A Semi-Automatic Groundtruthing Tool for Mobile-Captured Document Segmentation

Joseph Chazalon*, Marçal Rusiñol†*, Jean-Marc Ogier* and Josep Lladós†
*L3i Laboratory, Université de La Rochelle
Avenue Michel Crépeau
17042 La Rochelle Cédex 1, France
†Computer Vision Center, Dept. Ciències de la Computació
Edifici O, Univ. Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain

***Abstract*—This paper presents a novel way to generate ground-truth data for the evaluation of mobile document capture systems, focusing on the first stage of the image processing pipeline involved: document object detection and segmentation in low-quality preview frames. We introduce and describe a simple, robust and fast technique based on color markers which enables a semi-automated annotation of page corners. We also detail a technique for marker removal. Methods and tools presented in the paper were successfully used to annotate, in few hours, 24889 frames in 150 video files for the smartDOC competition at ICDAR 2015.**

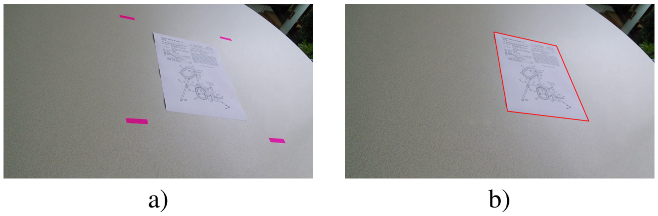

Fig. 1. Our annotation method allows for fast and reliable generation of segmentation ground-truth in videos based on color markers. a) Sample input frame. b) Sample output frame: markers are removed and segmentation is represented as a colored quadrilateral.

I. Introduction

In the past few years, the smartphone market expanded very quickly: according to the International Data Corporation [1], vendors sold a total of 1,004.2 million smartphones worldwide during the year 2013, which accounted for 55.1% of all mobile phone shipments. As a consequence of users having at nearly any moment of their life a powerful and handy device capable of taking increasingly better pictures triggered a new (and somehow unexpected) use: smartphones are progressively replacing desktop scanners, for both personal and professional use. This emerging trend is a source of new challenges for the document analysis and recognition community.

As a *camera-based* acquisition process, *mobile document capture* entails challenges which were already identified and studied by Liang et al. [2], but the most frequent or though ones may not be the same as *camera-based document capture*. Given the final goal of producing a document image suitable for many later uses (human readable, noise-free, OCR-able, indexable, etc.) like a scanned document page would be, we found that the most important tasks are, in decreasing order: background removal, perspective correction, lighting normalization, focus and motion blur avoidance. On the other hand, our recent experiments learned us that low resolution, non-planer surfaces, lens distortions, intensity and color quantization, sensor noise and compression are not majors issues for most applications. Regarding the need for lightweight algorithms, even if battery saving and real-time processing are mandatory, the processing power of smartphones makes those devices absolutely suitable for fast image processing.

Another important difference between camera-based document capture and *mobile document capture* is due to the intrinsic properties of the *mobility situation*, which induces specific constraints on the capture process. In the case of a receipt, for instance, a customer would rather digitize such document right after it is printed, instead of eventually losing it before checking his or her account balance, or getting refunded for a business expense. Therefore, we believe that mobile document capture solution should:

- be fast and simple to use;
- ensure that the quality of produced images are suitable for later processing: human reading, OCR, classification, etc.

We also believe that such features can only be obtained by assisting the user during the capture process, while he or she is pointing at the document to digitize. Ideally, the capture should be triggered at the best moment, as soon as the document is detected.

While working on a method for detecting and segmenting a page object within preview frames acquired during the capture process, we realized that evaluating such kind of approaches required a specific dataset, metrics and evaluation tools. Metrics based on overlapping areas are easily implemented, but to generate a large-scale and realistic dataset at minor costs required a new approach.

The contribution of this paper is to propose such approach, capable of generating large datasets of realistic preview videos and the associated segmentation ground-truth for document capture sessions. We introduce and describe a simple yet powerful technique based on color markers which enables a fast and robust annotation of objects corners, as illustrated in Figure 1.

Our method is based on two simple steps after acquiring video samples:

1) a semi-automated page segmentation step based on color markers, along with the removal of such markers;
2) a manual inspection and correction step to eliminate small errors and ensure a high quality dataset.

The method described in this paper was successfully implemented and used to annotate, in few hours, 24889 frames in 150 video files for the challenge 1 ("page object detection and segmentation in preview frames") of the smartDOC competition at ICDAR 2015. This method exhibited 3 main advantages:

1) it is very simple to set up and use, both for the semi-automated step and the inspection step;
2) it is inexpensive, both in terms of required material and manual workload;
3) it is robust to many distortions and is therefore suitable for the generation of realistic data.

This paper is organized as follows: Section II reviews similar datasets and approaches; Section III introduces the semi-automatic ground-truth generation method; Section IV details how color markers are removed from source videos; Section V discusses user experience and annotation costs; Section VI summarizes our contributions.

II. Related Work

Several datasets related to mobile document capture have been proposed, but none of such approaches can be used to generate the ground-truth for the preview videos we introduced in the previous section.

The first dataset, to our knowledge, of camera-captured document images is the "DFKI-1 dataset", proposed by Shafait et al. [3] and its evolution, the "IUPR dataset" [4]. Each dataset is composed of approximately 100 images, and represent many different documents classes captured in different situations with various distortions: perspective, warping, lighting, curl, skew, etc. The ground truth associated with each image is manually created and contains, along with a scanned version of each page, pixel-level accurate tagging of lines, text zones, and content type. ASCII text is also available. During the dewarping contest at CBDAR 2007 [3], dewarping methods were compared against what a commercial OCR system would produced on dewarped images. Such approach is not viable in our case, as we cannot rely on the presence of a majority of textual content within each document. Furthermore, the manual annotation of each document image is not feasible when dealing with tens of thousands of video frames.

Bukari et al. [5] recently proposed a method which makes use of the pixel-level content of images to compute a matching score using the Earth Movers Distance. While such technique was proved to perform well on scanned images for document classification, its application to evaluate the quality of the dewarping of some camera-captured document image seems difficult. Indeed, sensors noise and perspective distortion produce uneven image qualities after perspective correction, and we fear that it may not be possible to discriminate a well-segmented low-quality frames from badly-segmented high-quality frames because the pixel-level matching measure may confuse them.

Another direction was taken with the approach of Liang et al. [6]: the authors proposed a method for the geometric rectification of camera-captured document images, and used a synthetic dataset to evaluate their system. To avoid pixel-level comparison, they evaluated their method using the OCR accuracy after image rectification. While artificial image generation is attractive, it limits the complexity of the capture conditions, mostly in terms of lighting and device reaction to hand motion with six degrees of freedom.

Finally, one last approach, which specifically targets mobile document capture, was recently proposed by Kumar et al. [7]. To evaluate their system for image quality assessment, the authors created a dataset of images containing various amounts of out-of-focus blur. Using a transparent support, the mobile was kept at a fixed position relative to each document, and variations of the focal length were used to generate the images. While the fixed position of the device could allow for the replication of the page coordinate within each image, such setup does not support enough distortions and motion for our situation.

None of those methods was suitable to generate a dataset of preview videos with realistic content and reasonable annotation time. During the design of a new approach, we decided to avoid this risk and the computational complexity of methods based on pixel-level matching, and resorted to use intersection-based metrics which are content-agnostic. This forced us to make two compromises in order to enable a semi-automated generation of segmentation ground truth:

1) we introduced color markers in the scene to define a referential which permits to recover the page position;
2) we dropped the support for warped documents in our method, as it would have prevented us from using simple quadrilaterals when detecting page content and matching results.

The next sections will demonstrate that, first, such approach produces high quality segmentation with very little manual work, and second that the color markers can generally be safely removed, avoiding any bias in segmentation methods and resulting in a minor perturbation in the end. The acquisition step, which will not be detailed, is very simple: as illustrated in Figure 1, we position markers around the document to capture, and start recording a video while trying to center the page like we would with a scanning application.

III. Semi-automatic Groundtruthing

In order to evaluate the performance of different document capture methods in video streams, we need to generate the ground-truth of our collection consisting of a quadrilateral that defines the borders of the paper sheet in the scene for each frame of the video. Since manually annotating each video frame is a tedious and expensive task, we decided to use a semi-automatic groundtruthing approach. One of the assumptions we made to generate the dataset is that the paper sheet lies on a flat surface and does not move. By placing color markers in the table that can be easily segmented, we could bootstrap the page segmentation process by computing the transformation that exists between the markers position and the paper sheet with a minimum human intervention.

The proposed semi-automatic approach works as follows. First we have to correctly segment the color markers that we see in Figure 2a). Given a video stream composed of n frames $F_0, ..., F_i, ..., F_{n-1}$, we present the first frame F_0 to the user and ask him to click on the four markers. The RGB values of each of the selected points together with a tolerance parameter are then used as thresholds in order to obtain the marker segmentation that we can see in Figure 2b). In order to be tolerant to illumination changes, the RGB values that serve as thresholds are updated iteratively for each frame, i.e. the RGB value of each of the markers centroid at the ith frame is used to segment the $i + 1$th frame.

Let us denote M_i the quadrilateral formed by the four marker centroids $M_i = \{A_i, B_i, C_i, D_i\}$ and P_i the quadrilateral formed by the four page corners $P_i = \{W_i, X_i, Y_i, Z_i\}$ for the ith frame, as we can see in Figure 3. A reference coordinate system described by four points $M' = \{A', B', C', D'\}$ is defined and we can then compute for each frame a perspective transform [8] H_i that transforms the point set M to M' using

$$M' = H_i M_i.$$

When we compute H_0 for the first frame of the video we build the wrapped image F'_0 that is presented to the user for selecting the four corners $P' = \{W', X', Y', Z'\}$ of the page (see Figure 2c)). Since there is no movement between the page and the markers in the real world, ideally the wrapped images F' will all look alike and no matter the camera position and the perspective effect of the frame F_i, the page will always be located at the same P' corners in the reference coordinate system. By backwards projecting the points from P' using the inverse perspective transform H_i^{-1},

$$P_i = H_i^{-1} P',$$

we will find the corners of the page P_i at any frame i.

Using such approach, we were able to annotate the whole dataset by just asking the user eight clicks at the first frame of each video. Four clicks for pointing out the color of the markers and four other clicks for determining the corners of the paper page in the wrapped image.

Finally, we can use the marker segmentation mask in order to "erase" the markers from each frame using an inpainting technique.

IV. Marker Removal

We have used the approach by Telea [9] to remove the markers from the original video frames. Figure 4 illustrates the marker removal workflow: from a given frame containing color markers to be removed (Figure 4a), and the mask of the previously detected color markers (Figure 4b), we produce the final frame, for which the segmentation is known, and the markers are removed by an *"inpainting"* step (Figure 4c).

According to [9], digital inpainting provides a means for reconstruction of small portions of an image. Like most inpainting approaches, Telea's approach aims at progressively propagating color information (values and gradients) toward the center of the regions to reconstruct. The inpainted value of a point p is computed using a weighted sum of the extrapolated intensities from all the points q in the neighborhood of p.

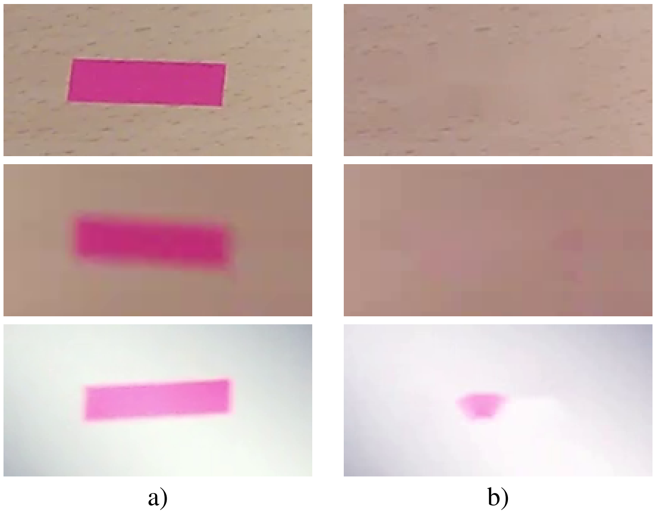

Fig. 5. Sample results for marker removal. a) Original content. b) Inpainted content. First line shows a sharp marker, second line a blurry marker, and third line a marker with highlight on it.

Telea's methods has several advantages which encouraged us to use it over other methods: is it fast enough to be capable of processing video frames in real time on a modern laptop; and due to its simplicity and publicly available code, many implementations are available.

In our case, the application of Telea's method requires a mask of the areas of the image to restore, and the size of the neighborhood to take into consideration for each pixel to restore. For the mask, we used a slightly dilated version of the mask produced by the color detection presented in Section III. It permits to tolerate a certain amount of blur, as well as some of the compression artifacts and noise. Regarding the size of the neighborhood, a few pixels (around 5) are usually sufficient to get results which are visually unnoticeable.

Figure 5 presents some details of results obtained with this marker removal method.

V. User Experience

The proposed method was successfully used to generate the sample and test sets for the Challenge 1 ("page object detection and segmentation in preview frames") of the SmartDOC competition[1] at ICDAR 2015, generating 150 inpainted videos (for 5 different backgrounds) with the associated segmentation ground-truth, giving a total of 24889 frames. Another set of 30 videos was processed, but its edition costs were too important, and we stopped the process to keep only a few valid outputs which were released as a sample of the test set.

In this section, we describe the evaluation of the user workload for each step of the process (semi-automatic groundtruthing, and manual inspection and correction) and give an overall estimation of the time required to generate the full dataset.

[1] https://sites.google.com/site/icdar15smartdoc

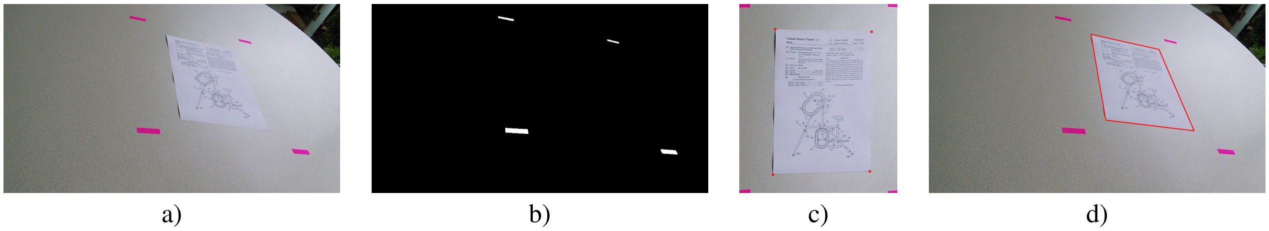

Fig. 2. Semi-automatic groundtruthing approach. a) Original frame. b) Marker segmentation. c) Wrapped frame in which the user marks the four document corners. d) Intermediate result with the document segmentation indicated as a red quadrilateral.

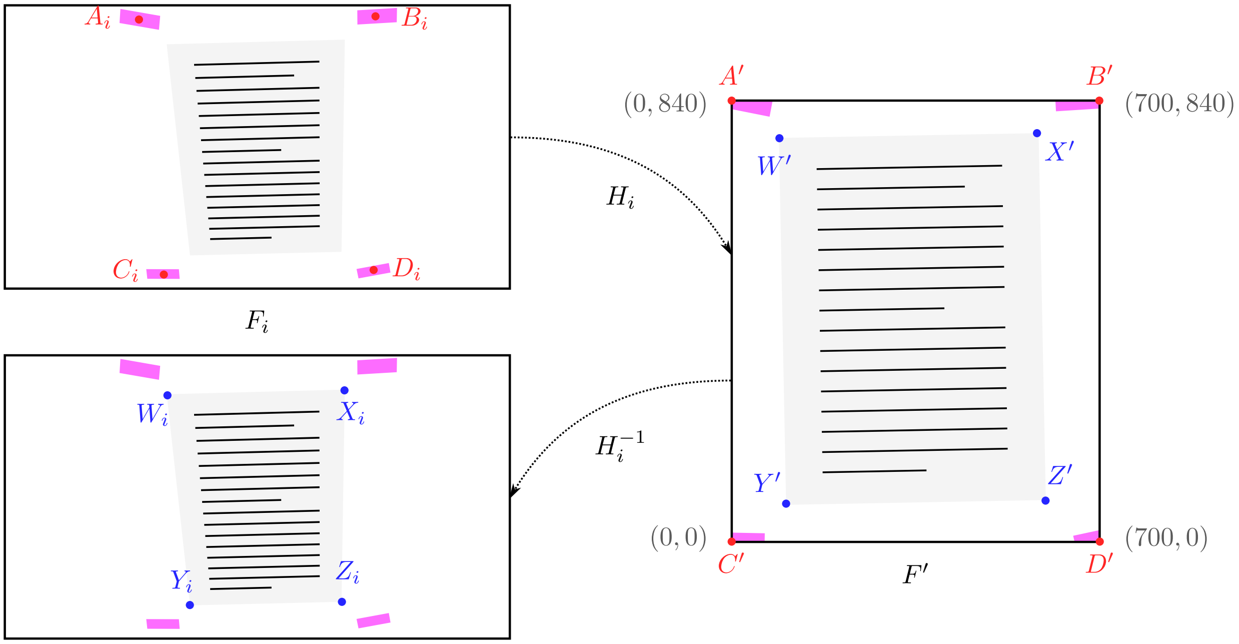

Fig. 3. Markers-to-corners approach. From any frame F_i, marker centroids $\{A_i, B_i, C_i, D_i\}$ are detected, and the forward transformation H_i which maps those points to know coordinates $\{A', B', C', D'\}$ is found. The position of the page corners $\{W', X', Y', Z'\}$ within this referential $F'$are constant, and the inverse transformation of those coordinates with H_i^{-1} gives the real coordinates $\{W_i, X_i, Y_i, Z_i\}$ of the corners in frame F_i.

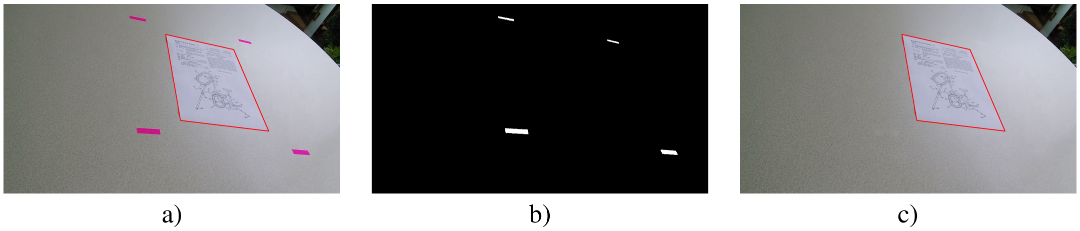

Fig. 4. Semi-automated marker removal. a) Frame with the document segmentation indicated as a red quadrilateral. b) Marker segmentation. c) Final output with the document segmentation and marker inpainting.

A. *Semi-automatic groundtruthing*

This first stage requires from the user, for each video file to process, to select from the first frame of the video (by clicking) the positions of the four markers, as well as the position of the four corners of the page.

During the frame-by-frame automatic processing, when some marker cannot be found in a frame, the system asks for a manual correction and pauses the processing of the current video. The user has to manually select again which are the marker positions. This is used to generate a new color mask and resume the process. We processed videos in parallel and pooled interaction requests at each stage to prevent the user from waiting for work.

Table I summarizes the cost of this first step. First frame actions refer to mandatory coordinate selections in the first frame of each video, and error correction actions refer to coordinate selections in an erroneous frame. It shows that we were able to reduce, at this step, the manual annotation of 24889 frame to 1232 coordinates selections (clicks). However, in order to produce a high quality ground-truth, manual inspection was still required.

TABLE I. MANUAL COST FOR SEMI-AUTOMATIC GROUNDTRUTHING.

Background	Frames	Actions	
		First frame	Error correction
background01	6180	240	0
background02	6011	240	4
background03	5952	240	24
background04	4169	240	0
background05	2577	240	4
TOTAL	24889	1200	32

B. Manual inspection and correction

Manual inspection and correction of the inpainted frames and the associated ground-truth required three kind of actions:

1) navigation actions: keystrokes or mouse wheel for forward and backward frame advance;
2) segmentation correction actions: choose four new coordinates for each frame where the segmentation page corner coordinates were erroneous;
3) inpainting correction actions: select four quadrilaterals to surround marker regions and define explicitly the inpainting mask (resulting in 16 coordinate selections).

Table II summarizes the costs for this second step. While navigation actions exhibit an important amount of user action, they are very fast and simple to perform. Segmentation corrections were simple and non frequent actions (only 21 frames were corrected). Inpainting actions were the costliest, requiring around 1500 clicks for the edition of 94 frames.

TABLE II. MANUAL COST FOR INSPECTION AND CORRECTION.

Background	Actions		
	Navigation	Inpainting	Segmentation
background01	6993	320	0
background02	6254	224	4
background03	6647	320	60
background04	4251	144	0
background05	2918	496	20
TOTAL	27063	1504	84

C. Overall evaluation

The global process of semi-automatic groundtruthing and inpainting, and a subsequent manual inspection and correction step, was really fast. For the first semi-automatic step, half an hour was sufficient to process all videos on a modern laptop with 8 cores and 8 GB of RAM. The second step lasted longer: each group of 30 videos took between 30 and 45 minutes to be reviewed and eventually corrected. It represents roughly 1 minute of manual inspection and edition for video recordings of around 10 seconds each. As it was mentioned in introduction of this section, another set of 30 videos was discarded because of the important processing cost it would have required.

VI. CONCLUSIONS

We presented a novel method based on color markers to generate a realistic dataset of preview frames, along with the associated segmentation ground-truth, suitable for the evaluation of assisting methods in the context of mobile document capture.

While our method does not permit to handle curled pages, it exhibits three main strengths. First, it is very simple to set up and use, both for the semi-automatic and the inspection steps: as soon as processing tools are available, unexperienced users can easily capture and process video samples without any particular knowledge. Second, it is inexpensive, both in terms of required material and manual workload: a smartphone and only a few color markers are required, and the overall edition cost is very reasonable regarding the important amount of frames which are automatically annotated. Third, it is robust to many distortions and is therefore suitable for the generation of realistic data: even in the context of low light or rapid motion, with various backgrounds, this method was capable of localizing the color markers and efficiently segment the page objects. Multiple orientations and size of pages can be captured, and there is no restriction on the content.

This method was successfully used to generate the full dataset for the Challenge 1 (“page object detection and segmentation in preview frames”) of the SmartDOC competition at ICDAR 2015, proving the accuracy and the efficiency of the overall process.

ACKNOWLEDGMENT

This work has been partially supported by the Spanish Ministry of Education and Science under projects TIN2014-52072-P, and TIN2012-37475-C02-02 and by the People Programme (Marie Curie Actions) of the Seventh Framework Programme of the European Union (FP7/2007-2013) under REA grant agreement no. 600388, and by the Agency of Competitiveness for Companies of the Government of Catalonia, ACCIÓ.

REFERENCES

[1] *Worldwide Smartphone Shipments Top One Billion Units for the First Time, According to IDC*, International Data Corporation, Jan. 27, 2014, retrieved on Oct. 30, 2014. [Online]. Available: http://www.idc.com/getdoc.jsp?containerId=prUS24645514

[2] J. Liang, D. Doermann, and H. Li, “Camera-based analysis of text and documents: a survey,” *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 7, no. 2, pp. 84–104, 2005.

[3] F. Shafait and T. M. Breuel, “Document image dewarping contest,” in *2nd Int. Workshop on Camera-Based Document Analysis and Recognition (CBDAR)*, 2007, pp. 181–188.

[4] S. S. Bukhari, F. Shafait, and T. M. Breuel, “The IUPR dataset of camera-captured document images,” in *Camera-Based Document Analysis and Recognition 2011*, ser. Lecture Notes in Computer Science, M. Iwamura and F. Shafait, Eds. Springer Berlin Heidelberg, 2012, no. 7139, pp. 164–171.

[5] S. S. Bukhari, M. Ebbecke, and M. Gillmann, “Business forms classification using earth mover’s distance,” in *11th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 2014, pp. 11–15.

[6] J. Liang, D. DeMenthon, and D. Doermann, “Geometric rectification of camera-captured document images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 30, no. 4, pp. 591–605, 2008.

[7] J. Kumar, P. Ye, and D. Doermann, “A dataset for quality assessment of camera captured document images,” in *International Workshop on Camera-Based Document Analysis and Recognition (CBDAR)*, 2013, pp. 39–44.

[8] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[9] A. Telea, “An image inpainting technique based on the fast marching method,” *Journal of graphics tools*, vol. 9, no. 1, pp. 23–34, 2004.