

Combining Color and Layout Features for the Identification of Low-resolution Documents

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. The combined color and layout features are arranged in a symbolic file, which is unique for each document and is called the document's visual signature. Our identification method first uses the color information in the signatures 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 search space. Finally, our experiment considers slide documents, which are often captured using handheld devices.

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 existing low-resolution captured document identification system makes use of either global image matching methods or of the document textual content using OCR [1]-[3]. However, these methods are generally time-consuming and inadequate for low-resolution images. Furthermore, such systems need a non-empty textual content and a uniform background in order to function efficiently. Slide documents, presented during meetings, conferences, seminars, etc., have generally a poor textual content or textured background, and thus existing methods give back unsatisfactory performances.

The identification method, we propose in this paper, benefits from both the color and the layout features of documents, which are both robust features not only for low-resolution images but also to color deformations due to the various properties of handheld capture devices and to the

varying lighting conditions during capture.

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 and are achieved along with the captured audio/video [4]-[6]. In our smart meeting application [7], such documents are captured using handheld devices and identified by comparing them with their corresponding electronic documents (e.g. PDF, PowerPoint) from which they are generated. After identification, the relevant portion of meeting/lecture/conference could then be retrieved by querying captured document images from the handheld devices on the multimedia repository. The current focus of our work is on the identification of captured projected slides.

In the following section, two different state-of-the-arts approaches, i.e. content-based and layout-based, are introduced along with a brief discussion. Content-based methods exploit the textual content, using OCR, and/or bi-level global image matching methods (pixel-by-pixel) in order to compute the best match. Such global image-based methods do not include low-level visual features such as color, shape, texture, etc. which are often used in image and video retrieval. On the other hand, layout-based methods work on high-resolution document images (≥ 300 dpi) in order to extract both the physical and logical structures of documents, i.e. a decomposition of the documents in logical blocks such as title, abstract, section, figures, etc. The pros and cons of each of those two approaches are listed along with a brief discussion on how they are applied to the identification of low-resolution captured documents.

Due to the poor resolution of the captured documents, and of the poor textual content of slides, state-of-the-art methods do not perform well and for this reason, we propose in this article a method that combines both the color and layout features in order to represent a low-resolution document with a visual signature.

A. Content-based Identification

Content-based identification of captured low-resolution documents has been handled by various research projects. Such projects use, mostly the textual content of the documents rather than the layout features, for linking with other captured medias. Mukhopadhyay *et al* described a method for the identification of slides captured as a video stream using a low-

Manuscript received March 15, 2005. This work was supported by the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2).

Ardhendu Behera is 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@unifr.ch).

Denis Lalanne is with the Computer Science Department, University of Fribourg, CH-1700, Switzerland (e-mail: denis.lalanne@unifr.ch).

Rolf Ingold is with the Computer Science Department, University of Fribourg, CH-1700, Switzerland (e-mail: rolf.ingold@unifr.ch).

resolution camera [1]. The method dilates the binarized segmented slide images and video frames to highlight the text regions and then uses the *Hausdorff distance* to compute the similarity between the text lines. The drawback of this method is that the slide region must be accurately segmented and it works well only for the slides containing texts. Erol *et al* proposed the method for identification of captured slides from the digital cameras using the textual layout and text string [2]. The method uses the binarized image of both captured and original slides and extracts the text strings using OCR. The identification is then based on the matching of text string and the line profiles (horizontal and vertical runs) of the textual part of the slides. This approach performs well for the slides with high-resolution and having some textual contents. Chiu *et al* presented a method for linking of slides with the captured video using a DCT-based image matching [8]. Unfortunately, the method is suitable for high quality slide images, and thus degrades performance for the low-resolution images and also if the image is not properly segmented. Furthermore, partial occlusion and presence of blur could also degrade its performance. Ozawa *et al* demonstrated a method for identification of slides in lecture movies by matching the characters and slide images [9]. The method uses the OCR to recognize the text and an image matching technique (pixel-by-pixel) for matching slides extracted from the video with the original slides. This approach functions only with high-resolution slide images having some textual content of minimum font size of 24 points.

B. Layout-based Identification

Though the layout-based identification is not yet included for the linking of documents with other captured medias but it is widely used in the area of document analysis for the classification, understanding and recognition of the document images [10]-[13]. So, we believe that this feature could be useful and should not be neglected for the identification of the slide documents though it is not feasible to extract the complete layout structure due to poor resolution. In section IV, we have shown how it is useful for the identification of low-resolution captured documents (slides).

The aim of the document layout analysis is to partition the documents into homogeneous regions. Traditional approaches for the layout analysis are typically referred as top-down approach. Such approaches look for the global information of the documents and partition it into blocks and classify into texts or graphics [14]-[16]. Often, each text block is further analyzed till the word level and sometimes character. The above-mentioned approaches performed well for the documents assumed to be rectangular in shape with relatively uniform font and size. However, the performance of such approaches degrades significantly when different components overlap or are closely adjacent to each other.

On the other side, bottom-up approach starts with the local information such as foreground pixels or the connected components. The connected components are extracted from the image and then those of the same type are iteratively

grouped together to form progressively higher-level descriptions of the documents (e.g. words, text lines, blocks, paragraphs, etc.) [17]-[19]. Such methods suffer from the traditional problem of incorrect segmentation due to early groupings. Furthermore, the time complexities of such approaches are higher as compared to the top-down approaches due to the identification analysis and grouping of the connected components. The combinations of the top-down and bottom-up methods are called hybrid methods and one of such algorithm is the split-and-merge algorithm [20]. All the above-mentioned methods work in the image (spatial) domain. However, these algorithms work on a particular layout only.

Often the texture-based analyses are performed to partition the documents into different regions according to the type of texture in that region. The various components of a document image such as text, images or graphics are of different textures. A significant amount of research has been done on the use of multi-channel filtering techniques and the design of Gabor filters for texture segmentation [21]-[23]. The drawbacks of such approaches is that the time complexity is high and in some cases, the regions of different types having similar texture could be confused or merged.

The above-mentioned approaches are mainly used for the documents having the resolution of 300 *dpi* or higher, which is suitable for low-level processing. In our case, the perceived captured document is from a low-resolution capture device and compressed with the quality-losing format such as JPEG (50 – 100 *dpi*) in order to reduce the storage space and speed up the processing. Unfortunately, this result in the loss of some useful details and more noise are brought in. Furthermore, the variation in capture environments (lighting conditions, distance from object, use of flash, etc.) and capture devices create difficulties in the document analysis using one of the above-mentioned approaches. For this reason, the proposed approach considers both the layout as well as color content for the identification of low-resolution captured documents.

The organization of the rest of the paper is as follows: Section II describes the low-level color feature, which is rarely applied to document analysis and specifically, the color histogram-based identification of the documents. In Section III, the density estimation of the normalized 2-D color histogram and the difficulties in the matching of the density surface are presented. The extraction of visual signature containing the color and layout features that represent the respective color density surface and physical layout structure are explained in Section IV. Section V presents the matching of the signatures and evaluation, while the results are presented in Section VI. Finally, we conclude with Section VII.

II. COLOR BASED RETRIEVAL

Since most of the slide images in a slideshow have similar low-level visual features (color, texture and shape) our slide identification system should consider not only the layout

structure of the slide images but also the low-level visual features. Assuming that no textual information about the content of the perceived image is given, the low-level visual features such as color [24]-[26], as well as texture [27] [28] and shape [29] [30], are 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 [31]. Furthermore, such features are very dependent on illumination conditions, shading and compression and for this reason we believe that distribution of features is a better visual representation i.e. more robust to all the cited effects, than an individual feature vector. In our case, we considered color as one of the feature for our signature instead of the texture and shape, since often the slides in different slideshow vary in color rather than in texture and shape. Additionally, the goal is to identify the slideshow of the queried slide using the color feature rather than the exact slide. The exact identification of slides from slideshow is carried out using the layout features.

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 [32]. 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, $0 \leq R, G, B \leq 2^m - 1$ 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 $P = \{r_{ij}, g_{ij}\}$ could be represented with the chromatic values, where $i = 1 \dots n_1$ and $j = 1 \dots n_2$. The reduced color histogram $h(r, g)$ in rg -space is obtained as:

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

Finally, the similarity between any two images I_p and I_q is very often measured by computing the similarity distance between their respective histograms h_p and h_q . *Minkowski distance* is one of the most popular method used to measure the similarity distance and is defined as

$$D(I_p, I_q) = \sum_{x=0}^M \sum_{y=0}^M \{ [h_p(x, y) - h_q(x, y)]^n \}^{n^{-1}} \quad (2)$$

The different values of parameter n gives us different distance measures, for example when $n = 1$ we get the *Manhattan* distance, and for $n = 2$, the *Euclidean* distance. Another measure of the similarity distance of the two histograms is expressed as the intersection of the histograms [32] and is

defined as

$$D(I_p, I_q) = 1 - \frac{\sum_{x=0}^M \sum_{y=0}^M \min\{h_p(x, y), h_q(x, y)\}}{|h_p|} \quad (3)$$

In the histogram representation the drawback is that the shape of the histogram strongly depends on the number of pixels and the method used for image representation. If the image size is small, then there are very few points available for the histogram, which gives rise to erroneous results for the histogram-based comparison. In order 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 [33].

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 $f(x) = \# \text{ samples in one bin} / h$. 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 one dimensional sample space $X = \{x_i\}$, where $i = 1 \dots N$, the kernel density at any point x is estimated as:

$$\hat{f}(x) = \sum_{i=1}^N w_i K\left(\frac{x - x_i}{h}\right) \quad (4)$$

Where K is the kernel function, which determines the shape of the ‘‘bumps’’ placed around the data points in the sample space, h is the bandwidth of the kernel and w_i is the weighting coefficients. Normally, the value of w_i is constant and is $1/(Nh)$. The multivariate kernel density in case of d -dimensional sample space is defined as:

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^N \frac{1}{h_1 \dots h_d} \kappa\left(\frac{x_i - x_1}{h_1}, \dots, \frac{x_i - x_d}{h_d}\right) \quad (5)$$

Where $h_1 \dots h_d$ the bandwidths for each dimension and κ is the d -dimensional kernel function. The d -dimensional kernel functions are commonly represented as the product of the one-dimensional kernel functions i.e.

$$\kappa(u_1, u_2, \dots, u_d) = K(u_1)K(u_2) \dots K(u_d) \quad (6)$$

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:

$$\hat{f}(x) = \frac{1}{Nh^2} \sum_{i=1}^N \left\{ \prod_{j=1}^2 K \left(\frac{x_{ij} - x_j}{h} \right) \right\} \quad (7)$$

The estimation of the kernel density depends on the kernel function and the bandwidth, h . The kernel function decides the shape of the “bumps” placed around the sample for a given bandwidth. 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 [34].

$$K(u) = \begin{cases} \frac{1}{2} c_d^{-1} (d+2) (1-u^T u) & \text{if } u^T u < 1 \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

Where c_d is the volume of the unit d -dimensional sphere and u is the d -dimensional data point. For the density estimation, the shape of the density function is heavily dependent on the chosen bandwidth. The small values of h result in spiky density estimation, which shows the spurious features. On the other hand, too large values of h lead to over-smoothed density estimation that masks the structural features. In our case, we considered the h of 2.5 and 2.0 for the respective original and captured images after evaluating 100 different images for the different values of h ranging from 1.5 to 3.5. Furthermore, we evaluated the above-mentioned density estimation using the optimal Gaussian kernel [35] and obtained similar results. Fig. 1 illustrates the *KDE* of a sample slide document for the bandwidth of 2.5.

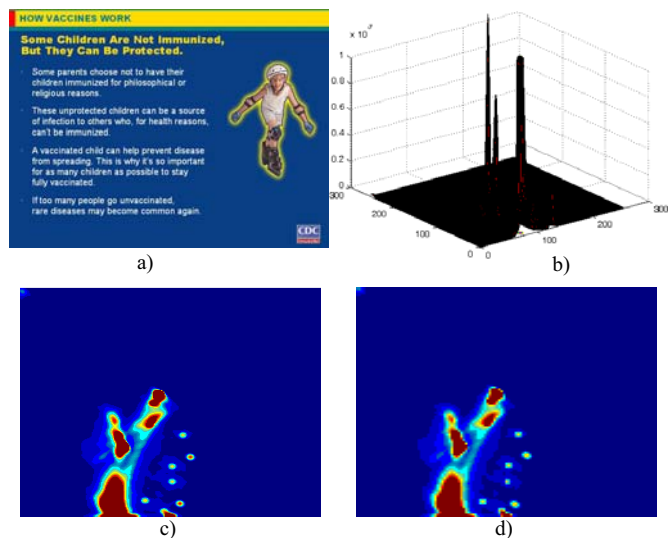


Fig. 1 a) Original image; b) *KDE* of the color distribution in the rg color space; c) its pseudo-color representation for the true color (24-bits) and d) reduced color (21-bits).

Jones and Rebag reported that 77% of the possible 24-bit RGB colors were never encountered on images collected from the web [36]. Furthermore, we observed no perceptible 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. Since the color feature is not used in our method to identify the

exact matching of the slide but instead to identify the slideshows or groups of slides having similar background pattern and color. Therefore, it is judged reasonable to consider for the *KDE*, the 7 most significant bits (*msb*) of each of the RGB channels. This reduces the sample space to $\frac{1}{4}$ th of the actual one, and thus heavily speeds-up the computation time of the *KDE*.

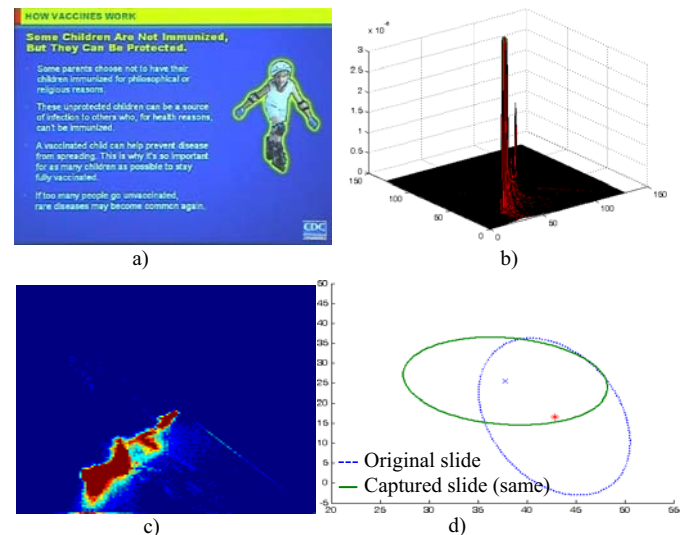


Fig. 2 a) Captured image of Fig. 1, b) its *KDE* of the color distribution, and c) pseudo-color representation for 21-bits in the rg -space and d) equivalent ellipses of the density surfaces of both the original and the captured slide.

The similarity between two images could be measured by computing the distance between their respective *KDE* of the histograms using the equation either (2) or (3). This distance-based similarity measurement is known to perform well for images of the same size with negligible color distortions [4]. In this scenario of captured documents, one faces a problem of non-uniform color shifting in the captured image as compared to the original image. This shift of color is due to the presence of color cast, which is the predominant superimposed color. The color cast is because of variations in the lighting conditions or to the capture device properties. This causes the position(s) of the peak(s) and valley(s) in the density surface is to be different in both the original (Fig. 1b) and captured images (Fig. 2b) and thus, the standard histogram-based similarity distance would not perform efficiently. Furthermore, our aim is to represent the documents with their corresponding signature and identification is based on the matching of the signatures. It is not wise to keep all values of the density surface, which not only takes more matching time but also storage space. So, the reduction of feature space is a better option for both storage and fast matching. In the next section, the extraction of both color and layout features for the building of the signature of both the captured and image format of original electronic documents is described.

IV. DOCUMENT'S VISUAL SIGNATURE

In our identification method, each of the captured and original electronic documents is represented with a signature

containing, mainly of two parts: a) the documents' color distributions and b) the documents' shallow layout structure with the respective labeling. The image of the original electronic slides does not require any preprocessing but the captured document requires rectification to the projected part as the captured documents contain not only the targeted projected part but also the surrounding background areas, which need to be removed.

A. Preprocessing of Captured Documents

The captured images from the handheld devices contain the documents as well as the background. It is thus, necessary to remove the non-document region and rectify the images in case of any skew. The capture devices are assumed to have low radial distortion. Therefore, one needs to consider the four corners of the quadrangle of the projected part and it is mapped to the rectangle of common resolution. This is done using the perspective transform and bi-cubic interpolation [1]. Currently, the selection of the corner points is done manually by clicking on a captured image during a one-time calibration step, although this could be done automatically. If the capture device is fixed, then the calibration is done only once and the same corners of the quadrangle and the transformation matrix are used for the images from that device. Fig. 3 shows one of the projected slides captured from a digital camera and the corresponding rectified projected part of the document. The rectified image is further low-pass filtered for the removal of noise.

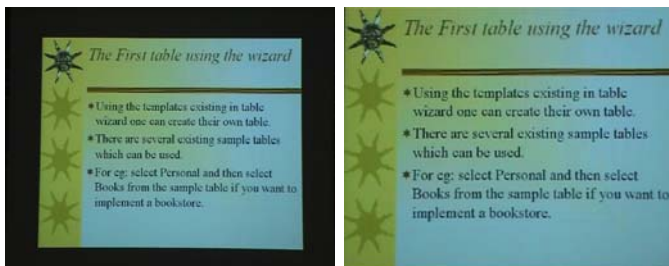


Fig. 3 a) Captured image of a projected slide and b) the corresponding corrected projected part.

B. Color Features Extraction

The color features are computed from the estimated density surface of the normalized color using the Kernel Density Estimation (*KDE*). Once this is done, the distribution of the density surface 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:

$$\begin{aligned} \mu_r &= \int_r r K_d(r, g) dr, \quad \mu_g = \int_g g K_d(r, g) dg \\ \sigma_r^2 &= \int_r (r - \mu_r)^2 K_d(r, g) dr, \quad \sigma_g^2 = \int_g (g - \mu_g)^2 K_d(r, g) dg \end{aligned} \quad (9)$$

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 . Although the density surface of the original and captured images are not the same but often, it

is observed that most of the properties (eccentricity, orientation, etc.) of the *Equivalent Ellipse* of both the captured and original images are preserved and that only the *Equivalent Ellipse* location is shifted (Fig. 2d). 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. When the axes (semi major and semi minor) are equal, then the *Equivalent Ellipse (EE)* becomes *Equivalent Circle (EC)*. In this case, the orientation angle θ is not considered as this is not applicable for a circle. However, practically it is not feasible. Fig. 4 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 adjacent to each other but the orientation and axes are dissimilar, which help to differentiate slides having different colors.

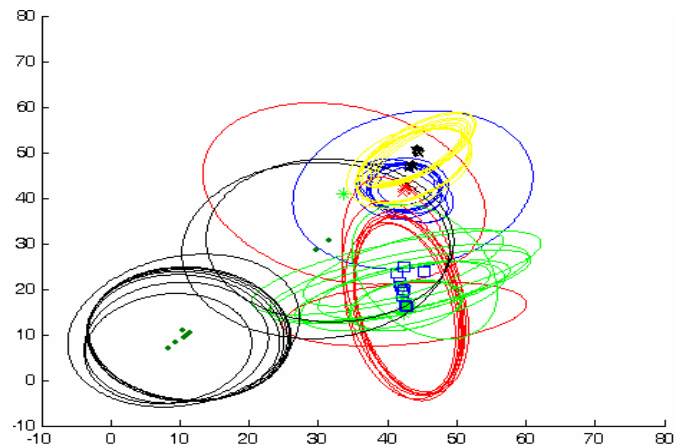


Fig. 4 Equivalent ellipse representation of the estimated color densities in the reduced rg -space of slides randomly picked from 5 different slideshows.

C. 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 compared to other type of documents (e.g. newspaper, articles, etc.). The extraction of the layout signature follows the top-down approach. It, first considers the full slide as a page and partition the page into different blocks (images, text, bars, etc.) and then subsequently, the textual blocks to the word level. Due to poor resolution, it is not feasible to go up to character level as long as the adjacent characters are overlapped in the captured documents. The preprocessed captured documents and the image format of the original

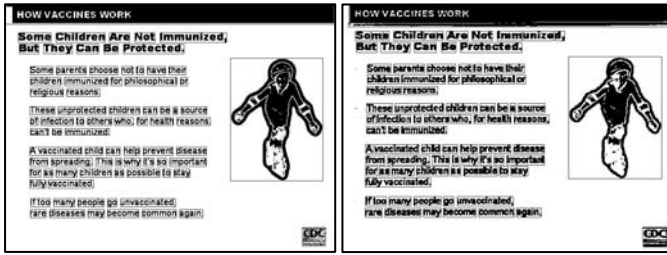


Fig. 5 Layout signatures i.e. bounding boxes for each visual features of the original slide (left) and its corresponding captured slide image (right).

electronic documents are binarized and passed through the Run Length Smearing Algorithm (RLSA) in both horizontal and vertical direction for the extraction of bounding box for each of the homogeneous region [14]. Subsequently, the properties (height, width, black runs, correlation coefficients of the pixels of adjacent lines, etc.) of the regions are looked for and are labeled accordingly. Then the textual parts are separated with individual horizontal and vertical lines and number of words per line using the horizontal and vertical line profiles [37]. The presence of bullets in the text line is checked and if present, they are separated. Graphical objects are differentiated with different labeling such as image, horizontal bars, vertical bars, etc. The detailed extraction procedure for the layout signature of each original electronic slide document and captured slide image is explained in [38]. The layout signature of each slide contains one or more features from the set of features $\{f_1, f_2, \dots, f_8\}$. 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 priority of the features 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. 5 illustrates a document, where each bounding box represents a feature of the visual signature. Fig. 6 shows the XML representation of the visual signature of the documents that combine both the color and layout features.

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_j) close (distance inferior to a threshold T_c) to the color feature of the queried image are considered. Let $S = \{s_1, s_2, \dots, s_n\}$ be the set

```
<VisualSign>
<ColorSignature>
<y="43" x="12" a="25" b="16" theta="-1.45" density="0.485" />
</ColorSignature>
<BoundingBox NoOfBb="10">
<Text NoOfLine="7">
<HasHorizontalText NoOfSentence="7">
<S y="53" x="123" width="436" height="25" NoOfWords="4" PixelRatio="0.40" />
</HasHorizontalText> <HasVerticalText NoOfSentence="0" />
</Text>
<HasImage NoOfImage="3">
<Image y="1" x="16" width="57" height="533" PixelRatio="0.88" />
</HasImage>
<HasBullet NoOfBullets="2">
<Bullet y="122" x="141" width="12" height="12" PixelRatio="1.0" />
</HasBullet>
<Line NoOfLine="0"> <HasHLine NoOfLine="0" /> <HasVLine NoOfLine="0" /> </Line>
<BarWithText NoOfBar="0"> <HBarWithText NoOfBar="0" /> <VBarWithText NoOfBar="0" />
</BarWithText>
</BoundingBox>
</VisualSign>
```

Fig. 6 Visual Signature of a slide document in XML format. The first node of the XML structure represents the color feature and the second one is the layout feature.

of signatures in the repository. After the color matching, a new set $S_c = \{s_1, s_2, \dots, s_m\}$ is derived from S where $m \leq n$.

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. Here, each of the nodes represents one of the layout features of the document, which is the sub-node of the node *BoundingBox* in Fig. 6. The similarity distance vector $D = \{d_j\}$, where $j = 1 \dots m$, 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 adaptively computed during the matching rather than using the predefined weight as described in [38]. In case of pre-defined weight assignment if the error Δ is introduced during the computation of feature score, then it becomes $w_i \times \Delta$ during the computation of distance. To minimize this error, the weight is assigned after considering the number of elements at each feature node of the original electronic documents. This gives the higher priority to the feature node having more number of elements than those having less. The weight, w_i and feature score, f_i at its i^{th} feature node of the signature $s_j \in S_c$ is computed as:

$$\left. \begin{aligned} w_i &= \frac{\# \text{elements at node } i}{\# \text{existing features in } s_j} \\ f_i &= \frac{\# \text{matched elements at node } i}{\# \text{existing elements at node } i \text{ of } s_j} \end{aligned} \right\} 1 \leq i \leq 8$$

For the weight assignment, the original electronic documents are considered rather than the captured images. The probability of introducing error during the extraction process is much less in case of original electronic documents rather than the captured documents due to its high resolution. 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 $d_q^i(l, m) =$

$\|V_q^i(l) - V_j^i(m)\| < T_v$ 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.

TABLE I
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

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 queried on a repository, containing about 1500 slides from 45 different slideshows, in order to retrieve 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 color and layout features and then the corresponding signatures are built. The captured document is pre-processed and then the corresponding signature is extracted. The extracted signature of the captured documents is queried to the system to get the corresponding original matched signature. In this evaluation, the original slides of all the queried captured slide images exist in the repository and the following metrics have been used for measuring our system performances:

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

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

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 physical content of the corresponding document. 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 number of elements in the search space could be the same as in 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.

VII. CONCLUSION

In this article, we proposed a novel document identification method that combines color and layout features. This features combination is reflected in a symbolic file called the visual signature of the document. In order to extract this signature, we proposed using the color density estimation in the rg -color space, which is independent of the illumination geometry. We further proposed to represent the color density surface as an *Equivalent Ellipse*, which not only reduce the feature space but also save the comparison time and signature size. On the other hand, layout features are extracted and structured in order to represent the shallow layout structure of the document. Finally, both color and layout structures are finally used in order to identify documents. The evaluation we performed, querying more than 300 slide images on a repository containing 1500 original documents, has demonstrated that our combined method solves the low-resolution and color deformation problems associated with the 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 colors in the images. Furthermore, the spatial distribution of colors in the documents would also be added to the color-based identification, which should considerably prune the search space, and focus the matching and documents with similar colors and similar spatial distribution. This, should not only speed-up the identification but also improve its performance.

REFERENCES

- [1] S. Mukhopadhyay, and B. Smith, "Passive capture and structuring of lectures," in *Proc. of ACM Multimedia*, 1999, pp. 477-487.
- [2] B. Erol, and J. Hull, "Linking presentation documents using image analysis," in *Asilomar Conf. on Signals, Systems, and Computers*, Nov. 9-12 2003, Pacific Grove, CA.
- [3] D. Franklin, S. Bradshaw, and K. J. Hammond, "Jabberwocky: you don't have to be a rocket scientist to change slides for hydrogen combustion lecture," *Intelligent User Interface*, 2000, pp. 98-105.

- [4] D. Lee, B. Erol, J. Graham, J. J. Hull, and N. Murata, "Portable meeting recorder," in *ACM Multimedia Conference*, 2000, pp. 493-502.
- [5] G. D. Abowd, "Classroom 2000: An experiment with the instrumentation of a living educational environment," *IBM Systems Journal, Special issue on Pervasive Computing*, vol. 38, No. 4, pp. 508-530, 1999.
- [6] P. Chiu, A. Kapuskar, and L. Wilcox, "Meeting capture in a media enriched conference room," in *2nd International Workshop on Cooperative Buildings*, 1999, pp.79-88.
- [7] D. Lalanne, R. Ingold, D. von Rotz, A. Behera, D. Mekhaldi and A. Popescu-Belis, "Using static documents as structured and thematic interfaces to multimedia meeting archives," in *1st Intl. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2004, Martigny, Switzerland, LNCS, vol. 3361, pp. 87-100.
- [8] P. Chiu, J. Foote, A. Girgensohn, and J. Boreczky, "Automatically linking multimedia meeting documents by image matching," in *Proc. of ACM Hypertext '00*, 2000, pp. 244-245.
- [9] N. Ozawa, H. Takebe, Y. Katsuyama, S. Naoi, and H. Yakota, "Slide identification for lecture movies by matching characters and images," in *Proc. SPIE-Documen Recognition and Retrieval XI*, 2004, vol. 5296, pp. 74-81.
- [10] J. Hu, R. Kashi, and G. Wilfong, "Document classification using layout analysis," in *Proc. International Workshop on Database and Expert Systems Applications*, 1999, pp. 556-560.
- [11] C. Shin and D. Doermann, "Classification of document page images based on visual similarity of layout structures," in *Proc. SPIE - Document Recognition and Retrieval VII*, 2000, pp. 182-190.
- [12] A. Dengel, and F. Dubiel, "Clustering and classification of document structure - a machine learning approach," in *Proc. Second International Conf. on Document Analysis and Recognition*, 1993, pp. 587-591.
- [13] E. Appiani, and A.M. Colla, "Automatic analysis and indexing of variable-layout documents," in *Proc. RIAO2000*, Paris, France, April 12-14, 2000, pp. 980-987.
- [14] K. Y. Wong, R. G. Casey, and F. M. Wahl, "Document analysis system," *IBM Journal of Research and Development*, vol.26, pp. 647-656, 1982.
- [15] G. Nagy, and S. Seth, "Hierarchical representation of optically scanned documents," in *Proceedings of International Conference on Pattern Recognition*, 1984, Vol. 1, pp. 347-349.
- [16] H. S. Baird, S. E. Jones, and S. J. Fortune, "Image segmentation by shape-directed covers," in *Proceedings of International Conference on Pattern Recognition*, June 1990, pp. 820-825.
- [17] L. O. Gorman, "The document spectrum for page layout analysis," *IEEE Trans. on PAMI*, vol. 15, pp. 1162-1173, 1993.
- [18] K. Kise, A. Sato, and M. Iwata, "Segmentation of page images using the area voronoi diagram," *Computer Vision and Image Understanding*, vol. 70, pp. 370-382, 1998.
- [19] F. Wahl, K. Wong, and R. Casey, "Block segmentation and text extraction in mixed text/image documents," *Graphical Models and Image Processing*, vol. 20, pp. 375-390, 1982.
- [20] T. Pavlidis and J. Zhou, "Page segmentation and classification," *CVGIP* vol. 54, pp. 484-496, 1992.
- [21] T. Weldon and W. Higgins, "An algorithm for designing multiple gabor filters for segmenting multi-textured images," in *IEEE International Conference on Image Processing*, Chicago, October, 1998, pp. 4-7.
- [22] A.K. Jain, and S.K. Bhattacharjee, "Address block location on envelopes using gabor filters," *Pattern Recognition*, vol. 25, no.12, pp. 1459-1477, 1992.
- [23] A. K. Jain, and Y. Zhong, "Page segmentation using texture analysis," *Pattern Recognition*, 1996, vol. 29, pp. 743-770.
- [24] X.Wan, and C.C.J. Kuo, "Color distribution analysis and quantization for image retrieval," in *Proceedings of SPIE*, vol. 2670, February 1996.
- [25] M. Stricker, M. Orengo, "Similarity of color images," in *SPIE Conference on Storage and Retrieval for Image and Video Databases III*, February 1995, vol. 2420, pp. 381-392.
- [26] 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.
- [27] B.S. Manjunath, W.Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837-842, 1996.
- [28] B.S. Manjunath, J.R. Ohm, V.V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 703-715, 2001.
- [29] A.K. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognition*, vol. 29, no. 8, pp. 1233-1244, 1996.
- [30] J.E. Gary, and R. Mehrotra, "Similar shape retrieval using a structural feature index," *Information Systems*, 18, 7, pp. 525-537, October 1990.
- [31] M. Petkovic, "Content-based video retrieval," in *7th International Conference on Extending Database Technology*, March 27-31, 2000, Konstanz, Germany, pp 74-77.
- [32] M. Swain and D. Ballard, "Color indexing," *Intl. Journal of Computer Vision*, vol. 7, no. 1, pp. 11-32, 1991.
- [33] D. W. Scott, *Multivariate Density Estimation*. New York: John Wiley, 1992.
- [34] B. W. Silverman, *Density Estimation for Statistic and Data Analysis*. New York: Chapman and Hall, 1986.
- [35] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, vol. 33, pp. 1065-1076, 1962.
- [36] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *Intl. Journal of Computer Vision*, vol. 46, no. 1, pp. 81-96, 2002.
- [37] R. Cattoni, T. Coianiz, S. Messelodi, and C. M. Modena, "Geometric layout analysis techniques for document image understanding a review," Technical Report, ITC-IRST, Trento, Italy 1998.
- [38] 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.



Ardhendu Behera (M'05) became a member (M) of ENFORMATIKA in 2005. He received his B.E. degree in Electrical Engineering from the National Institute of Technology, Allahabad, India in 1999, and M.E. degree in System Science and Automation from the Electrical Engineering Department of Indian Institute of Science, Bangalore, in 2001. He is currently pursuing his PhD in the Department of Informatics, University of Fribourg, Switzerland. Previously, he has worked as a Member of Technical Staff (MTS) in the Multimedia group of Sun Microsystems, Bangalore, India. His research interests are Document Image Processing, Image and Video Analysis, Documents Analysis and Recognition.



Denis Lalanne is a senior researcher in the Department of Informatics of the University of Fribourg, Switzerland. He received his B.S. degree in Computer Science from Grenoble, France in 1993, M.S. degree in Cognitive Science from INPG (Institut National Polytechnique Grenoble), France in 1994, and PhD in Computer Science from the Swiss Federal Institute of Technology Lausanne (EPFL) in 1998. Dr. Lalanne has worked as a research member in the USER group (User System Ergonomics Research) of IBM Almaden Research Center, California, as a usability officer in Iconomic systems, a startup based in Switzerland, and as a research/teaching assistant in the University of Avignon (France). His major areas of expertise are Human Computer Interaction, Information Visualization, Artificial Intelligence, Multimedia, and Multimodal Content Management.



Rolf Ingold is a professor in the Department of Informatics, University of Fribourg, Switzerland, as well as director of the DIVA group (Document, Image and Voice Analysis) and head of IM2.DI, an individual project of the National Center of Competence in Research IM2. The DIVA research group covers several topics from the following areas: image processing and analysis, pattern recognition, document analysis and recognition, speech processing. He received his PhD degree in Computer Science from the Swiss Federal Institute of Technology Lausanne (EPFL) in 1989, Switzerland. Prof. Ingold is a member of the editorial board of several international journals as well as a member of the board of directors of the French association GRCE (Groupe de Recherche sur la Communication Ecrite). Current research themes are concentrated on image analysis, content-based image retrieval, as well as multimodal document alignment. His most significant achieved results cover font recognition, structure analysis of composite documents, document modeling, in which the University of Fribourg has become an incontestable leader.