Visual Detection of Business Cards: Segmentation

Domen Tabernik
Faculty of Computer and Information Science, Ljubljana
January 2015

***Abstract*—This study explores Graph Cut method for the segmentation of business cards captured from a sequence of images. The process of segmentation is needed as a part of an application for detection and recognition of business cards. We explore Graph Cut method in detail, and show it can be applied to business card segmentation. We show how in Graph Cut method foreground and background regions can be successfully initialized using business card key-point detector from our previous study. Furthermore, we show how a sequence of images can be used to improve foreground/background initialization and how by merging multiple frames final segmentation can be improved. We demonstrate our proposed approach on a set of four business cards sequences, using two textured and two non-textured business cards.**

I. Introduction

This report is part of a study to build an application for detection and recognition of business cards. The main purpose of the application is digital storage of business cards. The application is designed such that it requires the user to show and wave a business card in front of a camera. The application will proceed by analyzing the video to: (i) localize the card, (ii) segment it out, and (iii) match it with the digital signature of existing card or store it to the system if not seen before.

In one of the previous study [12], we evaluated key-point detector suitable for business card detection, and in [11], we developed a filtering of the non-stable key-points laying outside of the business card by finding the dominant planar region over a sequence of images using the RANSAC [6] method. Having designed the system to obtain a set of key-points contained on the business card, we continue this series of the studies with the segmentation of the business card. In this study we explore the use of Graph Cut method [7] to perform the segmentation. We show how to utilize key-points on the business card for the successful initialization of the background and foreground needed by the Graph Cut method. Furthermore, we show that having business card visible from multiple frames can be utilized to improve the final segmentation. We utilize it in two ways. Key-points from multiple frames are projected onto each frame to increase the density of the key-points and improve the initialization of the foreground, and segmentation from multiple frames is merged to increase the robustness around poorly visible edges of the business card. We evaluate the proposed methods on a sequence of four different business card, two textured and two non-textured, and show our approach obtains a suitable segmentation for the application of business cards detection and recognition.

The remaining of this paper is organized as follows. In Section II the related work regarding the segmentation is explored. In Section III Graph Cut method is detailed, while our approach to business card segmentation is proposed in Section IV and evaluated in Section V. Concluding remarks are presented in Section VI.

II. Image segmentation

The problem of image segmentation can be addressed with several different approaches. One of the simplest and earliest approaches is gray-level thresholding. In [10], Otsu presented segmentation as reduction of gray-scale image to binary image. It assumes intensity histogram will have two peeks, one corresponding to foreground and one to background pixels. By finding the optimal separation in histogram an image can be separated into two regions, representing a final binary segmentation. Another gray-level thresholding method is watershed algorithm proposed by Beucher and Lantuejoul [1] and later improved by Meyer [2]. In watershed algorithm gray-scale image is thresholded by intensity values. Increasing thresholds can be considered as slowly flooding the relief from the minimal values. Relief is represented as image gradients, thus edges form a natural barriers between different basins. By increasing the water level different basins will start to merge. This is prevented by adding artificial barriers between basins when they start to merge, and a set of all added artificial barriers represents the final segmentation.

Another type of methods rely on analyzing feature space. In [4], Comaniciu and Mee use Mean Shift algorithm to find clusters in colors space, which they then use for the segmentation. This approach can be considered a generalization of Otsu's segmentation, as Mean Shift uses centroids of clusters over the feature space to separated images into different segments. In Otsu's segmentation intensity values are feature space and only two clusters are used. However, the Mean Shift algorithm searches for multiple clusters and works iteratively by using K-Means clustering to find the set of clusters in n-dimensional feature space.

Graphical models have also proven successful tool for segmentation. Greig et al. [7] proposed Graph Cut method in which image pixels are organized as a graph with pixels as vertices and their neighbors representing the edges between them. By adding a sink and a source they show that optimal segmentation can be defined as a minimal cut that maximizes the flow from source to sink in the graph. Further details of this algorithm are presented in the next section. Another graph based model has been presented by Felzenszwalb and Huttenlocher [5]. They model the image pixels with a graph as well, but enable edges to be dynamically changed based on degree of variability in neighboring regions of the image. They incorporate non-local information of regions by measuring intensity differences across the boundary and comparing them to intensity differences between neighboring pixels of each region. They implement this model with a greedy algorithm but show to correctly capture the global boundary information.

Segmentation with active contour model has been introduced by Kass et al. [8]. This approach can be considered as an interactive one, as initial contour needs to be provided. In active contour the contour is modeled as deformable spline, and by minimizing the energy of the spline a correct segmentation can be found. The energy is composed from internal elastic energy and external edge-based energy. The former controls the deformation of the contour and ensures its smoothness, while the later uses image intensities. Specifically, gradients and intensity values are used for external edge-based energy. Formulation of the energy allows optimization with the gradient descend methods.

III. Segmentation with Graph Cut

Segmentation with the Graph Cut has been introduced by Greig et al. in [7], where they proposed to use maximum a posteriori (MAP) estimate to find the segmentation of a simple, but corrupted, binary image. An image segmentation problem can be defined as classification of each pixel y_i into two classes, background, $x_i = 0$ and foreground, $x_i = 1$:

$$f : (y_1, ..., y_N) \to (x_1, ..., x_N).$$

Considering pixels $(y_1, ..., y_N)$ are conditionally independent given x, we can write likelihood function for x independently for each pixel value y_i. We can also model prior distribution as a pairwise interaction between each two neighboring pixels. Greig et al. [7] proposed a pairwise interaction of Markov Random Field where a difference between two neighbors is considered, but only if they belong to different class x. Conditional likelihood of x given y can be now written as:

$$L(x|y) = \sum_{i=1}^{N} \lambda_i x_i + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \beta_{ij} \{x_i x_j + (1 - x_i)(1 - x_j)\},$$

where λ_i is log-likelihood ratio at pixel i, and β_{ij} is zero for any non-neighboring i and j, non-negative for $x_i = x_j$ and non-positive for $x_i \neq x_j$. The image $\tilde{x}$ that maximizes $L(x|y)$ is MAP estimate, and Greig et al. showed this can be computed efficiently by interpreting the problem as graphical model. The problem is formulated as graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where vertices $\mathcal{V}$ are pixels and edges $\mathcal{E}$ represent connections between pixels that are dependent on each other, in this case between neighboring pixels. Prior distribution term from $L(x|y)$ is assigned as the cost of edge between the neighboring pixels, i.e. cost of edge $\mathcal{E}(i, j) = \beta_{ij}$. However, to assign the likelihood term $\lambda_i x_i$ an additional vertices are added, a sink t, and a source s. Edges between source s and pixel i are set as $\mathcal{E}(s, i) = \lambda_i$, if $\lambda_i > 0$; otherwise, if $\lambda_i < 0$, edges between sink t and pixel i are set as $\mathcal{E}(t, i) = -\lambda_i$. In this graph maximum a posteriori estimate is now equivalent to minimum cut $\mathcal{C}(x)$ of graph $\mathcal{G}$. Such cut $\mathcal{C}(x)$ maximizes the flow through the network from the source to the sink and can be efficiency computed with Ford-Fulkerson algorithm.

Binary segmentation can now be considered in an abstract way as an optimization of an energy cost function $E(x)$ composed from two terms:

$$E(x) = \lambda R(x) + B(x),$$

where $R(x)$ captures the cost of assigning each pixel y_i to specific class x_i, and $B(x)$ captures the cost of assigning an opposite class to each neighbor of a pixel y_i. For both terms multiple methods of computation exist. For $R(x)$ the cost needs to capture a similarity of the pixel to either background or foreground pixels. A common technique is to compute a histogram of pixel intensities for both background $\mathcal{H}(y|x = 0)$ and foreground $\mathcal{H}(y|x = 1)$, and then use negative log-likelihood of histogram at specific intensity. $R(x)$ can now be defined as:

$$R(x) = \sum_{i=1}^{N} \left[-ln\mathcal{H}(y_i|x_i = 0) - ln\mathcal{H}(y_i|x_i = 1) \right].$$

Other definitions of $R(x)$ can also be used, such as modeling the distribution of background and foreground pixels with the Gaussian Mixture Model. This can also be done in multiple dimensions so color space can be used. Alternatively, pixel similarity can be computed with a texton that also incorporates local information [9]. The second term, prior distribution $B(x)$, needs to capture the cost of adding a neighboring pixel to foreground or background class. A commonly used neighbor system is 4 way connectivity, but 8 way connectivity can also be used to produce segmentation with a more smooth and round contours. Difference in intensity between neighbors are normally used for the cost of $B(x)$:

$$B(x) = \sum_{i,j} \exp(-\frac{(y_i - y_j)^2}{2\sigma^2}),$$

but a more complex, Laplacian zero-crossing, gradient direction or color mixture model, can also be employed.

IV. Business card segmentation

We employ a Graph Cut method similar to [3] for the segmentation of business card. For the cost of assigning specific class to each pixel, $R(x)$, we use a negative log-likelihood of intensity histogram, while for neighboring cost, $B(x)$, we use difference of pixel intensities in a 4 way neighboring connectivity system. Both exact terms, as used in our implementation, are described in previous section. For computing intensity histogram used in $R(x)$ a gray-scale image is used, and pixels belonging to foreground are sampled for $\mathcal{H}(y|x = 1)$, while pixels belonging to background are sampled for $\mathcal{H}(y|x = 0)$.

Next, we detail how foreground and background regions are defined from key-points, and then detail how multiple frames are merged to improve the final segmentation.

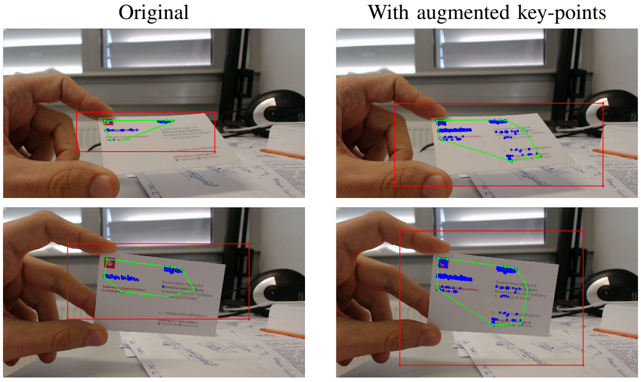

Figure 1. Examples of two frames with poor key-points density leading to poor background initialization (red line) in the left. However, as can be seen with the images on the right by projecting key-points from other frames we more evenly cover the whole area of the business card and improve background initialization.

A. Foreground and background initialization

In Graph Cut methods initialization of foreground and background regions can be accomplished with the modification of the $R(x)$ term. For pixels with high probability of belonging to the foreground the cost of assigning them to the foreground must be set to infinity and the cost of assigning them to the background must be set to zero. In the graphical model this is implemented by setting the costs, i.e. weights on edges, between foreground pixels and the source to infinity, and costs between background pixels and the sink to zero. Having zero cost between background pixels and the sink will ensure that those edges will be included in the minimal cut $\mathcal{C}(x)$, and will therefore never be assigned to the background. While having high cost between foreground pixels and the source will ensure that those edges are assigned to the source. The same concept applies for the pixels with high probability of belonging to the background, except that zero cost is assigned between edges of the source, and infinity cost between edges of the sink. Note that, we do not use infinity but instead, similar to [3], use the highest value between two neighboring pixels in prior distribution term $B(x)$ incremented by one.

In our case the initialization of foreground and background can be performed with the key-points detected on business card. In our previous study we have shown how key-point detection performs on business card [12], and how key-points with a high probability of occurring on business card can be selected [11]. We can now utilize those key-points for the initialization of foreground and background. For the foreground region we use a convex hull around the detected key-points. We avoid using outer bounding box since it will include background region when business card is rotated, while using inner bounding box would reduce to a too small region. We also did not shrink the size of foreground region as the key-point detector was reliable enough, and did not return any key-points outside of the business card. Nevertheless, if this would pose a problem, we could address it by simply shrinking the region to 90-80% of the original size, or alternatively, by fitting a 2D Gaussian to key-point locations and multiplying cost weights in $R(x)$ with it.

For the background region we use 2-times scaled bounding box around key-points, and define locations outside of that region as background region. Note, bigger scaling is needed if density key-points is not improved with multiple frames as shown in the next section. In some frames density of key-points was low and concentrated only over specific part of the business card. Using 2-times scaling was not sufficient in that case to avoid including business card regions to the background, and higher scaling of around 4-3x was needed. However, improving density of key-points as described in the next section allowed to reduce the scaling factor to 2x.

B. Improving segmentation with multiple frames

Segmentation as described in previous sections was performed on each frame independently. However, our application captures several frames of business card being visible from different orientations. In our previous study [11] we utilized this to select only key-points on the business card, and during this process obtained a set of transformations describing the motion of business card from one frame to another. We now utilize this information in two ways to further improve the segmentation.

Increasing key-points density: As described in previous section, some frames may contain only a few key-points and they may not cover the whole business card. In certain frames one area of the business card is more visible and thus more key-points detected there, while in other frames key-points are detected on other areas of the business card. To ensure key-points will cover the whole area of the business card we now utilize provided transformations of business card motion, and merge key-points from different areas of the business card detected in different frames. We project key-points from each frame into the first image, and them from there all collected key-points are projected back to each frame. This ensures each frame has a dense set of key-points covering the business card more evenly. Segmentation can now be initialized with a more dense set of key-points. This now allows to use a smaller scaling factor when initializing background region. Furthermore, by reducing this scaling factor background pixels used to calculate histogram can samples from a bigger set of pixels thus improving the cost in $R(x)$. Two examples where background initialization has been improved are depicted in Figure 1.

Merged segmentation from multiple frames: Similarly to how we projected key-points into the first frame, we can now project segmentation from all frames into a single image. As certain areas of business card are better segmented in one frame, and other areas in other frames, combining them together improves the overall segmentation of the business card. Segmentation merging also increases robustness to background clutter since certain edges around the business card may not be visible enough when background objects have similar color as the business card. In such cases segmentation may spill into the background, but if multiple segmentations are merged this does not pose a problem unless segmentation spill would be consistently occurring in the same area.

The method to obtain final segmentation of the business card can now be described with the following steps. First,

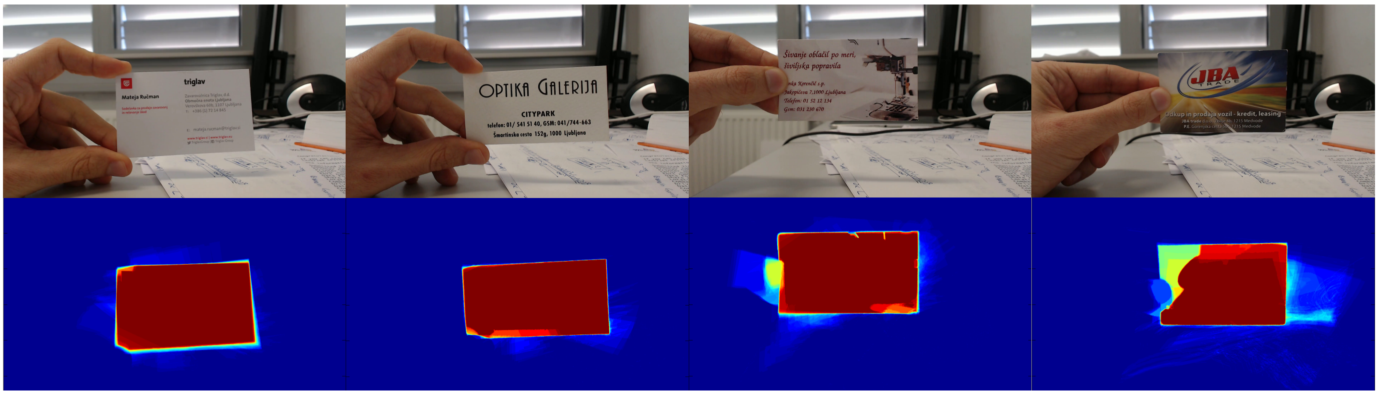

Figure 2. Example of four samples, from left column to right column: two non-textured P1 and P2, and two textured T1 and T2. Top row depicts first frame, while bottom row depicts segmentation probability map merged from summing segmentation of several frames.

segmentation with Graph Cut is performed on each frame as described in previous section. We use key-points from other frames to improve foreground/background initialization. Additionally, we perform connected component analysis on the outputted segmentation and discard small segments with less then 10 pixels. Then each segmentation is projected back to the first image and multiple segmentations are summed into one. By normalizing the map to max value of 1 we produce a final segmentation. Final segmentation is not binary mask but can be considered as a probability map, and for each pixel we can now say with a certain probability if it belongs to the business card or a background.

V. Evaluation

We evaluate the method presented in Section IV on a set of four business card sequences, two textured (referred to as P1 and P2) and two non-textured (referred to as T1 and T2). Evaluation is focused on segmentation and measures the pixel-wise accuracy as a ratio between area of both regions and the overall area of the regions:

$$acc = \frac{|G \cap P|}{|G \cup P|},$$

where G is groundtruth region, P is segmentation and $|\cdot|$ is a sum of all elements of the binary mask. We annotated first frame in each sequence to obtain an initial groundtruth and then projected it to each next frame to obtain groundtruth for all remaining frames.

	$s = 0.5$				
$\lambda\backslash\sigma$	1	5	10	50	100
10^0	0.61	0.62	0.62	0.62	0.62
10^{-1}	0.63	0.65	0.65	0.66	0.67
10^{-2}	0.71	0.74	0.75	0.76	0.75
10^{-3}	0.85	***0.91***	***0.91***	0.89	0.86
10^{-4}	0.86	0.88	0.88	0.87	0.85
10^{-5}	0.86	0.87	0.87	0.87	0.85
	$s = 0.25$				
$\lambda\backslash\sigma$	1	5	10	50	100
10^0	0.61	0.62	0.62	0.64	0.64
10^{-1}	0.64	0.67	0.68	0.69	0.70
10^{-2}	0.73	0.80	0.82	0.81	0.79
10^{-3}	0.84	0.87	0.88	0.88	0.86
10^{-4}	0.85	0.87	0.84	0.86	0.85
10^{-5}	0.85	0.85	0.83	0.86	0.85

Table I
Grid search results on Graph Cut hyper-parameters for business card segmentation. Evaluation was perform using five frames from each business card samples, therefore, calculating 20 segmentations for each combination. The reported results are segmentation accuracy averaged over 20 images. The highest accuracy was obtained with $\lambda = 10^{-3}$, $\sigma = 10$ and $s = 0.5$.

A. Graph Cut parameters

Two parameters are needed to be set for Graph Cut method, λ and σ. The first parameter, λ, defines the ratio between the influence of $R(x)$ and $B(x)$. The higher it is the more influence $R(x)$ has. Therefore, high λ will favor segmentation as defined by the initialization and by similarity measure of a pixel to foreground and background regions, while having small λ will favor the neighborhood boundary. In our case we found a suitable range values between 1 and 10^{-5}.

The second parameter, σ, is related to the difference between two neighboring pixels in $B(x)$. As stated in [3], this parameter can be estimated as a camera noise, as it will penalize for discontinuities between pixels of similar intensity when $|y_i - y_j| < \sigma$, but for pixels $|y_i - y_j| > \sigma$ penalty will be small. Having small σ will produce noisy segmentations with many discontinuities, while high σ will produce clear, long segments, but may miss small details. In our case the intensities y_i are between 0 and 255, and we find σ to work well between for up to 100.

We additionally added scaling parameter as we do not perform segmentation on original image, but instead scale down the image by at least a half. This has proven necessary as higher resolution images produce too big graphs that took too long to compute. Since original resolution of our samples is 1920x1080 reducing it by half does not increase the quality that much.

We get the most optimal combination of parameters by performing a grid search and varying all three parameters. We tested six values for $\lambda = [10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$, five for $\sigma = [1, 5, 10, 50, 100]$ and two scales $s = [0.25, 0.5]$. We evaluated parameters by segmenting a single frame and

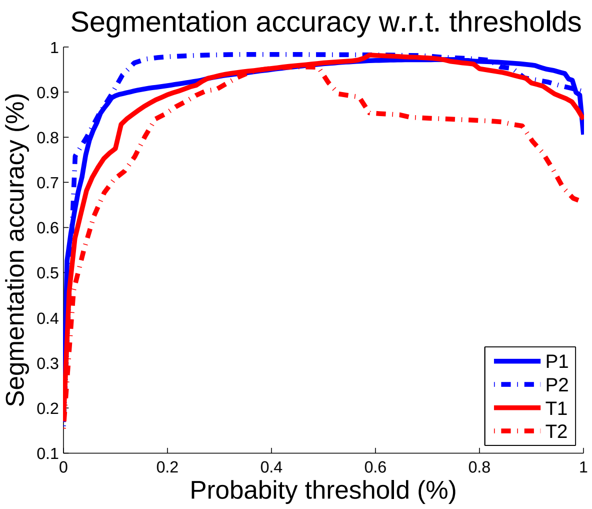

Figure 3. Final segmentation accuracy for four different business cards when probability map is thresholded at different levels.

information for multiple frames was used only for increasing key-point density. To asses each combination of parameters all four business cards were tested; however, only five frames from each sample were taken for this evaluation, thus using 20 images to evaluate each parameter. Selected five frames were evenly distributed over the sequence. Results are reported in Table I, which contain segmentation accuracy averaged over 20 images for each parameter combination. The highest score was obtained with $\lambda = 10^{-3}$, $\sigma = 10$ and $s = 0.5$; however, using $\sigma = 5$ could also be acceptable as it produces almost the same result.

B. *Segmentation merging*

The final segmentation of four business cards is shown in Figure 2. Figure depict a probability map below each business card composed by summing segmentation from multiple frames. Note that, perfect segmentation cannot be obtained due to fingers occluding the card, and in some cases edges are not necessary sharp since homography used to project the segmentation to the first frame may not have been perfect in all samples.

Obtaining a binary segmentation map from probability map requires a thresholding at specific probability level where pixels with probability of lower then the threshold are assigned to the background. In Figure 3 we plot segmentation accuracy of four samples thresholded at different levels. Highest probability for all four samples can be achieved by thresholding at between 0.3 and 0.5, where segmentation accuracy is between 0.95 to 0.98 for all four samples. However, we can notice that using higher probability for a threshold achieves even better results for P1, P2 and T1, but for sample T2 accuracy significantly drops. As can be seen from the probability map for T2, left-top area of the business card was consistently missing from the segmentation. A higher brightness of this specific area can explain poor segmentation. Compared to other areas of the business card which had many dark colors, this section contained only bright, white colors and may not have been properly captured by the intensity histogram for the foreground.

VI. Conclusion

In this paper we presented a segmentation method suitable for the proposed application of business card detection and recognition. We presented an overview of the Graph Cut [7] method, and showed how to utilize it for the segmentation of business cards. We used intensity histogram sampled from initial foreground/background region to capture the cost of assigning pixels to either foreground and background, while we defined prior distribution as difference in pixel intensity in a 4-way connected neighborhood. We showed that business card key-points returned by our detector from our previous study [12] is suitable for the initialization of the foreground and background region. Using convex hull around detected key-points has proven appropriate for foreground initialization, while 2-times scaled bounding box over the same key-points is appropriate for background initialization. We further searched for the most optimal Graph Cut hyper-parameters with a grid search, and have shown to achieve best results by reducing the image resolution in half and using $\lambda = 10^{-3}$ and $\sigma = 10$. Furthermore, we showed that homography describing the motion of business cards can be successfully utilized in two ways to improve the final segmentation. Firstly, the number of key-points per image can be increased by projecting key-points from other images to each frame and improving the background initialization. Secondly, segmentation of each frame can be projected into the first frame, and we have shown that by summing segmentation from multiple frames into one probability map we can improve segmentation, and achieve accuracy of between 0.95 and 0.98 for two textured and two non-textured samples.

Segmentation could also be further improved by exploiting the geometry of the business card. Since most business cards have rectangular shape Hough transform could be uses to fit several lines and find such combination of four lines that satisfies the geometry of the business card and at the same time maximizes the covered area.

References

[1] S. Beucher and C. Lantuejoul. Use of Watersheds in Contour Detection, 1979.

[2] S. Beucher and F. Meyer. The morphological approach to segmentation: the watershed transformation. *Optical Engineering-New York-Marcel Dekker Incorporated-*, 34:433–433, 1992.

[3] Y.Y. Boykov and M.-P. Jolly. Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, 1(July):105–112, 2001.

[4] Dorin Comaniciu, Peter Meer, and Senior Member. Mean Shift: A Robust Approach Toward Feature Space Analysis. 24(5):603–619, 2002.

[5] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, 59(2):167–181, sep 2004.

[6] Martin a Fischler and Robert C Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Apphcatlons to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[7] Dorothy M. Greig, Bruce T. Porteous, and Allan H. Seheult. Exact Maximum A Posteriori Estimation for Binary Images. *Journal of the Royal Statistical Society. Series B*, 51(2):271–279, 1989.

[8] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models, 1988.

[9] Jitendra Malik, Serge Belongie, Thomas K. Leung, and Jianbo Shi. Contour and Texture Analysis for Image Segmentation. *International Journal of Computer Vision*, 43(1):7–27, 2001.

[10] Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[11] Domen Tabernik. Visual Detection of Business Cards: Key-Point Correspondences Filtering. Technical report, 2015.

[12] Domen Tabernik. Visual detection of business cards: study of interest key-point detectors. Technical report, 2015.