Mobile phone camera-based video scanning of paper documents

Muhammad Muzzamil Luqman, Petra Gomez-Krämer, Jean-Marc Ogier

Mobile phone camera-based video scanning of paper documents

Muhammad Muzzamil Luqman, Petra Gomez-Krämer and Jean-Marc Ogier
L3i Laboratory, University of La Rochelle, Avenue M. Crépeau, 17042 La Rochelle, France
{muhammad_muzzamil.luqman,petra.gomez,jean-marc.ogier}@univ-lr.fr

***Abstract*—Mobile phone camera-based document video scanning is an interesting research problem which has entered into a new era with the emergence of widely used, processing capable and motion sensors equipped smartphones. We present our ongoing research on mobile phone camera-based document image mosaic reconstruction method for video scanning of paper documents. In this work, we have optimized the classic keypoint feature descriptor based image registration method, by employing the accelerometer and gyroscope sensor data. Experimental results are evaluated using optical character recognition (OCR) on the reconstructed mosaic from mobile phone camera-based video scanning of paper documents.**

I. Introduction

In recent years, the availability of camera equipped, processing capable, inertial sensors fitted and moderate priced mobile phones (a.k.a. smartphones), has attracted the attention of the research community to employ them for complementing the classical document scanning devices. The use of these devices for document scanning provides interesting advantages over the traditional document scanning devices. They can be used to scan thick books, historical documents that are too fragile to touch, text in scenes (walls, whiteboards, etc.), and large sized documents [1]. However, the use of smartphones introduces new challenges to document scanning which are not faced by classical document scanning devices. These challenges include uneven lighting, perspective distortion, non-planer surfaces, motion blur and low resolution of the cameras [1][2].

In this paper we present our ongoing work on mobile phone camera-based video scanning of paper documents. The video scanning of a paper document is achieved by swiping the mobile phone camera over the paper document and recording the accelerometer sensor data along with capturing the video frames. During the video scanning the orientation of the mobile phone camera is obtained from the gyroscope data and the user is provided with visual feedback on the orientation of the phone to avoid perspective distortion. A complete mosaic image of the paper document is reconstructed from the captured video frames by employing an optimized keypoint feature descriptor-based image registration techniques. The optimization is achieved by employing the recorded accelerometer sensor data. The resulting reconstructed mosaic has a higher resolution than a simple photo of the document taken by same camera.

In literature the camera-based scanning of paper documents has been approached by various works which are mainly motivated by panorama reconstruction and image mosaicing techniques from the computer vision research community. In [2] first image feature-based technique is used to estimate the camera motion and to assist the user to capture images of patches of document. The estimated camera motion is used with a keypoint feature descriptor-based technique for registration of captured image patches and reconstruction of a mosaic of the document. In [3] an algorithm for 2D scanning of a planar scene is proposed. The topology of the video frames are inferred on a 2D manifold by alignment of successive video frames and overlapping video frames. The aligned frames are merged by using a multi-resolution method for constructing a seamless mosaic. In [4] local likely arrangement hashing (LLAH), which is originally an image retrieval technique, is used for keypoint detection and feature description in frames. Images are aligned by matching LLAH feature descriptors and the feature correspondences are used for combining input frames for reconstruction of mosaic. In [5] first captured frames are rectified by removing perspective distortion using texture flow information. A hough transform-based voting scheme is used for finding translation and scaling between video frames. The reconstruction of the mosaic is achieved by a sharpness-based seamless composition of overlapping images. In [6] inertial sensors in mobile phones have been employed for constructing panoramas on mobile phones. In a first step the position and relative displacement of video frames are computed by inertial sensor data. Using the alignment estimation from inertial sensor data, a more precise alignment of the video frames is computed by using a keypoint feature descriptor-based technique and the mosaic image is constructed by using the feature correspondences.

The perspective distortion is very important to be handled in case of camera-based document scanning. A document image mosaicing technique should directly or indirectly rectify perspective distortion of the captured frames before reconstructing the mosaic image. A summary of methods for content based correction of the perspective distortion in camera captured document images is presented in [7].

In this paper we present our ongoing research on document image mosaicing. We are inspired by the work in [6] for employing the inertial sensors for document mosaic image reconstruction. However, we are working on elaborating a lightweight algorithm that could be implemented on smartphones. The two novel contributions of our work are the following: 1) We use the gyroscope sensor to give visual user feedback to avoid perspective distortion during the video scan of the paper document whereas perspective distorsion is not considered in [6]. 2) We compute the direction of swipe to optimize the keypoint feature descriptor-based image registration method by using only accelerometer sensor data whereas

the authors of [6] optimize a keypoint feature descriptor-based image registration method by computing the displacement of the mobile phone from accelerometer and gyroscope sensor data.

The remainder of this paper is organized as follows. We present a detailed description of our method of video scanning of paper documents in Section II. In Section III we discuss the experimental evaluation and the results. In Section IV we conclude our work and present future directions of research.

II. Mobile phone camera-based video scanning

In this section we present a detailed description of our method for mobile phone camera-based video scanning of paper documents. We first describe the capturing of video frames and accelerometer sensor data along with the gyroscope based visual feedback for avoiding perspective distortion. This is followed by description of our methodology for finding the direction of swipe (of video scanning), from the accelerometer data recorded with the captured frames in the video sequence. Finally we describe the image registration of the frames of the captured video and the reconstruction of the complete mosaic image of the paper document.

A block diagram of our method for mobile phone camera based video scanning of paper documents is presented in Fig.1.

A. Video scanning of paper documents

The video scanning of a paper document is achieved by a one-dimensional swipe of the mobile phone camera on the paper document. The swipe could either be from the top to the bottom of the document or from the bottom to the top of the document. During the video scanning we record the accelerometer sensor data along with capturing of the video frames.

Processing of accelerometer sensor data: An accelerometer in a smartphone measures the acceleration $\alpha = (\alpha_x, \alpha_y, \alpha_z)$ of the phone in each direction of the X, Y and Z-axis. The accelerometers in smartphones are usually not of very high quality (because of cost constraints) and thus the obtained acceleration data is very noisy. Hence, the raw accelerometer readings are full of random noise and are unusable in their original form. The rise in temperature of the mobile phone (resulting from camera and screen heat) increases the random noise in the accelerometer data. In order to make sure that the accelerometer reading is as close to the real value as possible, we compute the calibration offset of the accelerometer sensor by placing the phone on a flat surface and averaging x readings along each of the three axes separately and independently. We then subtract the calibration offset from the future readings of the accelerometer; hence obtaining calibrated readings. For removing random noise from the accelerometer readings we smooth the accelerometer readings by using a Kalman filter-based running average. The smoothed accelerometer readings are recorded with the video frames. Fig. 2 presents respectively the raw (noisy) accelerometer readings for the whole video scanning of a paper document, the calibrated and smoothed accelerometer readings for the whole video scanning of a paper document, the raw accelerometer data associated to the captured frames and the calibrated and smoothed accelerometer data associated to the captured frames, for the X-axis of the phone accelerometer sensor.

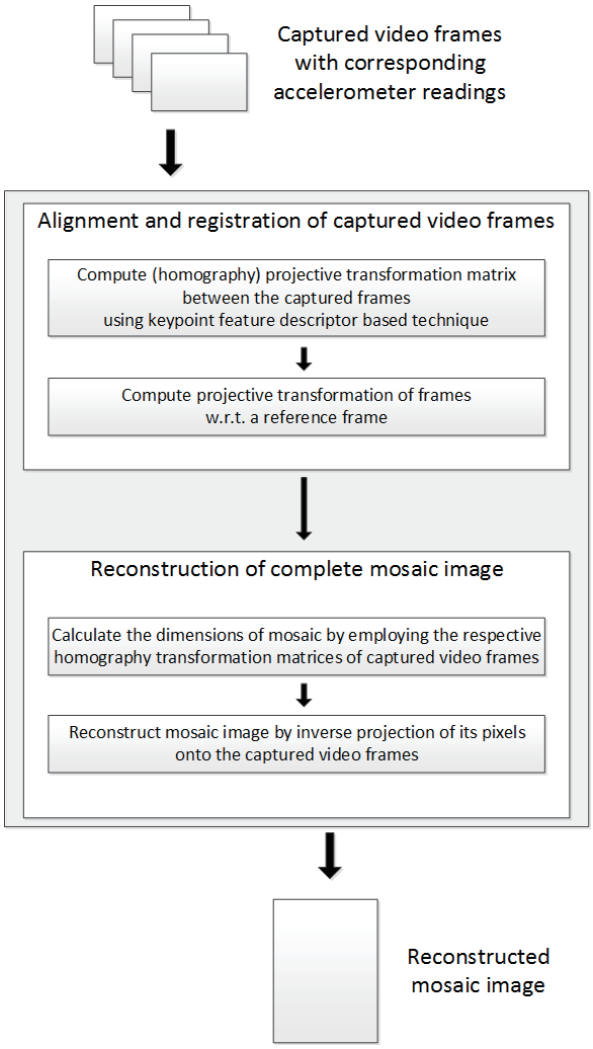
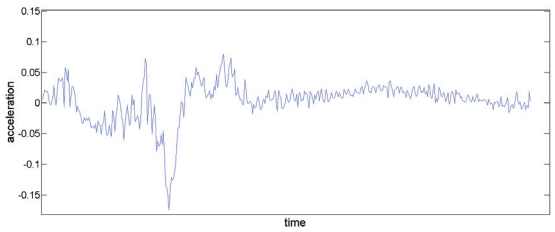
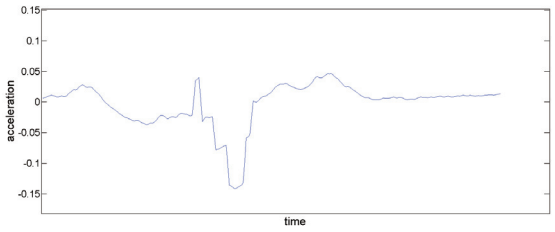


Fig. 1. Block diagram for mobile phone camera-based video scanning of paper documents.

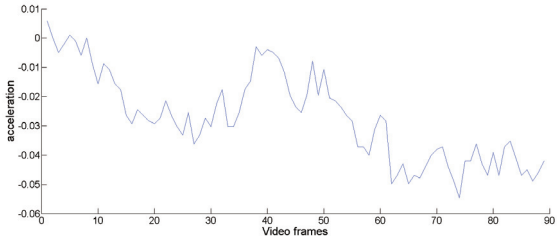
Use of gyroscope sensor data for visual user feedback: In order to avoid perspective distortion, we employ the gyroscope sensor data for obtaining the orientation of the mobile phone camera. Modern smartphones are fitted with 3D gyroscopes, which measure the angular velocity $\omega = (\omega_x, \omega_y, \omega_z)$ along the X, Y and Z-axis of the phone. The angular velocity along the X-axis is termed pitch, along the Y-axis yaw and along the Z-axis roll. We obtain the angular velocity from the gyroscope as an angle of rotation around each of the three axes. We then employ it for providing a visual feedback to the user in order to keep the mobile phone camera parallel to the document plane and so to avoid the perspective distortion. We show three triangles on the screen for the visual feedback to adjust the orientation of the device. The three triangles are mapped to the gyroscope data along the three axis, respectively. A change in orientation of the device along an axis, rotates its respective triangle on the screen. By aligning any two of the three triangles, the user can keep the mobile phone parallel to



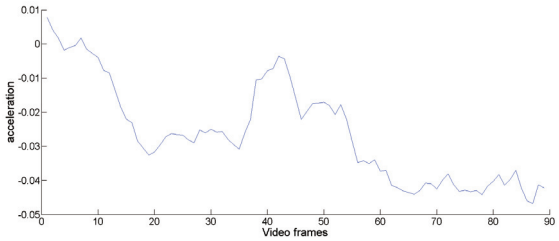
(a) Raw accelerometer readings of video scanning of a document.



(b) Calibrated & smoothed accelerometer readings of video scanning of a document.



(c) Raw accelerometer readings associated with video frames.



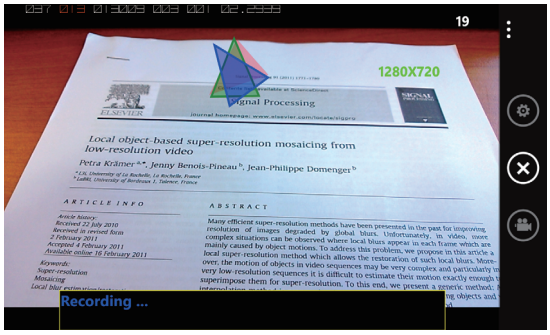
(d) Calibrated & smoothed accelerometer readings associated with video frames.

Fig. 2. Accelerometer data recorded for the X-axis of the mobile phone sensor during video scanning of a paper document.

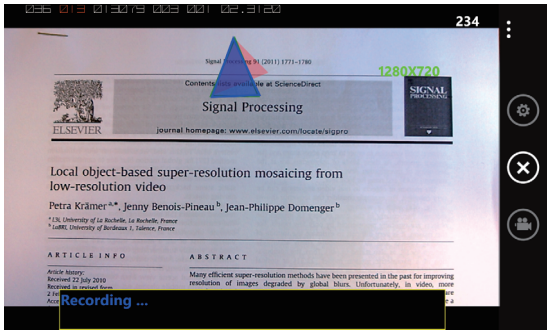
the XY, YZ or ZX plane. To illustrate it pictorially, the Fig. 3 presents some screenshots of the interface during the video scanning of a paper document.

B. Finding the direction of swipe by using accelerometer data

The accelerometer sensor data measures the acceleration of the phone along the X, Y and Z-axis. We use the calibrated and smoothed accelerometer data for inferring the swipe direction of the mobile phone during the video scanning of the paper document. The accelerometer readings along each of the three axes is between $-1g$ and $+1g$ (where $g = 9.8m/s^2$). The document page is placed on a planar surface (e.g. a table) and we use the phone in landscape mode (as shown in the screenshots of Fig. 3) for video scanning of documents. This



(a) Misaligned triangles resulting in perspective distortion.



(b) Well-aligned triangles avoiding perspective distortion.

Fig. 3. Screenshots of the interface during video scanning of a paper document.

setup makes the mobile phone's X-axis as the primary axis of swipe. The sensor coordinate system, indicating the X, Y and Z-axis of the mobile phone that we use for our research, is shown in Fig. 4.

To compute the swipe direction of the mobile phone (from the top to the bottom or from the bottom to the top of the document page) we employ a very simple methodology. We count the number of positive and negative readings in the recorded accelerometer data of the frames of a video scan. If there are more negative values than positive ones, this means that the phone is swiped in negative direction of the X-axis. And if there are more positive values than the negative ones, this means that the phone is swiped in positive direction of the X-axis. A swipe in negative X-axis direction corresponds

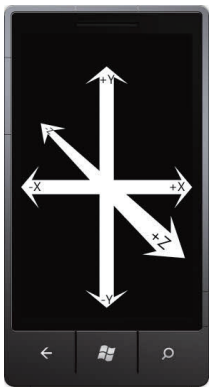


Fig. 4. Windows phone accelerometer sensor coordinate system.

to a top to bottom video scan of a paper document whereas a swipe in positive X-axis direction corresponds to a bottom to top video scan. This simple methodology is robust, efficient and very useful for detecting the direction of swipe during the video scanning of paper documents.

C. Image registration of the video frames

The captured frames from video scanning of the paper document are registered by employing a keypoint feature descriptor-based technique. Successive frames in the video sequence are aligned and the projective transform or homography is computed between them. Afterwards, the computed homographies between successive frames are employed for calculating the homography of each frame to a reference frame in the video sequence.

Optimized alignment of successive video frames: For the alignment of successive frames in the video sequence, we use a keypoint feature descriptor-based alignment technique. As discussed by [8], the feature based image alignment techniques have some very interesting advantages over pixel-based image alignment techniques. Namely for mobile phone camera-based video scanning of paper documents, the feature based image alignment methods are more efficient and robust than the pixel based methods; specially in case of uneven lighting and scene motions. There are many keypoint detectors and feature descriptors in state-of-the-art. They include the famous SIFT [9], SURF [10], FAST [11], ORB [12] and FREAK [13].

To align two successive frames, we first employ the SIFT keypoint detector to obtain a set of keypoints in two frames. Second, we extract the SIFT feature descriptors on each of the detected keypoints in the two frames. Third, we perform FLANN based feature matching [14] between the two frames and employ RANSAC [15] for refining the initial correspondences obtained by FLANN.

In order to optimize keypoint detection, feature descriptor computation and feature matching, we use the direction of mobile phone swipe during video scanning. During the keypoint detection and feature descriptor computation phases we use the direction of swipe to avoid processing the complete image and to ignore the top and bottom parts of successive frames (or vice versa depending upon the direction of swipe). As a result, it reduces the search space to be exploited during feature matching. The size of top and bottom parts of successive frames respectively to be ignored considering the direction of swipe is controlled by a parameter which is computed automatically for a video sequence and it takes into account the resolution of video frames (in pixels) and the speed of the swipe in the sequence (assumed to be constant during the video scanning). The speed of swipe is estimated from the total number of video frames in the video scan of a document page. The parameter for ignoring the top and bottom parts of successive frames is computed as:

$$G = \frac{h}{n} \tag{1}$$

where G denotes the number of pixel rows to be ignored on the top and bottom of successive frames respectively, h is the height of captured video frames in pixels and n is the total number of frames in video sequence.

Homography computation for two successive frames: We compute the planar homography or projective transform between two successive frames by minimizing the backpropagation error and further refine the computed homography by using the Levenberg-Marquardt method to minimize the backpropagation error [8]. The homography or projective transform between two successive frames is represented by a homography matrix.

For a captured video sequence of n frames given by:

$$V = \{f_1, f_2, f_3, \ldots, f_{n-1}, f_n\} \tag{2}$$

the set of homographies between successive frames is:

$$h = \{h_{(1,2)}, h_{(2,3)}, h_{(3,4)}, \ldots, h_{(n-1,n)}\} \tag{3}$$

where $h_{(i,j)}$ is the homography between frames f_i and f_j.

Homography between non-successives frames: Employing the well established properties of matrices, the homographies computed for successive frames are employed in a cascade matrix multiplication, for computing the homographies between non-successive frames. For example for computing the homography $h_{(1,5)}$ between the frames f_1 and f_5, the homographies $h_{(1,2)}$, $h_{(2,3)}$, $h_{(3,4)}$ and $h_{(4,5)}$ are matrix multiplied. This permits us to define a homography between any pair of frames in the video sequence (whether successive or non-successive).

D. Reconstruction of the complete mosaic

The complete mosaic image of the video scanned paper document is constructed by a projection of the pixels in captured frames onto a reference frame. Thus, we first select a reference frame in the captured video sequence. For simplicity, we suppose here that the first frame f_1 is selected as the reference frame. Then, we compute the size of the complete mosaic image by projecting the four corners of each frame in the video sequence using the corresponding homography matrix of the frame. To avoid holes or missing pixels in the complete mosaic image, the construction of the complete mosaic image is achieved by inverse projection of each of its pixels onto the sequence of frames. If a pixel of the mosaic is projected onto a subpixel in a frame, we use bilinear interpolation for computing the subpixel intensity from the intensities of the neighboring pixels. If a pixel of the mosaic is projected onto more that one frames of the video sequence we use the median value of the intensities of corresponding pixel of those frames.

III. Experimentation

In this section, we evaluate our method for mobile phone camera-based video scanning of paper documents on video frames captured at a resolution of 1280x720 pixels captured by a Nokia Lumia 920 smartphone. A custom application is developed for the capture phase of video scanning of paper documents. Some screenshots of this application are presented in Fig. 3. The video scanning application runs only in landscape mode to force the user to hold the smartphone with two hands and thus ensuring a stable orientation of the smartphone during capture.

During these preliminary experimentations the mosaic image reconstruction was performed on a laptop computer. The video scanning of paper documents was performed in an office environment with normal lighting conditions. The phone was kept parallel to the document plane by following the visual feedback on the orientation of the mobile phone camera, and the swipe was performed slowly and carefully.

The experimentation dataset comprises fifteen A4 sized pages of scientific research papers; containing mostly printed textual content (in English). Some document pages contain also tables and mathematical equations. The document pages were printed on A4 pages and were digitized in three different modes, as given below:

1) image scanned by a classic scanner at 300dpi grayscale
2) photo taken by the smartphone at a resolution of 1280x720 pixels
3) video scanned by the smartphone at a resolution of 1280x720 pixels

The image scanned by a classic scanner at 300dpi serves as reference for evaluating the quality of reconstructed mosaics. We use the Levenshtein distance [16] as metric for comparing the OCR results of the images with ground truth. The Levenshtein distance between two string sequences is the edit distance between them i.e. the minimum number of single character edits (insert, delete, replace) required to change one string sequence into the other.

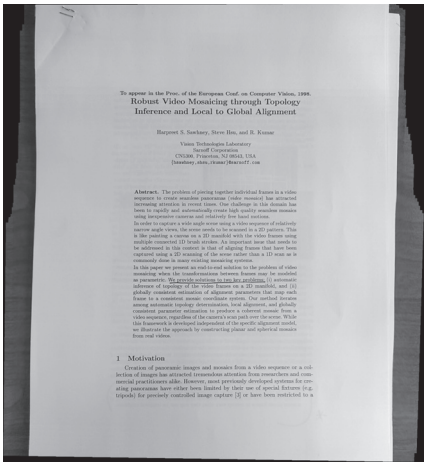
Table I presents a comparison of the Google Drive OCR results on documents for the three digization modes. The OCR results on the reconstructed mosaic from video scanning of the paper document at 1280x720 pixels, are better than the results on the single image captured at 1280x720 pixels. However they are not as good as those of the classical scanner scanned image. One important reason for this is that our method does not perform any camera calibration i.e. any wide angle lens correction on the captured video frames. The wide angle lens noise is thus inherited by the mosaic image and it eventually effects the OCR results. A second reason is that when the document pages were placed on the table for video scanning there can be small curvature at the corners whereas in case of a scanner this curvatures are flattened by closing the scanner lid. Our method does not perform any preprocessing of the document page.

Some examples of reconstructed mosaic images in grayscale from mobile phone camera-based video scanning of A4 sized paper documents are presented in Fig. 5.

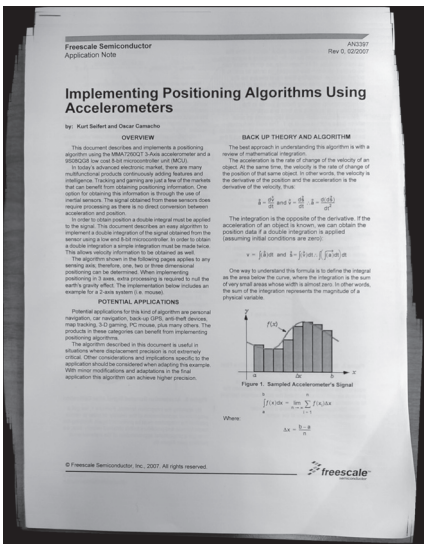
Apart from the advantages of portability and liberty of scanning the documents of different sizes, the video scanning of paper documents is interesting as it allows to obtain the mosaic image at a higher resolution than the resolution of the camera used for capturing the frames. A mobile phone camera with a resolution of 1280x720 pixels can take a single image of an A4 sized page at 98 dpi as given by:

$$dpi = \sqrt{\frac{1280 \times 720}{8.27 \times 11.69}} = 98 \qquad (4)$$

where, 8.27×11.69 is the size of an A4 page in inches.



(a) Resolution: 1492x1627 pixels, 159 dpi



(b) Resolution: 1508x1930 pixels, 174 dpi

(c) Resolution: 1407x1814, 163 dpi

Fig. 5. Reconstructed mosaic images for A4 sized paper documents from video scanning at 1280x720 pixels.

TABLE I. EXPERIMENTAL RESULTS

		Levenshtein distance between ground truth and Google Drive OCR results		
Page	# chars in page	Classical scanner image at 300 dpi	Single image at 1280x720 pixels	Mosaic from video scan at 1280x720 pixels
01	2569	36	1330	37
02	2311	44	808	57
03	1854	10	456	56
04	2353	10	441	125
05	2438	26	885	69
06	2495	22	1061	41
07	2171	20	678	542
08	2524	72	623	99
09	1422	296	785	354
10	2482	235	1417	487
11	2085	9	431	22
12	3286	382	3230	1442
13	4299	61	3786	111
14	3638	107	2112	1451
15	3924	300	3149	606
Mean	**2657**	**109**	**1413**	**367**

TABLE II. COMPARISON OF COMPUTATION TIMES

	Computation times (seconds) for mosaic construction from video scanning at 1280x720 pixels	
Page	Without optimization	With optimization
01	417	397
02	330	309
03	227	203
04	332	301
05	371	347
06	395	376
07	337	313
Mean	**344**	**321**

Whereas when the same mobile phone camera is used for video scanning of A4 sized pages, the reconstructed mosaic images as shown in Fig. 5 are at 159 dpi, 174 dpi and 163 dpi respectively.

Table II show a comparison of computation times for the reconstructed mosaics from video scanning at a resolution of 1280x720 pixels of A4 sized pages. Computation times are shown for mosaic construction with and without optimized registration using accelerometer sensor data. Using the optimiztion a mean speed up of 23 seconds could be realized.

IV. CONCLUSION

We have presented our ongoing research on the mobile phone camera based video scanning of paper documents. Our method employs the gyroscope sensor of the phone for providing a visual feedback to the user for avoiding perspective distortion, and the accelerometer sensor of the phone for optimizing the keypoint feature descriptor-based image mosaicing technique. Our preliminary experimentation shows that the video scanning of documents not only allows to reconstruct the full page mosaic image of a document page from its mobile phone camera-based video scanning, but also reconstructs the full page mosaic image at a better resolution than the physical limits of the camera lens. The work is in progress and we are working on the detailed experimental evaluation of the method along with an implementation on the smartphone platform.

Our ongoing research focus is on employing the gyroscope data for correcting perspective distortion of the frames in addition to the visual feedback. A second direction of ongoing research is to use super-resolution techniques for improving the quality of the mosaic image. In near future we will explore the use of the ambient light sensor for incorporating the lighting conditions of the video scan environment in mosaic reconstruction. In medium term we have planned to include a preprocessing step in our system for rectifying various geometric noises from mobile phone camera captured document image frames.

ACKNOWLEDGMENT

The piXL project is supported by the ”Fonds national pour la Société Numérique” of the French State by means of the ”Programme d’Investissements d’Avenir”, and referenced under PIA-FSN2-PIXL. For more details and resources, visit http://valconum.fr/index.php/les-projets/pixl.

REFERENCES

[1] J. Liang, D. Doermann, and H. Li, “Camera-based analysis of text and documents: a survey,” *International Journal of Document Analysis and Recognition*, vol. 7, no. 2-3, pp. 84–104, 2005.

[2] J. Hannuksela, P. Sangi, J. Heikkila, X. Liu, and D. Doermann, “Document image mosaicing with mobile phones,” in *International Conference on Image Analysis and Processing*, 2007, pp. 575–582.

[3] H. Sawhney, S. Hsu, and R. Kumar, “Robust video mosaicing through topology inference and local to global alignment,” in *European Conference on Computer Vision*, 1998, pp. 103–119.

[4] T. Nakai, K. Kise, and M. Iwamura, “Camera-based document image mosaicing using LLAH,” in *Document Recognition and Retrieval XVI*, 2009, pp. 1–10.

[5] J. Liang, D. DeMenthon, and D. Doermann, “Mosaicing of camera-captured document images,” *Computer Vision and Image Understanding*, vol. 113, no. 4, pp. 572–579, 2009.

[6] Q. Yang, C. Wang, Y. Gao, H. Qu, and E. Chang, “Inertial sensors aided image alignment and stitching for panorama on mobile phones,” *International Workshop on Mobile Location-based Service*, pp. 21–30, 2011.

[7] L. Jagannathan and C. Jawahar, “Perspective correction methods for camera based document analysis,” *International Workshop on Camera-based Document Analysis and Recognition*, pp. 148–154, 2005.

[8] R. Szeliski, “Image Alignment and Stitching,” in *Handbook of Mathematical Models in Computer Vision*. Springer, 2006, pp. 273–292.

[9] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] H. Bay, T. Tuytelaars, and L. Gool, “Surf: Speeded up robust features,” in *European Conference on Computer Vision*, 2006, pp. 404–417.

[11] E. Rosten and T. Drummond, “Machine learning for high-speed corner detection,” in *European Conference on Computer Vision*, 2006, pp. 430–443.

[12] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” in *International Conference on Computer Vision*, 2011, pp. 2564–2571.

[13] A. Alahi, R. Ortiz, and P. Vandergheynst, “FREAK: Fast Retina Keypoint,” in *International Conference on Computer Vision and Pattern Recognition*, 2012, pp. 510–517.

[14] M. Muja and D. Lowe, “Fast approximate nearest neighbors with automatic algorithm configuration,” in *International Conference on Computer Vision Theory and Applications*, 2009, pp. 331–340.

[15] M. Fischler and R. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, 1981.

[16] V. Levenshtein, “Binary codes capable of correcting deletions, insertions and reversals,” *Soviet physics doklady*, vol. 10, no. 8, pp. 707–710, 1966.