Real Time Camera Phone Guidance for Compliant Document Image Acquisition Without Sight

Michael P. Cutter and Roberto Manduchi
Computer Engineering Department
University of California at Santa Cruz
Email: {mcutter,manduchi}@soe.ucsc.edu

***Abstract*—Here we present an evaluation of an ideal document acquisition guidance system. Guidance is provided to help someone take a picture of a document capable of Optical Character Recognition (OCR). Our method infers the pose of the camera by detecting a pattern of fiduciary markers on a printed page. The guidance system offers a corrective trajectory based on the current pose, by optimizing the requirements for complete OCR. We evaluate the effectiveness of our software by measuring the quality of the image captured when we vary the experimental setting. After completing a user study with eight participants, we found that our guidance system is effective at helping the user position the phone in such a way that a compliant image is captured. This is based on an evaluation of a one way analysis of variance comparing the percentage of successful trials in each experimental setting. Negative Helmert Contrast is applied in order to tolerate only one ordering of experimental settings: no guidance (control), confirmation, and full guidance with confirmation.**

I. Introduction

Optical Character Recognition (OCR) enables a new frontier of printed document accessibility. Printed documents can be made accessible by applying OCR to an image of the document. Excellent (above 99.9% recognition accuracy) OCR has only been consistently achieved when the documents are in scanned in flatbed planar format. However, camera captured images of documents often suffer from perspective distortion, motion blur, and uneven lighting. Our software supplies real time feedback that helps someone capture an image of a document as if it was scanned.

In order for OCR results to be semantically meaningful the entire composition of the document must be visible in the viewing frustum of the camera. Further to maximize OCR quality resolution should be maximized. Perspective distortion should also be minimized as the greater the viewing angle changes from planar the minimum resolvable distance shrinks. Therefore, meaningful OCR results are more likely when the entire document fills the majority of the image plane and when the image is taken from overhead.

Our research addresses how best to guide a blind or visually impaired photographer to take an image where most the entire document is visible and recognizable. We aim to evaluate if an automatic guidance system is necessary, and whether such a system will be able to ensure that images taken by visually impaired photographers meet these requirements for complete and meaningful OCR results.

In Section II we address some of the related projects in this area. Then in Section III we derive the theoretical aspects of our model. Next we discuss the design of the experiment and the metrics in Section IV and in Section V we address our hypotheses with an evaluation of the user study. Finally, we motivate future work in the final Section VI.

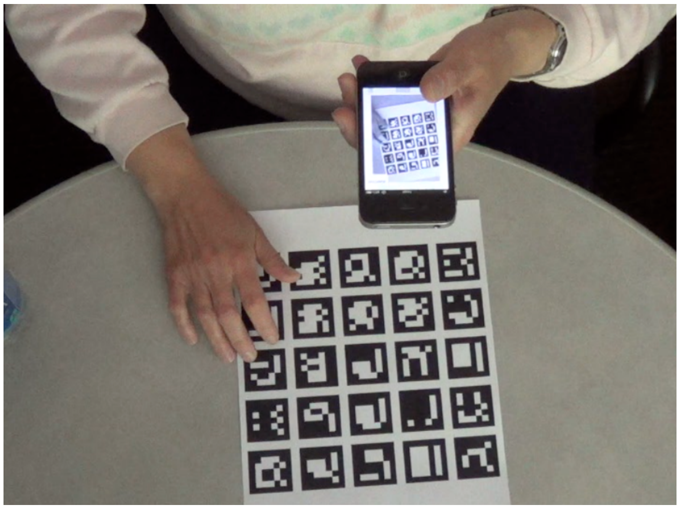

Fig. 1: A participant positioning the iPhone over the document printed with the ArUco fiducials.

II. Related Work

Several assistive technology mobile OCR applications exist [1], [2], [3], [4], [5], [6]. The KNFB Mobile reader [2] was perhaps the first powerful mobile OCR solution for the blind and visually impaired. It has the capabilities to read printed text documents. Zandifar et al. [3] created an end to end head mounted video based text recognition system for the visually impaired. A recent iOS based assistive OCR technology called Text-Detective [4] is available from Blindsight. Although, all of these applications address OCR none offer a real time feedback for document image acquisition assistance.

Recent work by Jayant et al. [7] includes the study and development of a mobile application, EasySnap, for assisted photography. EasySnap provides assisted photography feedback. They surveyed a large number of people and asked what their desired use of such a system would be. The largest desired use was text processing and OCR. Another point mentioned is that a significant number of their users had problems understanding the effect of distance and perspective in photography. A closely related assisted photography application was discussed by

Vázquez and Steinfeld [8]. They developed an application to help people take images of transit obstructions by spoken and auditory feedback based instructions in such a way that salient objects are in the center of the image. Both of these applications were successful and accepted by the users in the respective study.

III. Method

We have conducted an experiment to assess the potential of a document image acquisition guidance. Our contribution is a principled approach that maximizes the chance of a document being recognized by OCR. This method requires that the position of the camera of the phone be known with respect to the document. We determine pose optically for this experiment by placing fiducial marker on a document. We used an open source augmented reality package, ArUco [9], for fiducial marker generation and detection. ArUco is implemented in the OpenCV library [10].

Clearly, this an idealized case and a regular document does not contain fiducial markers that can be used for pose estimation. However, we note that previous studies have already considered adding tags to documents for augmented reality purposes such as in work by Guimbretière [11], Paper Augmented Digital Documents and therefore it is conceivable that are system could be employed in tagged documents. Other work by Nakai et al. [12], [13] address locating the position of a camera with 'locally likely arrangement hashing' [14] affine invariant features by posing the task as a retrieval problem. Liang et al. [15] work, Camera-Based Document Image Mosaicing, could also play a role in document image acquisition assistance.

Once the fiducial markers are detected, we know the correspondence between the 2D location of the marker on the image plane and the 3D-position of the markers on the document. Then we can use these correspondence to solve for the pose of the camera with respect to the page. Generally the problem of solving for pose given 2D image points to 3D world points on a plane (the document image) is known as the perspective-n-point problem [16]. An efficient non iterative solution, Efficient Perspective-n-Point Camera Pose Estimation [17], to the perspective-n-point is used for this purpose.

A. Compliant Space

Our real time guidance algorithm verbalizes instructions as a function of the current pose to guide the user to acquire an image of a document that can be completely read by OCR. We define this type of image as a compliant image. There exist a set of possible phone positions from where compliant images of the document can be acquired. We call this the Compliant Space, a bounded tetrahedron in 3D world coordinates.

The Compliant Space is formally defined as a set of poses where compliant images can be captured. A compliant image must satisfy the following conditions (1) when all four corners of the page are visible and (2) a small letter printed anywhere on the page maps to a sufficient number of pixels for accurate OCR. An example of Compliant Space with these assumptions is visible in Figure 1. The mathematical definition of these two requirements (1) in viewing frustum and (2) minimum reading distance are discussed next. Note that capital letters are used to denote 3D coordinates and matrices and lowercase letters are for 2D coordinates. Vectors and matrices are bold-faced. The four corners of the page are each denoted $\mathbf{\Pi}_i$ and π_i, in the world and image plane respectively. In the following section we will explain how the Compliant Space is derived.

1) In Viewing Frustum: The camera's viewing frustum is a function of its focal length, principal point, and pose, $\mathbf{P}$. Pose can be decomposed into a translation, $\mathbf{T}$, and a rotation matrix, $\mathbf{R}$, from the page to the camera. We assume the camera is calibrated [18] its unique intrinsic matrix $\mathbf{K}$ is therefore known. We assume no radial distortion. We set the reference frame centered in the middle of the document. All points on document are on the z plane origin. We can calculate the viewing frustum in world coordinates relative to the document by back-projecting the image plane corner points to 3D rays.

First we can solve for the camera location in world coordinates $\mathbf{C}_w = -\mathbf{R}^{-1} \cdot \mathbf{T}$ and for pixel $\mathbf{p}$ homogeneous (x,y,1) we can solve for the visible 3D location in world coordinates X_w, Y_w, Z_w. The set of points define the viewing frustum.

$$(X_w, Y_w, Z_w) = \mathbf{C}_w + \lambda \mathbf{R}^{-1}\mathbf{K}^{-1}\mathbf{p}$$

Since all the points on the page are on a plane where the z coordinate is equal to zero we can calculate λ and solve for X_w Y_w. In other words we can recover the 3D position of a pixel coordinate of the marker.

$$\lambda = \frac{Z_w - C_{w,z}}{r_3}$$
$$\text{where } (r_1, r_2, r_3)^T = \mathbf{R}^{-1}\mathbf{K}^{-1}\mathbf{p} \text{ [19]}$$

We can find the bounds of the viewing frustum in world (meter) coordinates using the following relation. Without loss of generality a corner of the image plane can be denoted as $\mathbf{p}_i$.

$$(X_w^i, Y_w^i) = \mathbf{C}_w - \frac{C_{w,z}}{r_3}\mathbf{R}^{-1}\mathbf{K}^{-1}\mathbf{p}_i \quad (1)$$

This functions maps the pixel corners, $\mathbf{p}_i$ to visible world coordinates (X_w^i, Y_w^i). With this relation the four corners of the image plane define the viewing frustum in world coordinates. Then by testing if all four corners of the page, $\mathbf{\Pi}_i \in n$, are within the viewing frustum we can determine if the entire page is visible.

2) Minimum reading distance: OCR requires a sufficient resolution for accurate recognition. For scanned documents the rule is that a document should be scanned at least at 300 dpi [20]. For camera captured images we will focus on the generally accepted rule of thumb that a text line should map to at least 12 pixels [3]. According to typographic standards for Latin script the height of a 'x', x-height, is a standard measure from the baseline to average height of a letter. A lowercase 'x' in 12 point Arial font (a font approved by the American with Disabilities Act [21]) has a height 4.23 millimeters which forms a constraint that anywhere a 'x' could be printed must map to 12 pixels in the image plane. We use this equality to define the Compliant Space.

A document can be recognized by OCR if an ‘x’ printed on each corner $\mathbf{\Pi}_i$ (location of corner of page in world coordinates) maps to 12 pixels of the image plane. This can be calculated by projecting two 3D points into rays for each corner; 3D point $\mathbf{\Pi}_i^U$ is the upper point of the theoretical ‘x’ and $\mathbf{\Pi}_i^L$ is the lower point of the theoretical ‘x’. Recall we have already calculated Pose, $\mathbf{P}$, where $\mathbf{P} = [\mathbf{R} \mid \mathbf{T}]$. $\mathbf{P}$ is a 4 by 4 matrix where the fourth row is (0 0 0 1). For this computation it is convenient to map to homogeneous coordinates to compute the projective transform with matrix multiplication. In order to compute 3D to 2D projection we matrix multiply by camera intrinsic $\mathbf{K}$ and camera extrinsic $\mathbf{P}$. In the following equations. Without loss of generality we can find $\pi_i^{U_x}$ from $\Pi_i^{U_x}$ with the following relation. Recall that $\Pi_i^{U_z} = 0$ as all points on the page are along the $Z = 0$ plane.

$$(q_1, q_2, q_3)^T = \mathbf{K} \cdot \mathbf{P} \cdot (\Pi_i^{U_x}, \Pi_i^{U_y}, \Pi_i^{U_z}, 0)^T$$
$$\pi_i^{U_x} = \frac{q_1}{q_3}$$

Therefore we can compute the x-height in pixels of a 12 point Arial ‘x’ that is theoretically located at each $\mathbf{\Pi}_i$ of the page.

$$\text{x-height}_i = \sqrt{(\pi_i^{U_x} - \pi_i^{L_x})^2 + (\pi_i^{U_y} - \pi_i^{L_y})^2} \tag{2}$$

If $\forall i \in n$ such that x-height$_i \geq 12$ pixels is satisfied for all four corners then minimum reading distance is observed.

3) Guidance Algorithm: We approximated the problem of finding the shortest path to the Compliant Space by instead finding the shortest path to a 3D-line within the Compliant Space. We call this 3D-line the Reduced Compliant Space Center Line, L, which is between $(0, 0, .28)$ and $(0, 0, .42)$ in world coordinates (meters). This approximation is sufficient and does not lead to non-compliant images because for any pose we can determine compliance in real-time through conditions (1) and (2).

The guidance algorithm computes the shortest path from the current position, $\mathbf{C}_w$, to the closest point on L by projecting the point $\mathbf{C}_w$ onto the 3D-line L. The closest point on L to $\mathbf{C}_w$ is called $\mathbf{O}_w$. The software then verbalizes the two axes in need of the most correction. The instructions come in centimeter units such as “move the phone up 15 centimeters and forward 9 centimeters”. Additionally, the software notifies the user if they are holding the phone at a significant tilt or angle relative to the page. This is important because the Compliant Space assumes the phone is held directly over the page.

IV. Evaluation

A blind person trying to take a snap shot of a document to be recognized by OCR typically has difficulties taking a compliant image of any document. Our goal is to the verify whether a guidance system as described in the previous section can facilitate a blind person in taking compliant images. To perform this, we establish two metrics to assess the ability of a person without sight to acquire a compliant image of a document.

We can extract measures for each user per trial and aggregate them per Experimental Setting (ES). The most important measure is the percentage of successful trials, which is equal to $\frac{1}{8} \cdot$ Count of Successful Trials. It is also interesting to measure the distance from the closest point in the Reduced Compliant Space Center Line the user reached. This is the euclidean distance from current camera position to the point $\mathbf{O}_w$, which is equal to $\|\mathbf{C}_w - \mathbf{O}_w\|$.

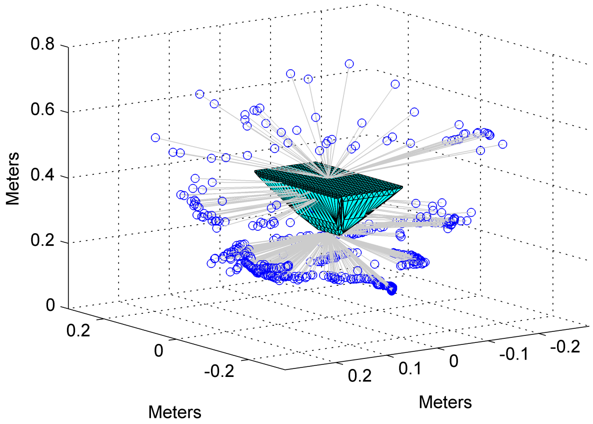

Fig. 2: Plot of a camera trajectory in 3D space. The compliant space is shown as the green tetrahedron. Each blue circle represents a camera position. At each detection an approximate shortest path (see section III-A3) can be found from camera position $\mathbf{C}_w$ to the closest point $\mathbf{O}_w$ on the Reduced Compliant Space Center Line.

A. Design of experiment

We used convenience sampling to recruit all eight of our participants. They are all adults who consented to being in the study. All participants are not blind and are therefore blindfolded, all but one participant was male. We understand that there will be differences between how blindfolded and blind users react to our system, therefore a follow up user study with the target community will soon commence.

The experiment began by handing the blindfolded participant a backpack that contained the document with fiducial markers. Then each participant heard the following instruction.

“Inside this backpack is a single piece of paper. Please take it out and feel for the side with a sticker. This side has the fiducial markers and should remain face up so it is visible to the camera. The sticker should be on the top left of the piece of paper with respect to you.”

The participant is then handed a phone and told to tap the screen to begin the trial. The phone vibrates to alert the user that the trial has begun. The participant has ninety seconds to complete. After the timeout the trial ends unsuccessfully.

Next we explain each of the Experimental Settings (ES) used in the experiment.

- **ES-Control** Provides no guidance. The user clicks the volume button when they believe the phone is in the correct position.
- **ES-Confirmation** There is no guidance. However, as soon as the phone has captured a compliant image the software alerts the subject and ends the trial.

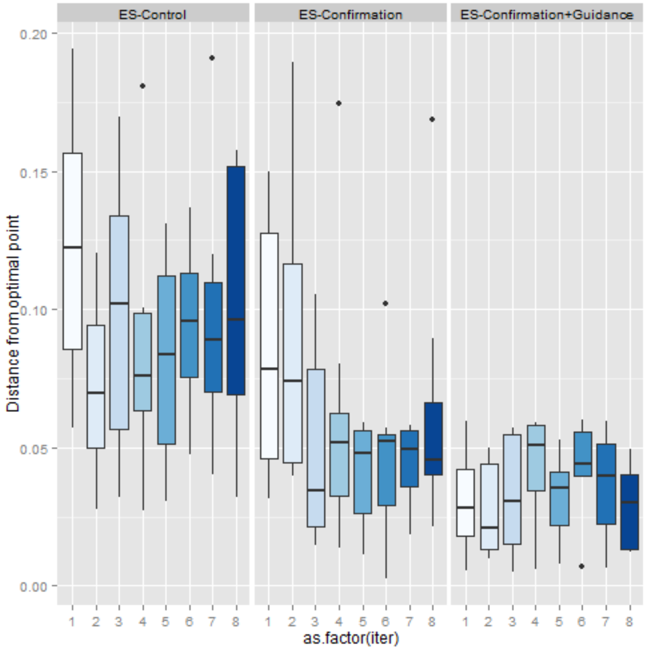

Fig. 3: The above plot is a facet box plot. The main facets are the experimental settings. Each experimental setting consists of eight trials differentiated by shading (best viewed in color). We can see a trend of small distances as users become more familiar with our system. These are per trial measures.

- **ES-Confirmation+Guidance** In this setting the user receives continuous guidance as described in section III-A3. As soon as the phone has captured a compliant image the software alerts the subject and ends the trial.

B. *Prototype software*

We implemented a real time marker recognition and guidance prototype iPhone application to obtain data to answer our hypotheses. The application is programmed to run each experimental setting for eight trials, the application logs the pose of the phone relative to the marker at approximately 3 hertz.

V. Results and Discussion

In this section we will summarize our findings based on the experiment conducted with our prototype software. Our hypothesis is that our guidance system will help people take a greater percentage of compliant images than they would be capable of without guidance. Starting with data exploration we can see a clear trend of lower distances from the optimal position by examining Figure 3.

In order to answer our hypothesis we turn to our measure of the percentage of successful trials. Since we knew we had a limited number of study participants we used negative Helmert Contrast [22] to design a tractable experiment that can tolerate only one orderings of experimental settings. Helmert Contrast compares the mean of a experimental setting with the means of previous experimental settings, and is often used in a medical context were an investigator is trying to discover the right dose of medicine. With this methodology we first to establish a baseline ES-control to get a measure of where the user is before intervention. Next, the user is given ES-just-confirmation and finally the user is provided eight trials of ES-full-guidance. This experiment design will allow us to answer whether or not guidance improved the user's ability to take a compliant image of the document. However, this experimental design does not rule out the effect of reordering experimental setting and therefore it is future work to redo the study with a Latin Square design.

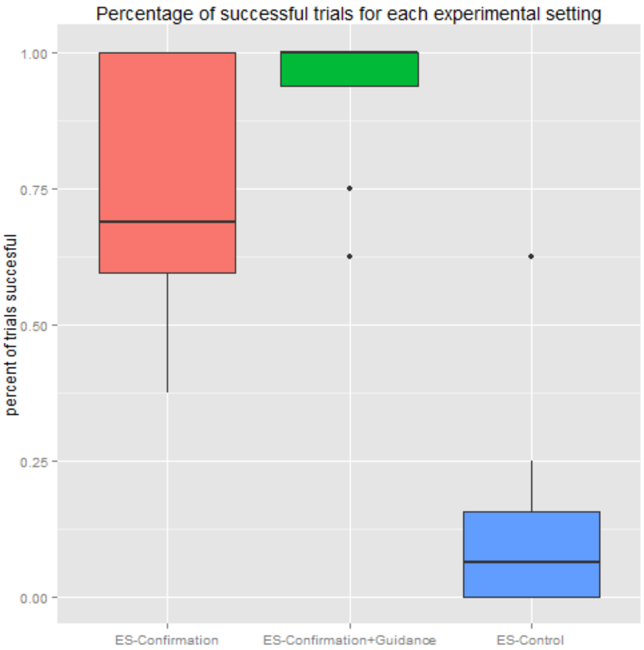

Fig. 4: This box plot shows users percentage of successful trials for each experimental setting.

Improvement trends are visible by examining Figure 4, were we can see a large improvement from the control in the subsequent trials. In addition, in order to statistically test for improvement we conducted the following analysis.

From the one way Analysis of Variance (see Table I) it is clear that between the ES-Confirmation+Guidance there is a significant improvement in the percent of trials that a compliant image is achieved. Between the ES-Confirmation and ES-Control there is a small enough p-value to indicate a trend but not small enough to be significant at the 95% confidence level.

A. *User Experience Report*

After completing the experiment we asked each of the users to describe their experience. The general consensus is that the ES-Confirmation+Guidance was preferable over ES-Control or ES-Confirmation. This is also supported by the average user's increase in accuracy after intervention.

Users reported frustration with one of the guidance features, which alerts the user if the phone was in the compliant

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	2.66	1.33	31.04	0.0000
group: C1	1	0.14	0.14	3.28	0.0845
group: C2	1	2.52	2.52	58.79	0.0000
Residuals	21	0.90	0.04		

TABLE I: This table is the summary of the one way ANOVA for the percent of succesful trials for each of the ES. ‘Group:C1 is measuring ES-Confirmation compared to the ES-Control. group:C2 is comparing ES-Confirmation+Guidance to ES-Confirmation and ES-Control. The latter is significant at the 95 percent level.

space while the marker was not in the viewing frustum (Equation 1). The message states “Make phone aligned and level” to indicate an improper orientation of the phone with respect to the document. However, this message is too vague since the user is unsure which of the three axes are improperly aligned. A superior version of the software would indicate if the phone is not level or if the phone is rotated around the optical axis.

A learning curve is noticeable as the participants gained a sense of where the acceptable space is located. Some users would spend the first few trials patiently probing possible positions to find the Compliant Space in ES-Just-Confirmation. However, once the user found the correct position they were much faster in returning to the Compliant Space in subsequent trials. Users typically became frustrated if they did not find the Compliant Space before the timeout occurred. Once they found the Compliant Space for the first time this frustration appeared to subside.

VI. Conclusion

We demonstrated that that our guidance system significantly improves compliant document image acquisition on camera phones. Our experiment provides evidence to support further research and development in adding non-obtrusive fiducial markers to documents. In addition, future work on an improved guidance system will take into account post-hoc user feedback before beginning. If successful this will provide far better document accessibility for the blind and visually impaired.

Acknowledgment

The project described was supported by Grant Number 1R21EY021643-01 from NEI/NIH. The authors would like to thank Professor Sri Kurniawan for helpful discussions.

References

[1] N. Ezaki, K. Kiyota, B. Minh, M. Bulacu, and L. Schomaker, “Improved text-detection methods for a camera-based text reading system for blind persons,” in *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, aug.-1 sept. 2005, pp. 257 – 261 Vol. 1.

[2] “Knfb reader mobile,” knfb Reading Technology, Inc., 2008, http://www.knfbreader.com/.

[3] A. Zandifar and A. Chahine, “A video based interface to textual information for the visually impaired,” in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ser. ICMI ’02. Washington, DC, USA: IEEE Computer Society, 2002, pp. 325–. [Online]. Available: http://dx.doi.org/10.1109/ICMI.2002.1167016

[4] “Text detective (blindsight),” Blindsight, Inc., 2011, http://blindsight.com/textdetective/.

[5] C. Yi and Y. Tian, “Assistive text reading from complex background for blind persons,” in *Proceedings of the 4th international conference on Camera-Based Document Analysis and Recognition*, ser. CBDAR’11. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 15–28. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-29364-1_2

[6] H. Shen and J. M. Coughlan, “Towards a real-time system for finding and reading signs for visually impaired users,” in *Proceedings of the 13th international conference on Computers Helping People with Special Needs - Volume Part II*, ser. ICCHP’12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 41–47. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-31534-3_7

[7] C. Jayant, H. Ji, S. White, and J. P. Bigham, “Supporting blind photography,” in *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, ser. ASSETS ’11. New York, NY, USA: ACM, 2011, pp. 203–210. [Online]. Available: http://doi.acm.org/10.1145/2049536.2049573

[8] M. Vázquez and A. Steinfeld, “Helping visually impaired users properly aim a camera,” in *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, ser. ASSETS ’12. New York, NY, USA: ACM, 2012, pp. 95–102. [Online]. Available: http://doi.acm.org/10.1145/2384916.2384934

[9] “Aruco: a minimal library for augmented reality applications based on opencv,” Universidad D Cordoba, 2012, http://www.uco.es/investiga/grupos/ava/node/26.

[10] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, 2000.

[11] F. Guimbretière, “Paper augmented digital documents,” in *Proceedings of the 16th annual ACM symposium on User interface software and technology*, ser. UIST ’03. New York, NY, USA: ACM, 2003, pp. 51–60. [Online]. Available: http://doi.acm.org/10.1145/964696.964702

[12] T. Nakai, K. Kise, and M. Iwamura, “Camera-based document image mosaicing using llah,” in *DRR*, ser. SPIE Proceedings, K. Berkner and L. Likforman-Sulem, Eds., vol. 7247. SPIE, 2009, pp. 1–10.

[13] K. Iwata, K. Kise, T. Nakai, M. Iwamura, S. Uchida, and S. Omachi, “Capturing digital ink as retrieving fragments of document images,” in *ICDAR*. IEEE Computer Society, 2009, pp. 1236–1240.

[14] T. Nakai, K. Kise, and M. Iwamura, “Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval,” in *Document Analysis Systems VII*, ser. Lecture Notes in Computer Science, H. Bunke and A. Spitz, Eds. Springer Berlin Heidelberg, 2006, vol. 3872, pp. 541–552. [Online]. Available: http://dx.doi.org/10.1007/11669487_48

[15] J. Liang, D. DeMenthon, and D. Doermann, “Camera-based document image mosaicing,” in *Proceedings of the 18th International Conference on Pattern Recognition - Volume 02*, ser. ICPR ’06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 476–479. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2006.352

[16] L. Quan and Z.-D. Lan, “Linear N-Point Camera Pose Determination,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 8, pp. 774–780, Aug. 1999. [Online]. Available: http://hal.inria.fr/inria-00590105

[17] V. Lepetit, F. Moreno-Noguer, and P. Fua, “Epnp: An accurate o(n) solution to the pnp problem,” *Int. J. Comput. Vision*, vol. 81, no. 2, pp. 155–166, Feb. 2009. [Online]. Available: http://dx.doi.org/10.1007/s11263-008-0152-6

[18] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000. [Online]. Available: http://dx.doi.org/10.1109/34.888718

[19] Y. Morvan, “Multi-view-coding-thesis,” 2009, http://www.epixea.com/research/multi-view-coding-thesis.html.

[20] 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. 84–104–104, Jul. 2005. [Online]. Available: http://dx.doi.org/10.1007/s10032-004-0138-z

[21] “Americans with disabilities act of 1990,” Pub. L. 101-336. 26, July 1990. 104 Stat. 328.

[22] J. M. Chambers and T. J. Hastie, Eds., *Statistical Models in S*. Pacific Grove, CA: Wadsworth and Brooks/Cole, 1992.