide contains one or more features from the set of features $\{f_1, f_2, \ldots, f_8\}$. These features are horizontal text line (f_1), image (f_2), bullet (f_3), horizontal solid line (f_4), vertical solid line (f_5), horizontal bar with text line (f_6), vertical text line (f_7) and vertical bar with text line (f_8). The final signature is organized according to the features priority containing the feature type, geometrical properties and pixel density. For the features with textual part, the number of words per text line is added to the feature's vector. For each feature f_i, it is represented with the vector $V = \{y, x, h, w, word, density\}$, where y and x are the minimum coordinates, height (h), width (w), number of words (*word*) and pixel density (*density*) of the feature's bounding box. Fig. 4 illustrates a document, where each bounding box represents a feature of the visual signature.

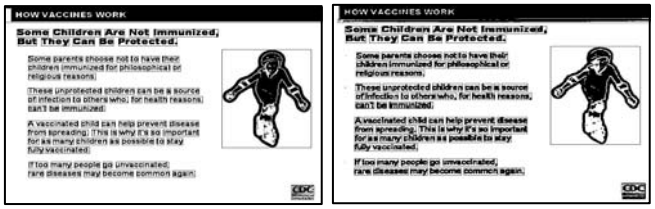

Fig. 4 Layout signatures i.e. bounding boxes for each visual features of the original slide (left) and its corresponding captured slide image (right).

V. Matching of Signatures

Our assumption is that most of the slides within a slideshow have similar background pattern and color, which means they share a similar distribution of the kernel density i.e. the properties of the equivalent ellipse in the *rg*-plane are similar. Once the queried image is identified from a particular slide show, further identification of the slide will be performed using the layout-based matching.

First, all the slide images in the repository are filtered out according to their color similarity, which reduces the size of the search space. The slides having the color feature (c_f) close (distance inferior to a threshold T_c) to the color feature of the queried image are considered. Let $S = \{s_1, s_2, \ldots, s_n\}$ be the set of signatures in the repository. After the color matching, a new set $S_c = \{s_1, s_2, \ldots, s_m\}$ is derived from S such that $m \leq n$.

Secondly, the layout-based feature matching is performed on the set S_c for the final detection of the queried slide images. The layout-based matching is basically matching of features between signatures by computing the features' score at each feature node (text, image, bars, bullets, etc). At each node some weight is added according to the position (priority) of features in the layout signature. The similarity distance vector $D = \{d_j\}$, where $j = 1 \ldots m$, is computed between the queried signature s_q and the signatures in S_c as $d_j(s_q, s_j) = \sum f_i w_i$, $(1 \leq i \leq 8)$. The required signature is the one having the maximum

similarity distance $d = max\ (D)$. The weight w_i is assigned according to the feature priority i.e. higher value to the features having frequent appearance in the image. The feature score f_i at the i^{th} feature node of the s_j is computed as:

$$f_i = \frac{\#matched\ elements\ at\ node\ i}{\#existing\ elements\ at\ node\ i\ of\ s_j} \quad 1 \le i \le 8$$

For each node, the number of matched elements between queried signature s_q and original signature s_j is computed by comparing the distance between the element's feature vectors to a threshold T_v. Let $V_q^i(l)$ and $V_j^i(m)$ is the l^{th} and m^{th} element of the i^{th} feature node of s_q and s_j. If the distance $d_{q,j}^{i}(l, m) = ||V_q^i(l) - V_j^i(m)|| < T_v$ then the matching is found and the l^{th} and m^{th} elements are removed from their corresponding i^{th} node, otherwise only the l^{th} element is removed from the i^{th} node of s_q. At each node i, the matching procedure above is carried out until the number of element becomes zero at i^{th} node of either s_q or s_j and then the f_i of that node is computed.

VI. Evaluation and Results

In our evaluation, 310 projected slides from 14 different slideshows, have been captured using a DV camera (Sony, DCR-TRV27E, PAL, 1 mega pixels) and have been then queried on a repository, containing about 1500 slides from 45 different slideshows, in order to find back the original document. For that purpose, all the electronic documents in the repository, mostly in PDF, have been first processed in order to extract their corresponding color and layout signatures. In this evaluation, all the queried captured slide images exist in the repository and the following metrics have been used for measuring our system performances:

$$Identification\ rate\ (I) = \frac{\#correct\ documents\ retrieved}{\#total\ documents\ queried}$$

$$Rejection\ rate\ (R) = \frac{\#documents\ rejected}{\#total\ documents\ queried}$$

Our combined identification method followed two steps: 1) the slides having a similar color distribution are filtered out and then 2) the original document within the remaining set having the most similar layout structure is returned. The first column of Table 1 represents the results for the matching of layout structure alone; whereas the second column shows the results for the combined method, i.e. color plus layout. The identification rate of the combined method is slightly better than the layout feature alone (90% and 88% respectively). Even if in the tested repository, most of the slides have little color variations, the average search space is already reduced to 42% when using the color feature, which is an encouraging result for more colorful repository.

For each signature the matching time is directly proportional to the number of elements in each feature node, which is dependent on the documents' physical content. For the color feature, the matching time is dependent only on the color content and thus the number of parameters is constant for each comparison. Therefore, in the combined features, not only the identification rate is improved but also the identification time is reduced due to the reduction in number of matching parameters. In the worst scenario, the search space could be equal to the whole repository when all the documents have similar color content. The above-mentioned evaluation has been performed on a 1.7 GHz Pentium 4 PC.

TABLE1
Documents Identification Methods Evaluation Results

Slideshow (# slides)	Layout only (Average)				Color + Layout (Average)			
	Search space	I	R	Time (s)	Search space	I	R	Time (s)
34	1.00	0.83	0.00	2.81	0.55	0.88	0.00	1.47
10	1.00	0.90	0.00	2.72	0.15	0.90	0.00	0.61
15	1.00	0.75	0.00	2.68	0.11	0.88	0.00	0.56
28	1.00	1.00	0.00	2.78	0.58	1.00	0.00	1.54
30	1.00	0.92	0.00	2.70	0.59	0.96	0.00	1.79
24	1.00	0.86	0.00	2.63	0.69	0.86	0.00	1.89
19	1.00	1.00	0.00	2.79	0.45	1.00	0.00	1.29
28	1.00	0.96	0.04	2.74	0.44	0.96	0.04	1.31
25	1.00	0.76	0.12	2.70	0.41	0.80	0.12	1.28
20	1.00	0.82	0.00	2.72	0.09	0.82	0.00	0.51
29	1.00	0.79	0.00	2.73	0.09	0.84	0.00	0.52
17	1.00	1.00	0.00	2.68	0.57	1.00	0.00	1.72
15	1.00	1.00	0.00	2.67	0.84	1.00	0.00	2.43
16	1.00	0.71	0.14	2.63	0.31	0.71	0.14	1.16
Total: 310	1.00	0.88	0.02	2.71	0.42	0.90	0.02	1.29

VII. Conclusion

In this article, a document identification method that combines color and layout features is proposed. The result of the evaluation shows that this method solves the low-resolution and color deformation problems due to document image capture from handheld devices. In the near future, our plan is to improve this method by considering one equivalent ellipse per effective peak in the density surface rather than a single ellipse for all, which should convey the number of major color in the images. Furthermore, the spatial distribution of colors in the documents would also be added to the color-based identification.

References

[1] A. Behera, D. Lalanne and R. Ingold "Visual Signature based Identification of Low-resolution Document Images," *ACM Symposium on Document Engineering*, Milwaukee, Wisconsin, 2004, pp. 178–187.

[2] D. Lalanne et al, ''Using static documents as structured and thematic interfaces to multimedia meeting archives", *1st Intl. Workshop on MLMI*, 2004, Martigny, Switzerland, LNCS, vol. 3361, pp. 87-100.

[3] P. Aigrain, H. Zhang, and D. Petkovic, "Content-Based Representation and Retrieval of Visual Media: A State-of-the-Art Review," Multimedia Tools and Applications, 1996, No. 3, pp. 179–202.

[4] M. Petkovic´, "Content-based video retrieval", *7th International Conference on Extending Database Technology*, March 27-31, 2000, Konstanz, Germany, pp 74-77.

[5] Swain and D. Ballard, ''Color Indexing'', *Intl. Journal of Computer Vision*, 1991, vol. 7, no. 1, pp. 11-32

[6] D. Zhang, G. Lu, "Evaluation of Similarity measurement for image retrieval", *IEEE Intl. Conf. on NNSP*, 2003, Nanjing, China.

[7] D. W. Scott, *Multivariate Density Estimation*. New York: John Wiley, 1992.

[8] B. W. Silverman, *Density Estimation for Statistic and Data Analysis.* New York: Chapman and Hall, 1986.

[9] M. J. Jones and J. M. Rehag, "Statistical Color models with Application to Skin Detection,'' *Intl. Journal of Computer Vision*, 2002, vol. 46, no. 1, pp. 81-96.