Click-Free, Video-Based Document Capture – Methodology and Evaluation

Waqas Tariq, Nazar Khan
Computer Vision & Machine Learning Group
Punjab University College of Information Technology
Lahore, Pakistan
Email: {waqas.tariq,nazarkhan}@pucit.edu.pk

***Abstract*—We propose a click-free method for video-based digitization of multi-page documents. The work is targeted at the non-commercial, low-volume, home user. The document is viewed through a mounted camera and the user is only required to turn pages manually while the system automatically extracts the video frames representing stationary document pages. This is in contrast to traditional document conversion approaches such as photocopying and scanning which can be time-consuming, repetitive, redundant and can lead to document deterioration.**

Main contributions of our work are i) a 3-step method for automatic extraction of unique, stable and clear document pages from video, ii) a manually annotated data set of 37 videos consisting of 763 page turn events covering a large variety of documents, and iii) a soft, quantitative evaluation criterion that is highly correlated with the hard F_1-measure. The criterion is motivated by the need to counter the subjectivity in human marked ground truth for videos. On our data set, we report an F_1-measure of 0.91 and a soft score of 0.94 for the page extraction task.

1. Introduction

The conversion of multi-page documents into their digital or even non-digital replicas is an ubiquitous process in the modern world. Applications range from low-level (personal copies) to mid-level (official copies) and even industrial scale [1], [2]. Current solutions include traditional photocopying, optical scanning and specialized camera-based conversion. Unfortunately, all current solutions suffer from at least one of the following problems:

1) time-consuming process,
2) repetitive,
3) redundant effort,
4) requirement of specialized hardware,
5) expensive setup,
6) document deterioration

The challenges in static image-based analysis of documents such as non-uniform lighting, optical distortions and non-planarity of documents have received tremendous attention over the last three decades [3], [4] and continue to do so [5], [6]. Video-based acquisition of documents poses some additional challenges such as memory requirement and dynamic effects. For instance, successive video frames can capture dynamic non-planarity changes due to i) document binding, ii) air disturbance, and iii) human interference to control i) and ii).

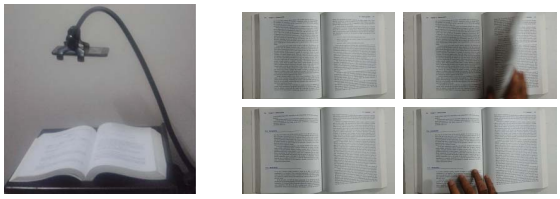


Figure 1. **Left**: An inexpensive setup for document video capture consisting of a camera enabled cell-phone and a cell-phone mount. **Right**: Four frames representing a page turn event (clockwise from top-left).

Compared to scanning-based solutions, camera-based solutions offer the flexibility of working in less constrained environments and digitizing documents that are not easily scannable (such as thick books or large and old documents). Combined with the mushrooming growth and increasing quality of consumer-grade cameras, many innovative solutions and applications have been introduced for camera-based document analysis [7], [8].

We present an inexpensive, video-based, multi-page document capture method that minimizes the user's effort and time without requiring specialized hardware. The work is targeted at the non-commercial, low-volume, personal user. The document is viewed through a mounted or handheld, downwards facing camera and the user is only required to turn pages manually (see Figure 1). The system automatically extracts video frames representing stationary document pages. Our main contributions include:

1) a 3-step method for automatic extraction of document pages from video,
2) a manually annotated data set of 37 videos consisting of 763 page turn events covering a large variety of multi-page documents and page turn methods, and
3) a soft, quantitative evaluation criterion that is highly correlated with the hard F_1-measure criterion.

On our data set, we report an F_1-measure of 0.91 and a soft score of 0.94 for the page extraction task.

2379-2140/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDAR.2017.344

2. Methodology

In general, document digitization is a four step procedure requiring i) an automatic or manual page turner, ii) digital page image capture, iii) page image dewarping, and iv) OCR of page contents. The focus of this work is the second step. Specifically, we present a mechanism for page image capture from videos of documents whose pages are being turned/replaced manually. The choice of working with manual page turns makes our solution accessible for low-volume, personal usage. We next describe our 3-step solution.

2.1. Step 1: Page Turn Removal (PTR)

The key assumption here is that while the page is being turned, there will be significant movement in the scene. In contrast, there will be minimal movement between two page turn events (PTE) and all frames in this range will be similar to each other. Therefore, if we compute the temporal gradient magnitudes at each pixel and compute their average value for each frame, it will yield a 1D signal of temporal gradient magnitude averages. In this signal, PTEs correspond to rapid changes and stationary page image frames correspond to regions of low signal variation. Figure 2 shows that small, non-overlapping chunks of frames corresponding to PTEs have a large standard deviation amongst their averages. This can be used to filter them out.

Accordingly, we assume that time between typical PTEs is at least half a second. For a video recorded at 30 frames-per-second, this amounts to 15 frames. Therefore, we pick non-overlapping chunks of 15 frames each and perform a temporal subsampling by factor 3 to obtain only 3 frames. This accounts for the temporal redundancy in videos and speeds up the processing pipeline. An additional benefit of limiting processing to small chunks is low memory requirement when dealing with arbitrarily long videos. We use these 3 frames to classify a chunk as belonging to a PTE or otherwise. We compute the temporal gradients at the first frame via forward difference, second frame via central difference and third frame via backward difference to obtain as accurate gradients as possible using 3 frames. Let the average of all temporal gradient magnitudes in each frame be denoted by $\bar{m}_1, \bar{m}_2$ and $\bar{m}_3$. If the standard deviation of these 3 values is greater than a fixed threshold τ_1, we mark the chunk as lying within a PTE and remove it.

2.2. Step 2: Candidate Frame Extraction (CFE)

Even after removing PTEs, the remaining stationary chunks of frames are of two types: with[1] and without hand presence. The former need to be removed and amongst the latter only one frame needs to be selected between any two PTEs. Each of these frames is called a candidate frame (CF). For both cases, let f and l denote the first and last frames of a chunk. Then the middle frame is obtained as $m = \lceil \frac{f+l}{2} \rceil$.

1. If the user pauses without removing his/her hand.

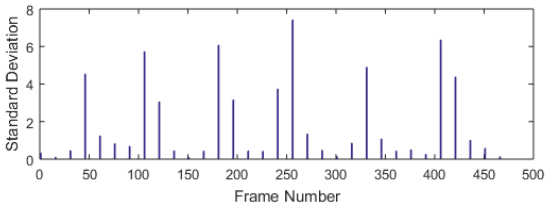


Figure 2. Chunks of frames corresponding to page turn events have a large standard deviation amongst their temporal gradients. This can be used to filter them out. This video contained 6 page turns.

If percentage of skin pixels in frame m is greater than a fixed threshold τ_2, the whole chunk is dropped. Otherwise frame m is selected as a CF. All CFs are passed onto the next module.

Since the hand detection problem is not the primary focus, we use a simple skin-color-based hand detector. Other more sophisticated hand detection modules [9], [10], [11], [12], [13] can be used to replace ours without affecting the overall pipeline.

2.3. Step 3: Duplicate Page Removal (DPR)

It is possible[2] for extracted CFs to represent the same page. To remove such duplicate pages, we compare pairs of CFs in sequence. After pre-computing SIFT descriptors [14] for each CF, we pick the first two CFs. We compute the percentage p_1 of descriptors in the first CF that match those in the second CF. Similarly, we compute the percentage p_2 of descriptors in the second CF that match those in the first CF. If the average matching percentage $\frac{p_1+p_2}{2}$ exceeds a fixed threshold τ_3, we remove the second CF and proceed towards the comparison between the first and the third CFs. Otherwise, we retain both and proceed towards the comparison between the second and third CFs. These comparisons are performed until all CFs are processed. Any frames remaining after this step are the key frames representing unique page images in the video. SIFT computation and matching was performed via the original code http://www.cs.ubc.ca/~lowe/keypoints/siftDemoV4.zip.

3. Dataset and Ground Truth

Figure 1 shows a simple, inexpensive physical setup that we used for data collection. Our equipment for video recording consisted of a camera enabled cell-phone (Samsung Galaxy Grand Prime) and a cell-phone mount. We recorded a total of 37 videos containing 763 PTEs. Books were recorded in 35 videos while $A4$ size paper documents were recorded in the remaining 2 videos. Some videos representing non-English language books contained left-to-right page turns. One of the $A4$ videos represented a stapled document and one unstapled. Each video was recorded at a resolution of 1920×1080 pixels but, for faster processing, were resized

2. For example, if the user performs half a page turn and goes back or there is page movement due to air or the document's rigidity.

to 25% of that size. No special illumination source was used during video recordings. Effort was made to perform most page turns in a normal manner and speed even though the presented algorithm is robust to atypical page turns. Figure 1 shows 4 sample frames of a page turn event. Figure 3 shows a variety of page turn methods and challenging cases in our dataset. Each video was annotated with ground-truth data containing i) the total number of pages and ii) for each desired page image, a valid range of key frames. The dataset and ground-truth can be downloaded from http://faculty.pucit.edu.pk/nazarkhan/datasets/pucit_page_turns.zip.

4. Evaluation Criteria

A straight forward evaluation criterion for any page extractor would be the F_1-measure

$$F_1 = 2\frac{PR}{P+R} \tag{1}$$

where P and R represent the precision and recall

$$P = \frac{tp}{tp+fp} = \frac{\text{valid}}{\text{valid}+\text{invalid}} \tag{2}$$

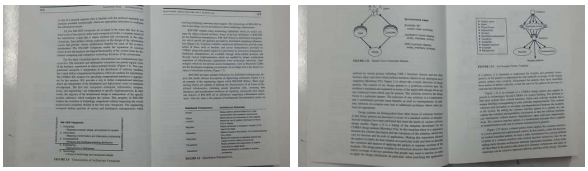
$$R = \frac{tp}{tp+fn} = \frac{\text{valid}}{\text{valid}+\text{missed}} \tag{3}$$

computed from the number of true positive page detections (tp), the number of false positives (fp) and the number of false negatives (fn). $F_1 = 0$ corresponds to the worst possible page extractor while $F_1 = 1$ corresponds to the best possible.

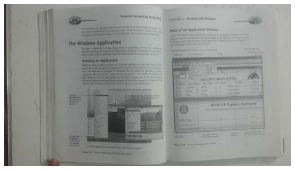
Weakness of F_1-measure The F_1-measure is a hard, binary evaluation criterion in the sense that a detected page frame is either correct or incorrect. For example, a detected key frame *just one frame* outside a human marked valid range will be treated as a false positive even though it might not contain any significant movement or hand presence. Such hard penalties become problematic considering the fact that in videos, temporal ground truths marked by humans can be somewhat subjective. For instance, Figure 4 shows some detected key frames that lie within but near the boundary of an invalid range as determined subjectively by the ground truth annotator. Despite representing valid, stationary and clear page images, such frames will be considered as false positives in the F_1-measure. In other words, the F_1-measure drops when reasonably acceptable detections happen to lie in invalid ranges just because of the subjective nature of human ground-truthing attempts.

Therefore, for the video-based page extraction task and its associated ground truth marking, we propose an alternative, soft criterion that assigns an adaptive, positive score to reasonably acceptable false positives instead of ignoring them completely as in the F_1-measure.

Soft evaluation criterion For a video with ground truth markings, let $\mathcal{V}_k$ denote the valid range of frames within which the k-th page image is allowed. Let $\mathcal{I}_k$ be the invalid range of frames between $\mathcal{V}_k$ and $\mathcal{V}_{k+1}$. Then $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \dots\}$ denotes the set of inter-PTE, valid frame ranges and $|\mathcal{V}|$ equals the ground truth estimate of the number of pages seen in the video. The set and number of detected key frames is denoted by $\mathcal{D}$ and $|\mathcal{D}|$ respectively. Any key frame detected within any range in set $\mathcal{V}$ will be considered as a true positive. Similar to the valid case, $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \dots\}$ denotes the invalid, intra-PTE frame ranges. For example, for a 100 frame video with a single PTE between frames 40 and 70, $\mathcal{V}_1 = [1..39]$, $\mathcal{I}_1 = [40..70]$, $\mathcal{V}_2 = [71..100]$, $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2\}$ and $\mathcal{I} = \{\mathcal{I}_1\}$. The corresponding ground truth representation is shown in Figure 5.

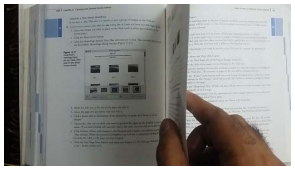


(a) Frames from handheld camera's video. Notice variation in viewing area.



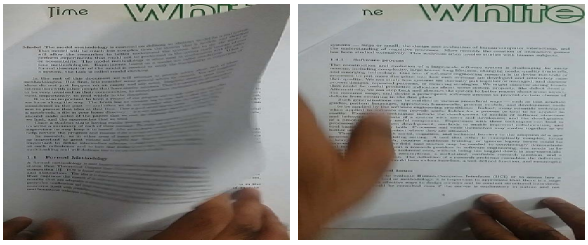
(b) Continuous hand presence to hold the book down.

(c) Left to right page turns.

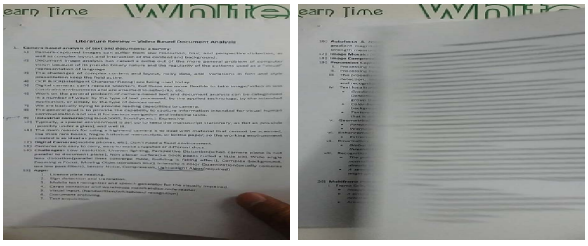


(d) Paused and aborted page turn. Notice absence of motion blur.

(e) Temporary hand presence for book flattening.



(f) Stapled document with bottom right to top left page turns.



(g) Unstapled, unbound pages removed from bottom right. No page turns.

Figure 3. Challenging cases and variations in the dataset.

TABLE 1. SOME SPECIAL CASES IN THE DATASET.

Description	Video No.	Figure
Handheld camera (no cell phone mount).	7	Figure 3a
Continuous hand presence to hold the book down.	12	Figure 3b
Left to right page turns.	14, 18, 23	Figure 3c
Paused and aborted page turn.	16, 21, 26	Figure 3d
Page layout correction.	24	Figure 3e
Stapled document with bottom right to top left page turns.	36	Figure 3f
Unstapled, unbound pages removed from bottom right. No page turns.	37	Figure 3g

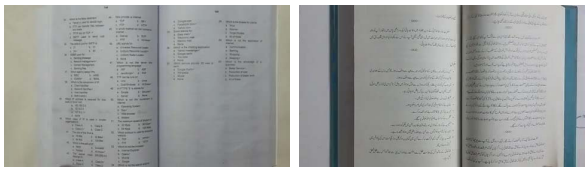


Figure 4. Detected key frames that lie within but near the boundary of a subjectively marked invalid range. Despite representing valid, stationary and clear page images, such frames will be considered as false positives in the F_1-measure. Our soft evaluation criterion assigns a positive score to such frames.

Let $d(\mathcal{V}_k)$ be the number of key frames detected within the k-th valid range and similarly for $d(\mathcal{I}_k)$. The score for the k-th valid range $\mathcal{V}_k$ can be computed as

$$S(\mathcal{V}_k) = \begin{cases} 0 & \text{if } d(\mathcal{V}_k) \equiv 0 \\ 1 & \text{if } d(\mathcal{V}_k) \equiv 1 \\ 1 - \frac{d(\mathcal{V}_k)}{|\mathcal{V}_k|} & \text{if } d(\mathcal{V}_k) > 1 \end{cases} \tag{4}$$

which ranges from 0 for no detection to 1 for 1 detection and becomes less than 1 for multiple detections. This is because the ideal detector should detect 1 and only 1 key frame within a valid range. With these definitions, we can define a soft score from all valid ranges as

$$S_{\mathcal{V}} = \sum_{\mathcal{V}} S(\mathcal{V}_k) \tag{5}$$

It can be verified that $0 \leq S_{\mathcal{V}} \leq |\mathcal{V}|$.

In addition to computing a score from valid ranges, we compute one from invalid ranges too. This is done to counteract the subjective nature of human marked ground truth. To see this, consider if in our earlier example of the 100 frame video, a detector marked frames $40, 55$ and 70 as key frames. These frames correspond to the start, middle and end of the page turn event. Clearly, marking 40 and 70 as key frames should be penalized less severely than 55 since they could have easily qualified as valid frames in a different human based ground truthing attempt[3].

Therefore, we also use key frames detected within invalid ranges to compute our soft score. This happens only when no key frame is detected in a valid range, in which case we check the adjacent invalid ranges. The score depends on two quantities of the detected key frame: i) location based score α_i which quantifies how far into the invalid range it lies, and ii) hand involvement score h_i. Both computations are described next.

If no key frame is detected in $\mathcal{V}_k$, then the system checks if any frame was detected in the adjacent invalid ranges $\mathcal{I}_{k-1}$ and $\mathcal{I}_k$. For $\mathcal{I}_{k-1}$ only the second half of its range is checked since the first half represents the previous page. Similarly, for $\mathcal{I}_k$ only the first half is checked. Therefore

$$\alpha_i = \frac{|m-i|}{m-f} \text{ where } i \in \begin{cases} \{m, \ldots, l\} & \text{for } \mathcal{I}_{k-1} \\ \{f, \ldots, m\} & \text{for } \mathcal{I}_k \end{cases} \tag{6}$$

and f, m and l are the first, middle and last frame numbers in the invalid range being checked ($\mathcal{I}_{k-1}$ or $\mathcal{I}_k$) and i is the detected frame number. It can be verified that $0 \leq \alpha_i \leq 1, \alpha_m = 0$ and $\alpha_f = \alpha_l = 1$. This is visualized in Figure 5.

The hand involvement score h_i is simply the percentage of skin pixels in the frame. This score is weighted by α_i and if there are multiple detections within range $\mathcal{I}_k$, we take their average. That is,

$$h_k = \frac{1}{d(\mathcal{I}_k)} \sum_{i \in \mathcal{D}(\mathcal{I}_k)} \alpha_i h_i \tag{7}$$

Normalized soft score for invalid range $\mathcal{I}_k$ is then calculated as

$$S(\mathcal{I}_k) = \max\left(\left(h_k - \frac{d(\mathcal{I}_k)}{|\mathcal{I}_k|}\right), 0\right) \tag{8}$$

where the $\max$ operator avoids negative scores. For all invalid ranges, the cumulative soft score can be computed as

$$S_{\mathcal{I}} = \sum_{\mathcal{I}} S(\mathcal{I}_k) \tag{9}$$

Finally, we can define our soft evaluation criterion as

$$S = \frac{S_{\mathcal{V}} + S_{\mathcal{I}}}{\max(|\mathcal{D}|, |\mathcal{V}|)} \tag{10}$$

where the $\max$ in the denominator ensures normalization in the case when number of detections $|\mathcal{D}|$ is greater than the number $|\mathcal{V}|$ of ground truth page frames. It can be verified that $0 \leq S \leq 1$.

5. Results

We cross-validated the three thresholds τ_1, τ_2 and τ_3 on one video to obtain their optimal values as shown in Table 2. These threshold values were used for all results shown in this paper.

3. Even involving the same human.

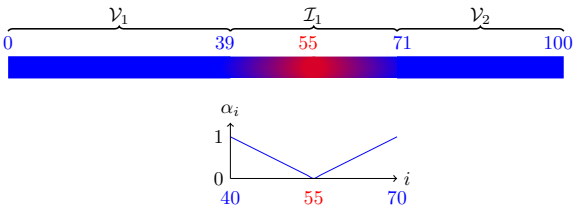


Figure 5. **Top**: Ground truth representing valid and invalid frame ranges. **Bottom**: Location based weight α_i for frames detected in invalid range $\mathcal{I}_1$ with first frame 40, middle frame 55 and last frame 70. Weight decreases to zero as detection approaches the middle of the invalid range from either side. Best viewed in color.

TABLE 2. CROSS-VALIDATED THRESHOLDS USED FOR ALL EXPERIMENTS.

Name	Module	Value
τ_1	PTR	1
τ_2	CFE	1%
τ_3	DPR	20%

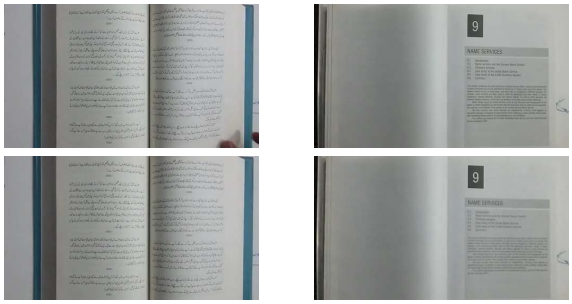
TABLE 3. PERFORMANCE FOR DIFFERENT CHUNK SIZES. USING CHUNKS OF 15 FRAMES GAVE THE BEST RESULTS ON OUR DATASET.

Chunk Size	Precision	Recall	F_1	S
7	0.83	0.91	0.87	0.90
15	**0.92**	**0.90**	**0.91**	**0.94**
22	0.97	0.81	0.88	0.83
30	0.96	0.73	0.83	0.75

In the methodology described so far, we picked non-overlapping chunks of 15 frames each. This number was motivated by the assumption that a typical PTE lasts half a second. In order to choose a more empirically justified chunk size, we compare 4 different chunk sizes equal to a quarter, half, three-quarters and full frame-rate of 30 frames-per-second. Table 3 demonstrates that using chunks of 15 frames was indeed the best choice for videos in the current dataset. For other videos with page turns performed at different speeds, the chunk size might need to be tuned but 15 frames is a reasonably useful starting point. Table 4 presents detailed per-video evaluation using chunks of 15 frames.

Video number 7 was recorded using a handheld camera and therefore contained constant jitter. As a result, chunks of 15 frames usually contained too much movement to be classified as stationary even when they did not represent a PTE. Unsurprisingly, performance on this video dropped. However, as demonstrated in Table 5, using a chunk size of 7 frames resulted in the extraction of all 9 ground truth page frames. This is because keeping a handheld camera stable for 7 frames is easier than for 15 frames and longer.

A similar observation was made for video numbers 31 and 34 which contained rapid page turns. Compared to the low performance for 15 frames, using chunks of 7 frames yielded the best possible values of $F_1 = 1$ as well as $S = 1$ for both videos. These results reinforce our suggestion that different recording conditions require appropriate tuning of parameters.



(a) Video 18. Hand presence and page movement.

(b) Video 24. Auto-focus.

(c) Video 26. Aborted page turn.

Figure 6. Duplicate frames detected as key frames due to (a) slight hand presence and page movement, (b) camera auto-focus, and (c) an aborted page turn attempt.

For video number 18, 4 duplicate frames were detected as key frames due to hand presence and/or page movement. For video 24, duplicates occurred due to camera auto-focus. Video 26 contained an aborted page turn with significant amounts of hand movement and shadow. Figure 6 shows some of these cases.

For all the experiments performed using all chunk sizes, overall Pearson's linear correlation coefficient between F_1 and S was 0.846 which shows that our soft evaluation rewards the same key frame detector properties as the hard evaluation via F_1. This is not surprising since our soft score was designed to mimic F_1. However, as has been explained in Section 4, the soft evaluation avoids the drastic drop in F_1 in the presence of reasonably acceptable false positives.

6. Conclusion

We have presented a low-cost, robust method for video-based document capture that is well-suited for low-volume, personal document digitization tasks. The method is shown to work on a wide variety of document types and page turn methods and extracts unique page frames. We also introduce a publicly available 37 video dataset comprising 763 page turn events with human marked ground truths. Lastly, to deal with subjectivity of human marked ground truth, we introduce a soft evaluation criterion that correlates highly with the traditional F_1 measure while avoiding some of it's weaknesses.

References

[1] "Google Books," https://www.google.com/googlebooks/about/history.html, Accessed: 2017-04-01.

[2] "Project Gutenberg," http://www.gutenberg.org, Accessed: 2017-04-01.

TABLE 4. PER-VIDEO PERFORMANCE EVALUATION USING CHUNKS OF 15 FRAMES.

				Hard						Soft		
Video No.	Duration (s)	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{D}\rvert$	tp	fp	fn	P	R	F_1	$S_\mathcal{V}$	$S_\mathcal{I}$	S
1	119	17	17	11	6	6	0.65	0.65	0.65	11.00	4.75	0.93
2	138	22	24	20	4	2	0.83	0.91	0.87	19.99	1.87	0.91
3	149	17	20	13	7	4	0.65	0.76	0.70	12.99	3.50	0.82
4	121	18	19	15	4	3	0.79	0.83	0.81	14.99	2.72	0.93
5	194	30	32	30	2	0	0.94	1.00	0.97	29.99	0.00	0.94
6	78	11	12	9	3	2	0.75	0.82	0.78	9.00	1.63	0.89
7	100	9	4	4	0	5	1.00	0.44	0.62	4.00	0.00	0.44
8	93	17	18	17	1	0	0.94	1.00	0.97	16.99	0.00	0.94
9	319	45	45	44	1	1	0.98	0.98	0.98	44.00	0.44	0.99
10	121	19	20	19	1	0	0.95	1.00	0.97	18.99	0.00	0.95
11	277	68	69	68	1	0	0.99	1.00	0.99	67.99	0.00	0.99
12	134	20	17	17	0	3	1.00	0.85	0.92	17.00	0.00	0.85
13	162	23	22	20	2	3	0.91	0.87	0.89	20.00	1.80	0.95
14	77	14	15	13	2	1	0.87	0.93	0.90	12.99	0.99	0.93
15	221	50	48	46	2	4	0.96	0.92	0.94	45.99	0.99	0.94
16	91	10	9	9	0	1	1.00	0.90	0.95	9.00	0.00	0.90
17	83	13	13	13	0	0	1.00	1.00	1.00	13.00	0.00	1.00
18	127	30	34	22	12	8	0.65	0.73	0.69	21.96	5.56	0.81
19	104	22	22	22	0	0	1.00	1.00	1.00	22.00	0.00	1.00
20	98	14	14	13	1	1	0.93	0.93	0.93	13.00	0.94	1.00
21	81	16	17	16	1	0	0.94	1.00	0.97	16.00	0.00	0.94
22	63	21	21	21	0	0	1.00	1.00	1.00	21.00	0.00	1.00
23	83	23	20	18	2	5	0.90	0.78	0.84	18.00	2.02	0.87
24	82	15	16	13	3	2	0.81	0.87	0.84	12.98	0.00	0.81
25	91	24	22	22	0	2	1.00	0.92	0.96	22.00	0.00	0.92
26	52	11	14	11	3	0	0.79	1.00	0.88	10.99	0.00	0.79
27	68	26	23	22	1	4	0.96	0.85	0.90	22.00	0.92	0.88
28	62	21	21	19	2	2	0.90	0.90	0.90	19.00	1.89	0.99
29	79	24	23	23	0	1	1.00	0.96	0.98	23.00	0.00	0.96
30	64	21	19	19	0	2	1.00	0.90	0.95	19.00	0.00	0.90
31	83	30	24	24	0	6	1.00	0.80	0.89	24.00	0.00	0.80
32	61	16	14	14	0	2	1.00	0.88	0.93	14.00	0.00	0.88
33	15	7	7	7	0	0	1.00	1.00	1.00	7.00	0.00	1.00
34	27	11	6	6	0	5	1.00	0.55	0.71	6.00	0.00	0.55
35	37	12	11	11	0	1	1.00	0.92	0.96	11.00	0.00	0.92
36	41	7	6	6	0	1	1.00	0.86	0.92	6.00	0.00	0.86
37	49	9	9	9	0	0	1.00	1.00	1.00	9.00	0.00	1.00
Total	**3844**	**763**	**747**	**686**	**61**	**77**	**0.92**	**0.90**	**0.91**	**685.84**	**30.02**	**0.94**

TABLE 5. PERFORMANCE COMPARISON ON VIDEO NUMBER 7 RECORDED USING A HANDHELD CAMERA.

			Hard						Soft		
Chunk size	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{D}\rvert$	tp	fp	fn	P	R	F_1	$S_\mathcal{V}$	$S_\mathcal{I}$	S
7	9	12	9	3	0	0.75	1.00	0.86	8.99	0.00	0.75
15	9	4	4	0	5	1.00	0.44	0.62	4.00	0.00	0.44
22	9	4	4	0	5	1.00	0.44	0.62	4.00	0.00	0.44
30	9	1	1	0	8	1.00	0.11	0.20	1.00	0.00	0.11

[3] G. Nagy, "Twenty years of document image analysis in PAMI," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 38–62, 2000.

[4] 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. 2-3, pp. 84–104, 2005.

[5] Y. Tian and S. G. Narasimhan, "Rectification and 3d reconstruction of curved document images," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 377–384.

[6] N. Stamatopoulos, B. Gatos, I. Pratikakis, and S. J. Perantonis, "Goal-oriented rectification of camera-based document images," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 910–920, April 2011.

[7] W. Newman, C. Dance, A. Taylor, S. Taylor, M. Taylor, and T. Aldhous, "Camworks: a video-based tool for efficient capture from paper source documents," in *Proceedings IEEE International Conference on Multimedia Computing and Systems*, vol. 2, Jul 1999, pp. 647–653 vol.2.

[8] D. Karatzas, V. P. DAndecy, M. Rusiol, A. Chica, and P. P. Vazquez, "Human-document interaction systems – a new frontier for document image analysis," in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, April 2016, pp. 369–374.

[9] J. Kumar, Q. Li, S. Kyal, E. A. Bernal, and R. Bala, "On-the-fly hand detection training with application in egocentric action recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2015, pp. 18–27.

[10] N. H. Do and K. Yanai, "Hand detection and tracking in videos for fine-grained action recognition," in *ACCV Workshops*, 2014, pp. 19–34.

[11] C. Li and K. M. Kitani, "Pixel-level hand detection in ego-centric videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3570–3577.

[12] A. Betancourt, M. M. López, C. S. Regazzoni, and M. Rauterberg, "A sequential classifier for hand detection in the framework of egocentric vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 586–591.

[13] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1949–1957.

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