A Coarse-to-Fine Approach for Layout Analysis of Ancient Manuscripts

Abedelkadir Asi, Rafi Cohen, Klara Kedem, Jihad El-Sana
Department of Computer Science
Ben-Gurion University
Beer-Sheva, Israel
abedas,rafico,klara,el-sana@cs.bgu.ac.il

Itshak Dinstein
Department of Electrical and Computer Engineering
Ben-Gurion University
Beer-Sheva, Israel
dinstein@ee.bgu.ac.il

***Abstract*—Many applications along the manuscript analysis pipeline rely on the accuracy of pre-processing steps. Perfectly detecting the main text area in ancient historical documents is of great importance for these applications. We propose a learning-free approach to detect the main text area in ancient manuscripts. First, we coarsely segment the main text area by using a texture-based filter. Then, we refine the segmentation by formulating the problem as an energy minimization task and achieving the minimum using graph cuts. The energy function is derived from properties of the text components. Spatial coherence of the segmented text regions is explicitly encouraged by the energy function. We evaluate the suggested method on a publicly available dataset of 38 historical document images. Experiments show that the suggested approach outperforms another state-of-the-art page segmentation method in terms of segmentation quality and time performance.**

***Keywords*-Layout Analysis; Page Segmentation; Historical Documents; Graph Cuts; Statistical Inference**

I. Introduction

Ancient manuscripts pose significant research challenges for the document analysis community [1]. Irregular layout format is a typical obstacle that researchers have been working to overcome [2], [3]. Many manuscripts include text in page margins (side-notes) as remarks on the text appearing in the main page frame (main-text). While the main-text is mostly horizontally oriented, side-notes can be of different orientations and locations on the page margins as shown in Figure 1. Accurate segmentation of both text regions from each other is of great importance for many applications along the manuscript analysis pipeline.

In general, image segmentation methods can roughly be divided into two categories: *top-down methods* [5]–[7] and *bottom-up methods* [8]–[11]. In top-down methods, the image is coarsely segmented and a subsequent refining process is applied. Bottom-up methods aggregate elementary units in the image, e.g., pixels or connected components, into larger regions which define distinct image classes. The aggregation process usually optimizes a cost function to meet a specific criteria. We refer the interested reader to the comprehensive survey on document structure analysis [3], [12].

Bukhari et al. [7] suggested a component-based technique to segment main-text from side-notes in Arabic historical manuscripts. They applied a multi-layer perceptron classifier

(a) (b)

Figure 1. Ancient manuscript images with complex layout format. Samples are publicly available at the Islamic manuscripts digitization project web page [4], Leipzig university library.

to obtain a coarse segmentation and then used a local voting step to produce the final segmentation. The refinement process relies on a pre-defined window size which is defined as a function of multiple parameters, such as image resolution, height and width of connected components.

We propose a learning-free approach for segmenting the main-text region from side-notes in historical documents. Our approach is based on a coarse-to-fine scheme that coarsely segments the main-text area using Gabor filter and then applies a global refinement scheme. The refinement scheme is based on minimizing an explicit energy function which is derived from properties of the text components, e.g., location, stroke width and area. The explicit energy function makes our algorithm flexible and can adapt to various energy terms.

The rest of this paper is organized as follows: Section II describes the energy minimization framework; Section III introduces the details of our approach; in section IV we present experimental results of the suggested technique. We conclude and present future work in section V.

II. Energy Minimization with Graph Cuts

Our approach relies on the energy minimization framework suggested by Boykov et al. [13] where they used graph cuts to approximate energy minimization of arbitrary functions. This framework perceives the segmentation problem as a labeling problem where every component, c, is assigned a label l. The goal is to find a labeling $\mathcal{L}$ that assigns each component c a label l_c, where $\mathcal{L}$ is both consistent with the observed data and spatial coherent. The energy function, $E(\mathcal{L})$, consists of two terms: the cost and the smoothness terms. The cost term, $D(c, l_c)$, expresses the cost of assigning the connected component c the label l_c. The smoothness term determines the coherence of the labels l_c and $l_{c'}$ with the spatial relation of the components c and c'. Let C be the set of components in the document and let $\mathcal{N}$ be the set of adjacent component pairs, according to this framework, minimizing the energy function in Equation (1) produces the appropriate labeling. The coefficient $d(c, c') \cdot \delta(l_c \neq l_{c'})$ in Equation (1) insures that the closer the components are the higher is the chance that they got assigned the same label, where, $d(c, c')$ represents a distance measure between components, and $\delta(l_c \neq l_{c'})$ is 1 if the condition inside the parentheses holds and 0 otherwise.

$$E(\mathcal{L}) = \sum_{c \in C} D(c, l_c) + \sum_{\{c,c'\} \in \mathcal{N}} d(c, c') \cdot \delta(l_c \neq l_{c'}) \quad (1)$$

In general this problem is NP-hard [14], however, for two labels, the optimal assignment could be obtained in polynomial time.

III. Our Method

Our method coarsely segments the main-text region and refine the segmentation by applying the minimization framework. The first stage relies on texture characteristics of text regions and the second stage exploits properties of the text, i.e., location, stroke width and component area.

According to historians the main-text and the side-notes were usually written by different writers. This observation motivates the use of Gabor filter which is widely used for texture segmentation as we assume that each writing style can be characterized by a distinct texture. The nearly rectangular shape of the main-text region is a significant prior that a human would rely on to segment this region. In addition, side-notes usually have a different texture with respect to the main-text, however, it can still appear in a horizontal orientation, as shown in Figure 1(a).

A. Coarse Segmentation

Gabor filters are particularly appropriate for capturing texture [15]. It was shown that the smooth terms of the Gaussian envelope of the Gabor filter play a major role in texture classification [16]. We define these terms as a function of the average heights of connected components in the document. Applying this filter generates a very high filter response in the main-text area and suppresses the responses from page margins (see Figure 2(b)). We employ hysteresis thresholding using two threshold values t_{low} and t_{high} to generate a coarse binary mask of the main-text area, as depicted in Figure 2(c). In hysteresis thresholding, pixels with responses higher than t_{high} are assigned the value 1 and pixels with responses below t_{low} are assigned the value 0. Pixels with responses between t_{low} and t_{high} are assigned the value 1 if they can be connected to a pixel with a response above t_{high} through a chain of other pixels with responses above t_{low}. This binary mask coarsely captures the nearly rectangular shape of the main-text region (see Figure 2(d)).

B. Global Refinement

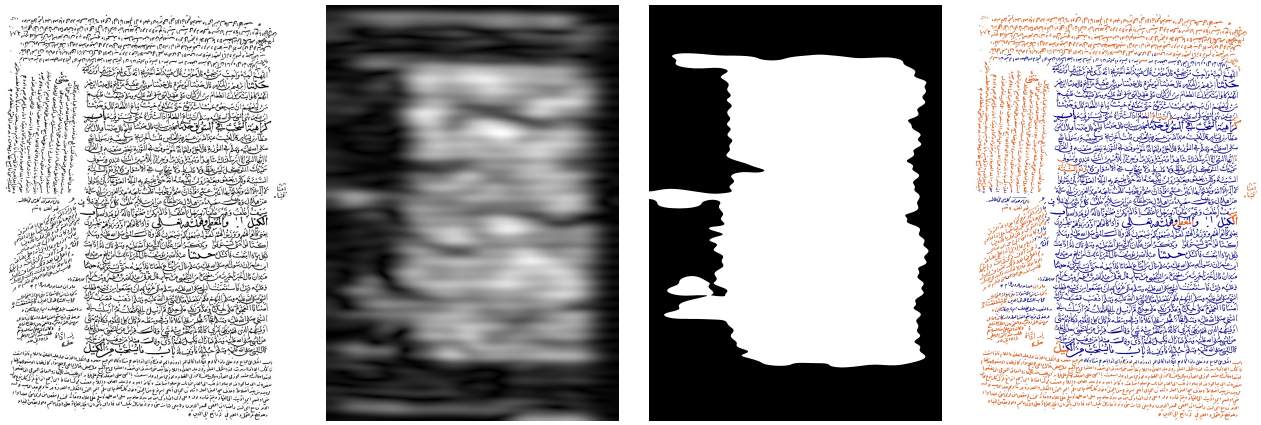
At this stage of the algorithm we aim to globally refine the coarse segmentation from the previous step. We employ the minimization framework by defining both the cost and the smoothness terms. As stated earlier, closer pairs of components are expected to have a higher probability to have the same label. Therefore, we use an 8-connected grid of connected components and define the distance $d(c, c')$ in Equation (1) according to Equation (2) (the spatial coherence strength decays exponentially with Euclidean distance [17]).

$$d(c, c') = \exp(-\alpha \cdot d_e(c, c')) \quad (2)$$

The term $d_e(c, c')$ is the Euclidean distance between the centroids of components c and c', and the constant α is defined as $(2 \langle d_e(c, c') \rangle)^{-1}$, where $\langle \cdot \rangle$ denotes expectation over all pairs of adjacent elements [18].

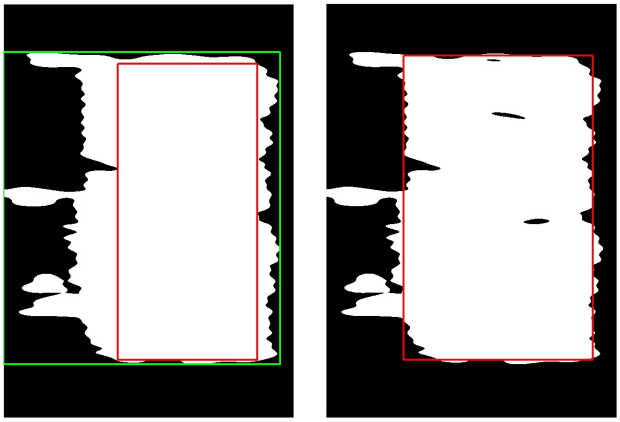
Since the shape of the main-text area is nearly rectangular, we use the mask extracted from the Gabor filter to approximate the rectangular shape. The bounded rectangle in the mask is usually too restrictive, whereas the bounding rectangle is too large, as shown in Figure 3(a). Therefore, we choose the rectangle $\mathcal{R}$ which maximizes the difference between foreground and background pixels, see Figure 3(b).

Now we are going to define the cost of assigning a particular component c with a specific label while distinguishing between the costs of the main-text and side-notes labels. We define the cost of assigning a main-text label, denoted ℓ_{mt}, to the component c as a function of the component location with respect to $\mathcal{R}$ boundaries. Namely, a component that resides within $\mathcal{R}$ has low cost, a component that is located far from $\mathcal{R}$ has a relatively high cost and components located around $\mathcal{R}$ boundaries have intermediate costs. These costs can be adequately modeled by a sigmoid function as depicted in Equation (3). The parameters h, w are the height and width of the document respectively, a and b are constants and $SDT(c, \mathcal{R})$ is the signed distance transform of c with respect to $\mathcal{R}$ boundaries and it is formulated in Equation (4). Figure 4(a) illustrates this cost on a particular document.



(a) (b) (c) (d)

Figure 2. Coarse segmentation of the main-text region: (a) Binary image (b) Gabor filter response (c) Binary mask by hysteresis thresholding (d) Coarse segmentation, blue and red colors represent main-text and side-notes labels, respectively.



(a) (b)

Figure 3. (a) The bounded (red) and bounding rectangles (green) (b) the rectangle which maximizes the difference between foreground and background pixels.

$$D_{rect}(c,\ell_{mt}) = \left(1+\exp^{\frac{-a\cdot SDT(c,\mathcal{R})}{min(h,w)}}\right)^{-1} \tag{3}$$

$$SDT(c,\mathcal{R}) = \begin{cases} d(c,\overline{\mathcal{R}}) & \text{for } c \in R \\ -d(c,\mathcal{R}) & \text{for } c \in \overline{\mathcal{R}} \end{cases} \tag{4}$$

The cost of assigning a side-notes label, denoted ℓ_{sn}, for a component c is small if it is outside $\mathcal{R}$ and large if it has a larger stroke width or larger area than the average stroke width or average area of all the components within $\mathcal{R}$. The stroke width of a component, c, is coarsely approximated by $SW(c) = \frac{|S|}{|D|}$, where $|S|$ is the number of foreground pixels of c, and $|D|$ is the number of pixels on it's contour. The component area is simply the number of foreground pixels belonging to c. The stroke width cost is defined according to Equation (5), where μ_{sw} and σ_{sw} are the average stroke width and the standard deviation of the components within $\mathcal{R}$. We define the area cost, $D_{area}(c,\ell_{sn})$, similarly. See Figure 4(b)-(c) for an illustration of the aforementioned costs terms. The cost of assigning a side-note label for a component c, is the sum of the three costs given in Equation (6), where w is a weighting constant.

$$D_{sw}(c,\ell_{sn}) = \exp\left(\frac{sw(c)-\mu_{sw}}{2\sigma_{sw}} - 1\right) \tag{5}$$

$$\begin{aligned} D(c,\ell_{sn}) &= 1 - D_{rect}(c,\ell_{mt}) \\ &+ w\cdot(D_{area}(c,\ell_{sn}) \\ &+ D_{sw}(c,\ell_{sn})) \end{aligned} \tag{6}$$

IV. Experimental Results

We evaluated our method on the dataset[1] and ground truth data presented by Bukhari et al. [7] to provide a comparative performance evaluation. The dataset includes 38 historical document images from 7 different manuscripts. The considered documents contain side-notes with various writing styles and orientations. The presence of side-notes on page margins makes the page layout complex and irregular. We adopted the F-measure metric to evaluate the performance of our algorithm. Precision and recall were estimated according to Equations (7) and (8). *True-Positive (TP)* and *False-Negative (FP)* are defined as the rate of a main-text component classified as main-text and side-note, respectively, and *False-Negative (FN)* is the rate of a side-note component classified as main-text.

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

[1]Available online at http://www.cs.bgu.ac.il/~abedas

(a) (b) (c) (d)

Figure 4. Illustration of various cost maps for a particular document. Cold colors (bluish) represent low costs and hot colors (reddish) represent high costs; (a) the distance from $\mathcal{R}$ cost (b) stroke width cost (c) the component area cost(d) the final segmentation for the document image from Figure 2.

In [7], 28 images where used for training and 10 for testing. Table I compares our results with the results of [7] on the 10 images of the testing set. Table II presents the performance of our learning-free method on the entire dataset.

	Our Method		Bukhari et al. [7]	
	Before	After	Before	After
Main-text	97.4%	**99.19%**	79.7%	95.02%
Side-notes	94.2%	**98.50%**	64.7%	94.68%
Average	95.8%	**98.84%**	72.2%	94.58%

Table I
SEGMENTATION ACCURACY (F-MEASURE) OF OUR METHOD COMPARED WITH RESULTS OF BUKHARI ET AL. [7], BEFORE AND AFTER THE REFINEMENT STEP.

As can be seen, the suggested method outperforms the approach suggested by Bukhari et al. [7]. The coarse segmentation of the main-text area already provides better results than the coarse segmentation of [7], however, after the global refinement step the performance gap becomes even larger. We examine the contribution of the global refinement framework for each of the 38 images in the dataset and report the results in Figure 5. It is obvious from this figure that the refinement phase has a prominent contribution as for some cases a prefect segmentation is achieved. Notice that sample 22 in Figure 5(a) and 5(b) has better segmentation accuracy before applying the refinement process. This happens due to the presence of components in the main-text area which are directly connected to other components in the side-notes region. These components disguise the minimization process and negatively affect the final segmentation, however, one can notice that this effect is marginal and does not severely affect the segmentation accuracy.

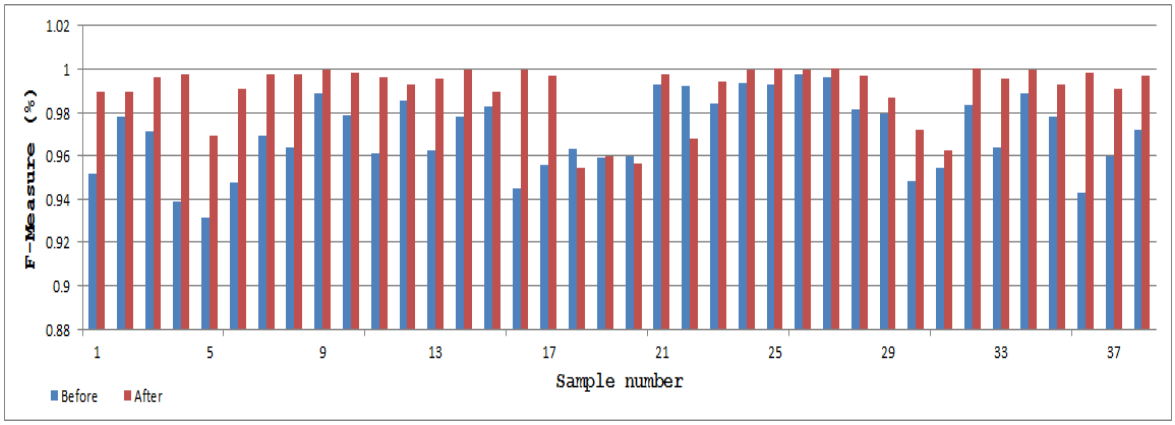
	Before	After
Main-text	97%	**98.98%**
Side-notes	93.1%	**97.75%**
Average	95.05%	**98.36%**

Table II
SEGMENTATION ACCURACY (F-MEASURE) OF OUR METHOD ON THE ENTIRE DATASET BEFORE AND AFTER THE REFINEMENT STEP.

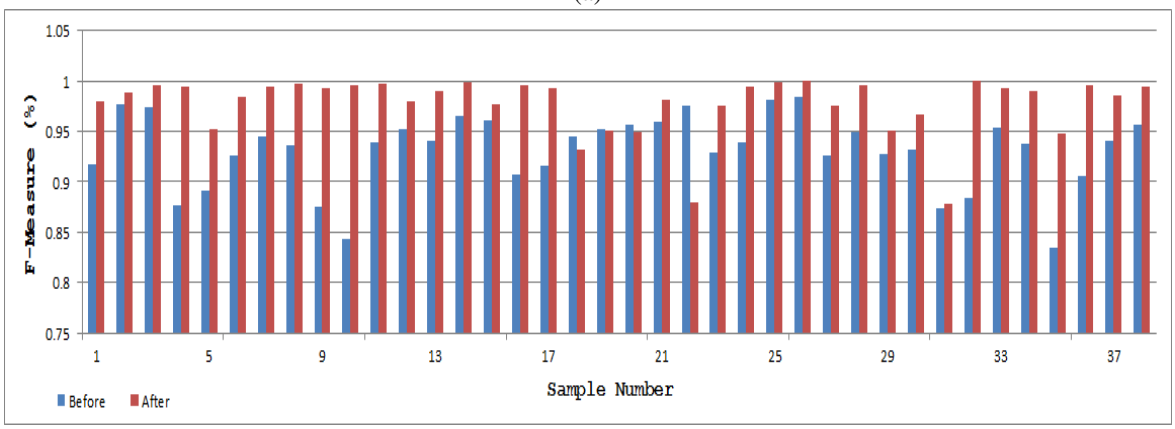
Figure 6 illustrates the coarse segmentation of a sample image from the dataset and its corresponding refinement. As one can notice, the refinement process produces spatially coherent regions and resolves the ambiguity across the borders between the main-text and side-notes regions. Incorrect labels are assigned when there is a significant proximity between a relatively small main-text component and the side-notes region.

(a) (b)

Figure 6. Depicts the (a) coarse segmentation of a sample image from the dataset and (b) its corresponding refinement produced by the energy minimization framework.



(a)



(b)

Figure 5. The contribution of the global refinement framework for the final segmentation accuracy (F-Measure) of (a) main-text and (b) side-notes.

We used MATLAB's built-in profiler to measure the running times of the two stages of the algorithm on an Intel Core 2 Duo running at 3.00GHz using a single core. The coarse segmentation stage, including the hysteresis thresholding, takes $10-90$ seconds per page. The time duration of this stage relies on the image resolution so that the highest time duration was obtained when applying the Gabor filter on high resolution images (2587x3913). Applying the global refinement stage on the entire dataset (38 images) takes 43 minutes, namely, it runs for 1.13 minutes per page on average. However, the work in [7] takes 2 hours to train the model (including the feature extraction phase), and 22 minutes to obtain the final segmentation of a single page. Comparing the time performance between the two methods using the testing set used in [7] yields that our method provides the final segmentation of all the pages in 20.3 minutes, while it took the other method 5.3 hours. We noticed that the profiler of MATLAB adds approximately 10% overhead that we have not substracted. Therefore, the actual times are slightly faster than reported. It is also important to emphasize that a C implementation of the graph-cut based optimization was used.

V. Conclusion and Future Work

We present a coarse-to-fine algorithm for segmenting main-text area in ancient manuscripts images. Exploiting Gabor filter abilities to capture different textures enables determining the approximate location of the main-text area. We generate a coarse binary mask of this area and apply a principled refinement approach by minimizing an explicit energy function. The minimization framework determines the global minimum of the function using graph-cuts. In contrast to previous algorithms, our approach is learning-free and does not include a local refinement step. Our experimental study shows that the suggested approach outperforms the segmentation quality of other segmentation methods. Moreover, a great improvement is achieved in term of time performance with respect to a state-of-the-art method.

Our future work plans include automatic detection of optimal smooth terms of the Gabor filter instead of defining it as a function of the average heights of connected components. We are planning to evaluate the robustness of our approach on a larger dataset. Examining possible directions for segmenting side-notes regions according to their orientation is a demanding challenge as well.

ACKNOWLEDGMENT

This research was supported in part by the DFG-Trilateral grant no. FI 1494/3-2, the Ministry of Science and Technology of Israel, the Council of Higher Education of Israel, the Lynn and William Frankel Center for Computer Sciences and by the Paul Ivanier Center for Robotics and Production Management at Ben-Gurion University, Israel.

REFERENCES

[1] A. Antonacopoulos and A. C. Downton, "Special issue on the analysis of historical documents," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 9, pp. 75–77, 2007.

[2] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher, "Historical document layout analysis competition.," in *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520, IEEE, 2011.

[3] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher, "Icdar 2013 competition on historical newspaper layout analysis (hnla 2013).," in *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1454–1458, IEEE.

[4] "DFG's "Cultural Heritage" programme.." http://www.islamic-manuscripts.net/content/below/index.xml. Online; accessed December, 2012.

[5] D. Comaniciu, P. Meer, and S. Member, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 603–619, 2002.

[6] Y. Wang, I. T. Phillips, and R. M. Haralick, "Document zone content classification and its performance evaluation," *Pattern Recogn.*, vol. 39, pp. 57–73, Jan. 2006.

[7] S. Bukhari, T. Breuel, A. Asi, and J. El-Sana, "Layout analysis for arabic historical document images using machine learning," in *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 639–644, 2012.

[8] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, pp. 167–181, Sept. 2004.

[9] N. Ouwayed and A. Belaïd, "Multi-oriented text line extraction from handwritten arabic documents," in *8th IAPR International Workshop on Document Analysis Systems (DAS)*, 2008.

[10] A. Garz, R. Sablatnig, and M. Diem, "Layout analysis for historical manuscripts using sift features," *International Conference on Document Analysis and Recognition*, vol. 0, pp. 508–512, 2011.

[11] R. Cohen, A. Asi, K. Kedem, J. El-Sana, and I. Dinstein, "Robust text and drawing segmentation algorithm for historical documents," in *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing (HIP)*, HIP '13, (New York, NY, USA), pp. 110–117, ACM, 2013.

[12] S. Mao, A. Rosenfeld, and T. Kanungo, "Document structure analysis algorithms: A literature survey," in *Proc. SPIE Electronic Imaging 5010*, p. 197207, 2003.

[13] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[14] O. Veksler, *Efficient Graph-Based Energy Minimization Methods in Computer Vision*. PhD thesis, Cornell University, Aug. 1999.

[15] I. Fogel and D. Sagi, "Gabor filters as texture discriminator," *Biological Cybernetics*, vol. 61, pp. 103–113, June 1989.

[16] F. Bianconi and A. Fernandez, "Evaluation of the effects of gabor filter parameters on texture classification," *Pattern Recognition*, vol. 40, no. 12, pp. 3325 – 3335, 2007.

[17] M. Kubovy and M. van den Berg, "The whole is equal to the sum of its parts: A probabilistic model of grouping by proximity and similarity in regular patterns," *Psychological review*, vol. 115, no. 1, pp. 131–154, 2008.

[18] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics (TOG)*, vol. 23, pp. 309–314, ACM, 2004.