Saliency-Based Detection of Identity Documents Captured by Smartphones

Minh On Vu Ngoc, Jonathan Fabrizio, Thierry Géraud

▶ **To cite this version:**

Minh On Vu Ngoc, Jonathan Fabrizio, Thierry Géraud. Saliency-Based Detection of Identity Documents Captured by Smartphones. IAPR International Workshop on Document Analysis Systems, Apr 2018, Vienna, Austria. hal-01816406

HAL Id: hal-01816406
https://hal.science/hal-01816406

Submitted on 15 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SALIENCY-BASED DETECTION OF IDENTITY DOCUMENTS CAPTURED BY SMARTPHONES

Minh Ôn Vũ Ngọc, Jonathan Fabrizio, Thierry Géraud †

EPITA Research and Development Laboratory (LRDE)
Le Kremlin-Bicêtre, France
Email: firstname.lastname@lrde.epita.fr

ABSTRACT

Smartphones have became an easy and convenient mean to acquire documents. In this paper, we focus on the automatic segmentation of identity documents in smartphone photos or videos using *visual saliency* (VS). VS-based approaches, which pertain to computer vision, have not be considered yet for this particular task. Here we compare different VS methods, and we propose a new VS scheme, based on a recent distance belonging to the scope of mathematical morphology. We show that our resulting saliency maps are competitive with state-of-the-art visual saliency methods, and that such approaches are very promising for use in identity document detection and segmentation, even without taking into account any prior knowledge about document contents. In particular they can perform in real-time on smartphones.

***Index Terms*—** Document detection, Visual saliency, Identity document, Mathematical morphology, Smartphone-based acquisition.

1. INTRODUCTION

Smartphones are able to easily capture images and take videos; thanks to this convenience, many users use smartphones as a tool to acquire documents instead of a traditional scanner. In this paper, we focus on the *detection* of identity documents, such as visas, passports, and identity cards, in photos or videos acquired by a smartphone. This detection task can be actually seen as the segmentation of the image into two parts: the document and the background—note that the term "document segmentation" usually refers to the segmentation of the document contents into several parts. Knowing the precise area of the document allows to guide the user during the image acquisition, to check for forgeries, to properly archive the document, and also to identify the model of document [1, 2]. There are many difficulties in such a real-world mobile-based application: the scene background is unknown; lighting conditions are highly variable (with poor contrast, and unreliable color tones); illumination defects can appear (inhomogeneity, shadows, specular reflections); last, some problems due to the acquisition can occur (out-of-focus blur, motion blur, optical distortions, and noise). In this paper, we assume that the type of identity document present in an image to process is unknown. Typically, we consider situations where we can have passports from different countries, such as in an airport. That implies that documents can have different kinds of contents (layout, text zones, pictures, background). As said before, we want to delineate precisely the document boundary, so its contents (presence of a face photo or of text) is actually of poor help.

To detect documents, the most classical approach is to extract lines from contours as candidates for being a document side [3] (see also [4], which presents a survey on camera-based analysis of documents, and the recent paper [5]). Here we put aside theses approaches, since we are going to explore a radically different approach, the visual saliency-based one. That is why Sec. 2 only focuses on salient object detection[1]. Many salient object detection methods, for use in computer vision, have been recently defined using the Minimum Barrier Distance (MBD) [6], the first ones being [7, 8], and the most recent one being [9]. This particular distance and a distance which derives from it [10], whose computation is very fast, are detailed in Sec. 2. That latter distance is the cornerstone of the method that we present in Sec. 3 to detect documents. This method computes a saliency map, that is, an intensity image where the pixels of salient objects are brighter than the other pixels. Then we binarize this map to obtain the final segmentation result.

The two main contributions of this paper are:

1. an extension to color images (Sec. 3.2) of the Dahu distance, originally defined on gray-level images [10], which allows for computing saliency maps,

2. and a study (Sec. 4) that compares different saliency-based methods for the segmentation / detection of identity documents.

2. SALIENT DOCUMENT DETECTION

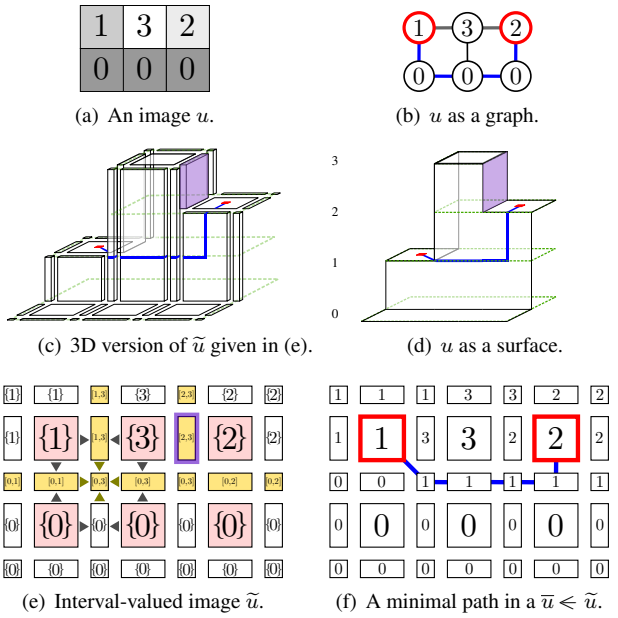
This section describes the saliency map we will use in the document segmentation method presented in Sec. 3.

2.1. Saliency based on the Minimum Barrier Distance

The Minimum Barrier Distance (**MBD**) has been defined in the seminal paper [6], and later studied in [11, 12]. Considering that the image domain is a graph, where vertices repre-

† This work has been conducted in the context of the MOBIDEM project, part of the "Systematic Paris-Region" and "Images & Network" Clusters (France). This project is partially funded by the French Government and its economic development agencies.

[1] It also explains that a comparison between saliency-based methods and some more classical line/contour-based methods is left as future work. Our intent here is only to see whether using saliency can be effective to document detection and segmentation.



(a) An image u. (b) u as a graph.

(c) 3D version of $\widetilde{u}$ given in (e). (d) u as a surface.

(e) Interval-valued image $\widetilde{u}$. (f) A minimal path in a $\overline{u} \lessdot \widetilde{u}$.

Fig. 1. Image representations for computing barrier distances.

sent the discrete points of the domain, we can define paths on this graph. A gray-level image (such as in Fig. 1(a)) is then a vertex-valued graph (such as in Fig. 1(b)). The barrier τ of a path $\pi = \langle ..., \pi_i, ... \rangle$ (π_i being a vertex of the graph) in a gray-level image u is defined by:

$$\tau_u(\pi) = \max_{\pi_i \in \pi} u(\pi_i) - \min_{\pi_i \in \pi} u(\pi_i). \tag{1}$$

The barrier thus represents the gray-level dynamics in u along a path. The minimum barrier distance between x and x' in u is then defined by:

$$d_u^{\text{MB}}(x, x') = \min_{\pi \in \Pi(x, x')} \tau_u(\pi), \tag{2}$$

where $\Pi(x, x')$ denotes the set of all paths between two vertices x and x'. The minimum barrier distance is thus the minimum gray-level dynamics that we can have along a path between two vertices.

A simple illustration is given in Fig. 1. In the graph-based representation depicted in Fig. 1(b), between the two red vertices, multiple paths are possible. The path π corresponding to the sequence of values $\langle 1, 3, 0, 0, 2 \rangle$ is such as $\tau_u(\pi) = 3 - 0 = 3$; it is not minimal since we can have paths with a lower barrier value. A minimal path w.r.t. the MBD is depicted in blue, and we have $d_u^{\text{MB}}(x, x') = 2$.

From a distance, we can derive a *saliency map*, that is, an image where the image value at a point x is the distance of x to a given set of points X'; formally:

$$S_u^{\text{MBD}}(x, X') = \min_{x' \in X'} d_u^{\text{MB}}(x, x'). \tag{3}$$

The computation of a saliency map using the exact MBD is costly [11]; yet some fast but approximate algorithms exist, based on the minimum spanning tree of the image [8]. The next section presents a variant of the MBD, which is also based on the notion of barrier (Eq. (1)), and which leads to an exact and efficient computation of saliency maps.

2.2. The Dahu Distance and the Tree of Shapes

In [10], a "continuous" version of the MBD has been defined, where a gray-level image is interpreted as a surface. An illustration is given in Fig. 1(d) for the image in Fig. 1(a). We can define paths *on this surface*, and a minimal path is depicted in blue in Fig. 1(d), having a barrier of 1 gray-level. This continuous representation of images thus leads to a slightly different distance. Let us now recall briefly how the continuous version of the MBD is defined in [10].

A gray-level image can be seen as a function $u : \mathbb{Z}^2 \to \mathbb{N}$, but such a function is inappropriate to represent a surface such as the one in Fig. 1(d). In [10] the authors have proposed to replace the domain $\mathbb{Z}^2$ by the topological space $\mathbb{H}^2$ of 2D cubical complexes, and the co-domain $\mathbb{N}$ by the set $\mathbb{I}_{\mathbb{N}}$ of intervals on natural numbers. Briefly put, a 2D cubical complex is a set of elements that have a geometrical interpretation: it is composed of squares (2D elements), of segments (1D elements), and of points (0D elements). Fig. 1(e) depicts these elements, where segments and points are respectively drawn as rectangles and tiny squares. The 2D elements correspond to the original pixels of the image (in salmon pink in Fig. 1(e)), whereas the other elements correspond to "what lies between the pixels". From a scalar-valued image u we construct an interval-valued image $\widetilde{u}$ which really represents the surface corresponding to u.

For instance, the scalar image u in Fig. 1(a) can be seen as the surface depicted in Fig. 1(d). The corresponding interval-valued image $\widetilde{u}$ depicted in Fig. 1(e), and in 3D in Fig. 1(c), is a way to represent this surface. Actually, the 0D and 1D elements of the complex which have non-degenerated interval values (in yellow in Fig. 1(e)) encode the vertical parts of the surface. For instance, the 1D element e with the purple border in Fig. 1(e) is such as $\widetilde{u}(e) = [2, 3]$; it represents the vertical part of the image surface depicted in purple both in Fig. 1(c) and in Fig. 1(d).

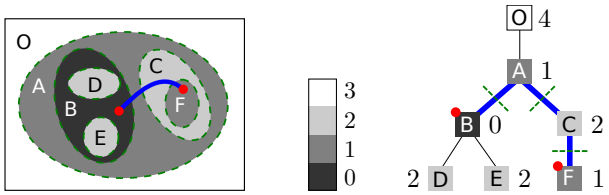
Let us denote by $\lessdot$ the relation between a scalar image and an interval-valued image stating that the values of the pixels of the former are "included" in the interval values of the pixels of the latter; formally: $\overline{u} \lessdot \widetilde{u} \;\Leftrightarrow\; \forall x,\ \overline{u}(x) \in \widetilde{u}(x)$. Fig. 1(f) depicts a scalar image $\overline{u}$ which is "included" in the interval-valued image $\widetilde{u}$ depicted in Fig. 1(e). The adaptation of the minimum barrier distance to an interval-valued image / function, called the Dahu distance [10], is the following:

$$d_u^{\text{DAHU}}(x, x') = \min_{\overline{u} \lessdot \widetilde{u}} d_{\overline{u}}^{\text{MB}}(h_x, h_{x'}) \tag{4}$$

$$= \min_{\overline{u} \lessdot \widetilde{u}} \min_{\pi \in \Pi(h_x, h_{x'})} \tau_{\overline{u}}(\pi), \tag{5}$$

where h_x and $h_{x'}$ are the 2D elements of the complex corresponding to x and x'. Fig. 1(f) depicts in blue a minimal path w.r.t. the Dahu distance; it is obtained with a particular scalar image included in the interval-valued image $\widetilde{u}$ depicted in Fig. 1(e). This minimal path corresponds to the one depicted on the surface of u in Fig. 1(d), and gives a distance of 1 gray-level between the two red pixels.

As compared to the minimum barrier distance (see Eq. (2)) there is an extra combinatorial layer with the minimization "$\min_{\overline{u} \lessdot \widetilde{u}}$". Yet, this new distance can be very easily and efficiently computed thanks to a tree-based representation of the image. The tree of shapes [13, 14] is a morphological decomposition of gray-level images into connected components, called *shapes*, which can be arranged into a tree; indeed, two



(a) Image u. (b) Gray scale. (c) Tree $\mathfrak{S}(u)$.

Fig. 2. The tree of shapes of an image allows to easily express and compute the Dahu distance and saliency maps.

shapes are either disjoint or nested. Quickly said, a shape is the interior of an iso-level line. In Fig. 2, an illustration of a tree of shapes is given and, for instance, the sub-tree $\mathsf{B} \cup \mathsf{D} \cup \mathsf{E}$ corresponds to a shape, the contour of which being a dark-gray iso-level line. In Fig. 2(a), the blue path between the two points $(x,\, x')$ indicated by red bullets in u starts from region B, then goes through A and C, and finally ends in region F. Such a path is minimal because every path in $\Pi(x,\, x')$ should at least cross this same set of level lines to go from x to x'; thus the Dahu distance corresponds to the level dynamics of this set of lines. Actually this path in the image space is exactly *the* path on the tree of shapes between the nodes t_x and $t_{x'}$ containing respectively the two red endpoints x and x'; see the blue path on the tree depicted in Fig. 2(c). In the following, a path on a tree is denoted by $\dot{\pi}$ (to distinguish it from paths in the image space).

The Dahu distance between x and x' can therefore be re-expressed directly on the tree of shapes $\mathfrak{S}(u)$ of u as being the minimum barrier distance between the nodes t_x and $t_{x'}$:

$$\begin{aligned} d_u^{\text{DAHU}}(x,\, x') &= d^{\text{MB}}_{\mathfrak{S}(u)}(t_x,\, t_{x'}) \\ &= \max_{t \,\in\, \dot{\pi}(t_x,\, t_{x'})} \mu_u(t) \;-\; \min_{t \,\in\, \dot{\pi}(t_x,\, t_{x'})} \mu_u(t), \end{aligned} \quad (6)$$

where $\mu_u(t)$ denotes the gray-level associated with the node t of $\mathfrak{S}(u)$. For instance, in Fig. 2(c), the blue path gives the sequence of node values $\langle 0, 1, 2, 1 \rangle$, so the Dahu distance is $2 - 0 = 2$. Eventually, there is *no need* to find the best scalar image $\overline{u} \lessdot \widetilde{u}$, nor the best path $\pi \in \Pi(x,\, x')$ in the image space; it thus means that the primary definition of the Dahu distance (Eq. (5)) is not used as is. The new expression of the distance (Eq. (6)) is just a barrier computation (such as Eq. (1)), but on the trivial path $\dot{\pi}(t_x,\, t_{x'})$ of nodes of the tree of shapes.

2.3. Saliency based on the Dahu Distance

A saliency map of an image u can be derived from this new distance, such as in Eq. (3), except that this new saliency map has a direct expression on the tree of shapes $\mathfrak{S}(u)$. With a set of points X', the corresponding set of nodes on $\mathfrak{S}(u)$ is:

$$T_{X'} \;=\; \{\, t_{x'} \,;\, x' \in X' \,\} \;\subset\; \mathfrak{S}(u). \quad (7)$$

The saliency map from X' based on the Dahu distance can then be expressed by:

$$S_u^{\text{DAHU}}(x, X') \;=\; \min_{x' \in X'} d_u^{\text{DAHU}}(x,\, x') \;=\; S^{\text{MBD}}_{\mathfrak{S}(u)}(t_x, T_{X'}). \quad (8)$$

The major difference with a classical saliency map, defined in the image space (such as the one of Eq. (3)), is that the tree structure is one-dimensional. Since the Dahu distance on the tree (given by Eq. (6)) has the form of a barrier "max - min", the saliency map $S^{\text{MBD}}_{\mathfrak{S}(u)}$ expressed on the tree can be computed by a two-pass procedure (here, downwards then upwards) like the very classical computation of a *chamfer distance map* [15]. Afterward, getting the 2D saliency map S_u^{DAHU} means reading for each x the value of $S^{\text{MBD}}_{\mathfrak{S}(u)}$ at t_x. Eventually, once computed the tree of shapes $\mathfrak{S}(u)$, the computation of a saliency map $x \mapsto S_u^{\text{DAHU}}(x,\, X')$ is instantaneous, whatever the set X'.

Last, let us mention that the representation of an image into a tree of connected components is not memory consuming and is very easy to manipulate [16]. The tree of shapes of an image can be computed in quasi-linear time complexity w.r.t. the number of image pixels [17], and can be parallelized [18].

3. PROPOSED METHOD

We now present a method that relies on saliency maps based on the Dahu distance (Eq. (8)) to detect identity documents.

3.1. Overview of the Method

The method we propose is composed of four steps. **1.** We rely on the SLIC algorithm [19] to simplify the image into superpixels (clusters of pixels, *i.e.*, very tiny regions). This step is interesting because it removes unnecessary image details, and the image can now been seen as a *graph of superpixels*, which has a reasonable size (instead of a huge matrix of pixels). That drastically reduce the number of elements to deal with for the next steps. **2.** To each superpixel we assign its average color, and a tree of shapes is computed from this graph. **3.** We then produce a saliency map from this structure, and we normalize this map (Sec. 3.3). **4.** Finally, we apply a detection step to obtain the resulting detection (Sec. 3.4).

Let us remark that steps 2 and 3 require to compute respectively the tree of shapes and Dahu distances on a color-valued graph; yet both this tree and this distance are originally defined on scalar data (gray-valued images and graphs). So, before giving the method details, we first have to extend these notions to color data.

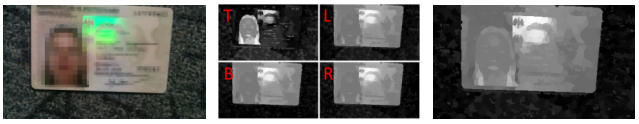
3.2. Extension to Color Data

The tree of shapes, primarily defined on gray-level images, has been recently extended to multi-valued data [20]; this extension is called the *Multivariate Tree of Shapes* (MToS). It yields that we can represent color images by a tree mapping the inclusion of shapes, that is, connected components without holes. Such a representation is of prime importance for computer vision [21] because it satisfies some strong invariance properties featured by natural images, such as local contrast changes [22].

However, the definition of the Dahu distance on the tree of shapes Eq. (6) cannot be used as is; it shall be adapted to take into account that we have color data. Let us now consider that $\boldsymbol{u}$ is a color image, t is a node of the MToS of $\boldsymbol{u}$, and $\mu_{\boldsymbol{u}}(t)$ is the color associated with node t. A superscript i is used to stand for taking one component of the color given by μ. We can then re-write the Dahu distance as follows:

$$\text{with } \tau^{(i)}_{\boldsymbol{u}}(\dot{\pi}) \;=\; \max_{t \in \dot{\pi}} \mu^{(i)}_{\boldsymbol{u}}(t) \;-\; \min_{t \in \dot{\pi}} \mu^{(i)}_{\boldsymbol{u}}(t), \quad (9)$$

$$d_{\boldsymbol{u}}^{\text{DAHU}}(x,\, x') \;=\; \textstyle\sum_{i \in \{R,G,B\}} \tau^{(i)}_{\boldsymbol{u}}\big(\, \dot{\pi}(t_x,\, t_{x'})\,\big). \quad (10)$$



(a) Input (b) 4 saliency maps (c) Fusion

Fig. 3. Effect of fusing four side-specific maps using Eq. (11).

This distance is therefore the sum of the lengths of the 3 sides of the minimum 3D bounding box of the set of colors corresponding to the nodes along the path between t_x and $t_{x'}$. This modified Dahu distance can now be used to compute the saliency map of Eq. (8). [2]

3.3. Obtaining a Relevant Saliency Map

We assume that the four sides of the image boundary are mostly composed of the scene background (*i.e.*, the document does not predominantly touch the image boundary). Hence, from each boundary side of the image, we compute a saliency map; for instance, with X_{top} being the set of pixels of the image top row, we have the saliency map $S_{\boldsymbol{u}}^{\text{DAHU}}(x,\, X_{\text{top}})$. We end up with 4 saliency maps, depicted in Fig. 3(b), that we combine in a pixel-wise way using:

$$S_{\boldsymbol{u}}^{\text{DAHU}}(x) \;=\; \sum_{i\in\{\text{top, left, right, bottom}\}} S_{\boldsymbol{u}}^{\text{DAHU}}(x,\, X_i) \;/\; 4. \quad (11)$$

An example is given in Fig. 3. As we can see in Fig. 3(a), the fact that the document touches the top row gives an irrelevant saliency map $S_{\boldsymbol{u}}^{\text{DAHU}}(x,\, X_{\text{top}})$, marked T in Fig. 3(b). However, after the fusion of the 4 maps, we obtain a satisfy result, which is depicted in Fig. 3(c).

Similarly to some previous works [23, 11, 12], we normalize the saliency map by using "a - b" normalization (with $a = 0.1$ and $b = 0.8$), followed by an adaptive contrast enhancement with a sigmoid mapping. The saliency map in Fig. 3(c) is depicted *after* normalization in the 2nd row of Fig. 5(f).

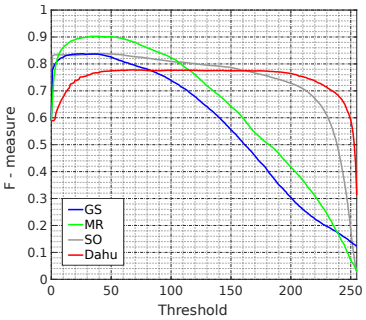
3.4. Final Detection Step

The final detection step consists in deducing a binary image from the saliency map obtained by Eq. (11). Our detection step is still experimental (briefly put, we only search for a threshold so that the result looks like a quadrilateral); it is not emphasized in this paper, since we focus on comparing gray-level saliency maps w.r.t. all possible thresholds in Sec. 4.2. Though, with this simple detection step, some preliminary results are depicted in Fig. 5(g) with the following color code: white for true positives, red for false negatives, and green for false positives.

[2] Please note that, although Eq. (10) looks simple, we have here a strong result. To be able to compute visual saliency maps (efficiently, and based on the very effective Minimum Barrier Distance) while taking into account colors, we need to compute a particular distance between two points. This distance is the one of an *optimal* path between two points in the image space, this path being such that the set of colors on the path has the smallest bounding box in the color space. Precisely, the distance between the 2 points is the diameter (with the L^1 norm) of this 3D bounding box. This is a highly combinatorial problem, far to be trivial, and which cannot be solved efficiently in the image space. Our contribution here is to turn this problem into an efficient straightforward computation in a tree space.

Method	MAE	$\overline{F_\beta}$
GS [24]	0.328	0.573
MR [25]	0.299	0.642
SO [26]	0.265	**0.7461**
Dahu	**0.178**	**0.7465**

(a) MAE and $\overline{F_\beta}$



(b) F_β w.r.t. saliency map thresholding

Fig. 4. Numerical comparison of saliency maps.

4. EXPERIMENTAL RESULTS

To know how our Dahu-distance-based saliency method performs in the context of identity document segmentation, we are going to compare it with some other similar approaches.

4.1. Some Other Saliency Detection Methods

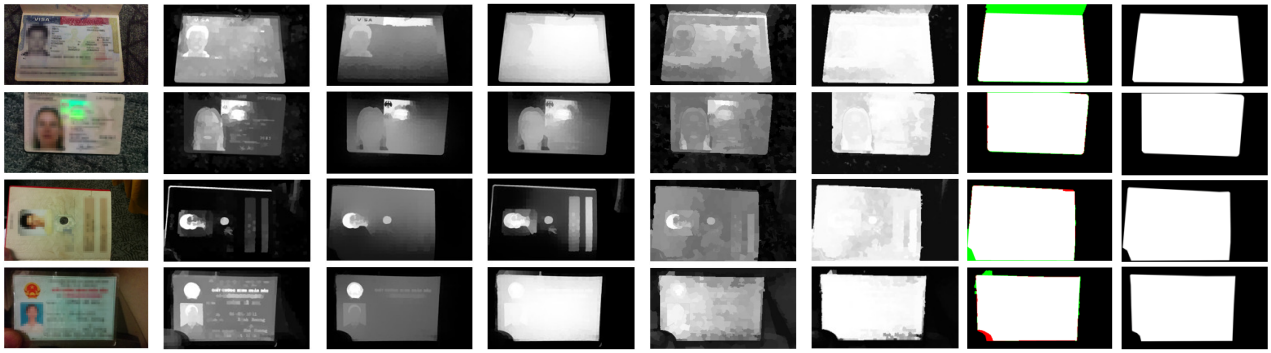
Let us now present three state-of-the-art methods of salient object detection, that we are going to compare our method with. In [24] the saliency detection is based on a geodesic distance (**GS**) which uses background priors. The major assumptions are that the background is usually large, homogeneous, and located near the boundary of the image. In [25] the saliency detection relies on a bottom-up approach to choose some regions by manifold ranking (**MR**) on a graph of superpixels. Such as in Sec. 3.3, the authors compute 4 maps and fuse them. In these maps, the superpixels are ranked w.r.t. the similarity with some seeds located in the image boundaries. In [26], a saliency optimization method (**SO**) is proposed which combines multiple saliency measures, one of them using the notion of "boundary connectivity". Note that all these methods also rely on a post-processing step to "normalize" the resulting saliency maps.

4.2. Dataset and Experiments

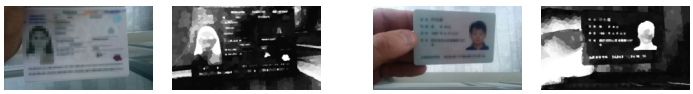
For our experiments, we have built a dataset of identity documents[3]. We have a dozen of different types of visas and passports from various countries. We recorded over 100 videos under different environment conditions, using several kinds of smartphones. From these videos, we selected 100 frames to create our dataset, so that it presents some realistic difficulties such as out-of-focus and motion blur, inhomogeneous illuminations, etc. Then, we generated semi-automatically the corresponding ground-truth images.

We compare our method with the state-of-the-art saliency-based detection methods presented in the previous section. We use two distinct measures: **1.** the Mean Absolute Error (MAE), which is the average difference between a saliency map S (gray-level image) and a ground-truth image GT (binary image): $\text{MAE} = (\sum_x |GT(x) - S(x)|) \,/\, N$, with N being the number of pixels, and **2.** an F_β-measure defined by: $F_\beta = (1+\beta^2) \times P \times R \,/\, (\beta^2 \times P + R)$, where P and R

[3] Available at `http://publications.lrde.epita.fr/movn.18.das`



(a) Input (b) GS [24] (c) MR [25] (d) SO [26] (e) Dahu-based... (f) ...normalized (g) Detection (h) Ground truth



(i) Some failure cases of the Dahu-based approach

Fig. 5. Comparison of our saliency maps with other classical or state-of-the-art methods.

are respectively the precision and the recall, and with $\beta^2 = 0.3$ (it is the classical setting in the visual saliency community). To compute the precision and recall scores, for each image to process, we simply binarize the corresponding gray-level saliency map with a threshold sliding from 0 to 255. Then, for every threshold, we compare the obtained binary map with the ground-truth map. For a given threshold, we depict in Fig. 4(b) the average F_β-measure obtained on the dataset of 100 images. The "global" F_β-measure, averaged for all thresholds (and all images), is denoted by $\overline{F_\beta}$. The values of $\overline{F_\beta}$ and the MAE scores for all the compared methods are depicted in the table in Fig. 4(a); note that the better a method is, the lower MAE values are, and the higher $\overline{F_\beta}$ values are. First, we can observe that, over the years, the state-of-the-art methods give better results (first GS, then MR, and last SO). Second, the Dahu-based approach gives the lowest MAE score, and slightly outperforms the SO method for the $\overline{F_\beta}$ criterion.

If we look at the F_β-measure curves for the different thresholds in Fig. 4(b), there are two main observations. First, the methods SO (in gray) and Dahu (in red) have stables / flat curves, which is an advantage, because the "best" threshold remains unknown and depends on the image. Conversely, for the GS and MR methods (respectively in blue and green), the curves are not stable, which means that taking a threshold might not be a very robust task. The second observation is that the "best" method with respect to the F_β-measure seems to be the MR method, with a rather low threshold (around 50). Though, the MR method is computationally expensive so it cannot run in real-time on smartphones, whereas the Dahu-based approach can.

Some qualitative illustrations on a few images (Fig. 5(a)) are depicted in Fig. 5. The prominent observation is that the compared saliency methods, from Fig. 5(b) to Fig. 5(f), have rather different behaviors. The one based on the Dahu distance, so on the principle of a *barrier* (see Eq. (1), Eq. (6), and Eq. (9)) is effective: the main barrier is visible around the documents, even *before* normalization; see Fig. 5(e). Also we can notice that the saliency values *inside* the documents are much more uniform with the Dahu-based method than with the other saliency-based methods.

4.3. Limitations and Perspectives

The major limitation of saliency-based methods is due to low contrast; some failure cases are depicted in Fig. 5(i). The left image is blurred and the contrast between the document and the background is poor, so the document cannot be detected. In the right image, the identity card has a color similar to the one of the background, so the salient objects are the hand and the portrait. Actually, as perspectives, the method we present can be improved through taking into account some extra prior information such as "text texture", and can be combined with more classical contour/line-based approaches.

5. RELATED WORK

Actually, there exists a short state of the art of document detection, contrasting from methods to extract lines as candidates for the document sides, and being related to the one presented here. In [27], after down-sampling, some seeds are located in the image, and the "geodesic object proposals" method [28] extracts from these seeds a set of regions; the best candidate region is then elected as being the document. In [29] and [30], the authors proposed a method based on the tree of shapes [20]. For each shape (node of the tree / connected component without hole), an energy is computed being the sum two terms: one measuring how the shape fits a quadrilateral, and the other one measuring the degree of "text texture" of the contents of the shape. The shape with the highest energy is considered as the candidate for document detection. This approach won the first challenge (detection of a document page in videos captured by smartphones) of the SMARTDOC competition, organized for ICDAR 2015 by Burie *et al.* [3]. The work presented in this paper, relying on a saliency map computed on the tree of shapes, is clearly derived from it.

6. CONCLUSION AND PERSPECTIVES

In this paper, we have presented an extension of the Dahu distance to color images, which allows for computing some saliency maps for object detection purpose. We have proposed a framework to detect identity documents in photos or videos captured by smartphones based on saliency maps, with very few prior knowledges about the documents and the images. We only take into account that the document looks like a quadrilateral and does not mostly touch the image boundary. Our main conclusion (and contribution) is that visual saliency approaches are relevant to document detection. Moreover, while remaining efficient (both in time and memory usage), which is critical in embedded software, we have the potential to offer better results than the one presented here, using some extra knowledge. Indeed, finding some text [31] or a face photograph can help the final decision step in locating the document, though that does not directly help delineating the document boundary. Last, we will also consider images acquired by tablets and webcams to test the robustness of the saliency approach.

Acknowledgments. The authors would like to thank Nicole Vincent and Jean-Christophe Burie for their valuable comments on a preliminary version of this work.

References

[1] L. de las Heras *et al.*, "Use case visual bag-of-words techniques for camera based identity document classification," in *Proc. of ICDAR*, 2015, pp. 721–725.

[2] R. Sicre, A. Awal, and T. Furon, "Identity documents classification as an image classification problem," Inria Rennes - Bretagne Atl.; IRISA; AriadNext, Tech. Rep. RT-0488, 2017.

[3] J.-C. Burie *et al.*, "ICDAR 2015 competition on smartphone document capture and OCR (SmartDoc)," in *Proc. of ICDAR*, 2015, pp. 1161–1165.

[4] J. Liang, D. Doermann, and H. Li, "Camera-based analysis of text and documents: A survey," *International Journal on Document Analysis and Recognition*, vol. 7, no. 2, pp. 84–104, 2005.

[5] K. Bulatov *et al.*, "Smart IDReader: Document recognition," in *Proc. of IAPR CBDAR*, 2017, pp. 39–44, to appear.

[6] R. Strand *et al.*, "The minimum barrier distance," *Comp. Vision and Image Understanding*, vol. 117, no. 4, pp. 429–437, 2013.

[7] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in *Proc. of ICCV*, 2015, pp. 1404–1412.

[8] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proc. of CVPR*, 2016, pp. 2334–2342.

[9] X. Huang and Y. Zhang, "Water flow driven salient object detection at 180 fps," *Patt. Rec.*, vol. 76, pp. 95–107, 2018.

[10] T. Géraud, Y. Xu, E. Carlinet, and N. Boutry, "Introducing the Dahu pseudo-distance," in *Proc. of ISMM*, ser. LNCS, vol. 10225, 2017, pp. 55–67.

[11] K. C. Ciesielski *et al.*, "Efficient algorithm for finding the exact minimum barrier distance," *Computer Vision and Image Understanding*, vol. 123, pp. 53–64, 2014.

[12] R. Strand *et al.*, "The minimum barrier distance: A summary of recent advances," in *Proc. of DGCI*, ser. LNCS, vol. 10502. Springer, 2017, pp. 57–68.

[13] P. Monasse and F. Guichard, "Fast computation of a contrast-invariant image representation," *IEEE Transactions on Image Processing*, vol. 9, no. 5, pp. 860–872, 2000.

[14] V. Caselles and P. Monasse, *Geometric Description of Images as Topographic Maps*, ser. LNM. Springer, 2009, vol. 1984.

[15] G. Borgefors, "Distance transformations in arbitrary dimensions," *CVGIP*, vol. 27, no. 3, pp. 321–345, 1984.

[16] E. Carlinet and T. Géraud, "A comparative review of component tree computation algorithms," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3885–3895, 2014.

[17] T. Géraud, E. Carlinet, S. Crozet, and L. Najman, "A quasi-linear algorithm to compute the tree of shapes of n-D images." in *Proc. of ISMM*, ser. LNCS, vol. 7883, 2013, pp. 98–110.

[18] S. Crozet and T. Géraud, "A first parallel algorithm to compute the morphological tree of shapes of nD images," in *Proc. of IEEE ICIP*, 2014, pp. 2933–2937.

[19] R. Achanta *et al.*, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[20] E. Carlinet and T. Géraud, "MToS: A tree of shapes for multivariate images," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5330–5342, 2015.

[21] F. Cao, J.-L. Lisani, J.-M. Morel, P. Musé, and F. Sur, *A Theory of Shape Identification*, ser. LNM. Springer, 2008, vol. 1948.

[22] V. Caselles, B. Coll, and J.-M. Morel, "Topographic maps and local contrast changes in natural images," *International Journal on Computer Vision*, vol. 33, no. 1, pp. 5–27, 1999.

[23] G. Wang, Y. Zhang, and J. Li, "High-level background prior based salient object detection," *Journal of Visual Communication and Image Representation*, vol. 48, pp. 432–441, 2017.

[24] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," *Proc. of ECCV*, pp. 29–42, 2012.

[25] C. Yang *et al.*, "Saliency detection via graph-based manifold ranking," in *Proc. of ICPR*, 2013, pp. 3166–3173.

[26] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. of ICPR*, 2014, pp. 2814–2821.

[27] L. R. Leal and B. L. Bezerra, "Smartphone camera document detection via Geodesic Object Proposals," in *Proc. of IEEE LA-CCI*, 2016, pp. 1–6.

[28] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *Proc. of ECCV*, vol. 5. Springer, 2014, pp. 725–739.

[29] Y. Xu, T. Géraud, and L. Najman, "Hierarchical image simplification and segmentation based on Mumford-Shah-salient level line selection," *Pattern Recognition Letters*, vol. 83, no. 3, pp. 278–286, 2016.

[30] Y. Xu, E. Carlinet, T. Géraud, and L. Najman, "Hierarchical segmentation using tree-based shape spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 3, pp. 457–469, 2017.

[31] L. D. Huỳnh, Y. Xu, and T. Géraud, "Morphology-based hierarchical representation with application to text segmentation in natural images," in *Proc. of ICPR*, 2016, pp. 4029–4034.