A Two-Stage Filtering Approach for Video-Based Document Digitization

Shunsuke Kubo[1 [0009−0000−3679−4971]], Cheng Tang[1 [0000−0002−8148−1509]], Tomonori Akashi[1], and Yuta Taniguchi[1 [0000−0003−3298−8124]]

Kyushu University, 744 Motooka, Nishi-ku, Fukuoka, 819-0395, Japan
kubo.syunsuke.585@s.kyushu-u.ac.jp
tang@ait.kyushu-u.ac.jp
akashi.tomonori.950@m.kyushu-u.ac.jp
yuta.taniguchi.y.t@gmail.com

Abstract. Traditional document digitization using specialized scanners is expensive, while manual camera photography is time-consuming. This study proposes a novel two-stage filtering framework for video-based document digitization using a fixed overhead camera to automatically extract static images of pages. The framework combines (1) temporal anomaly detection using the cosine similarity of lightweight CNN features between consecutive frames to identify page-turning events (PTEs) and (2) density-based clustering with OPTICS to group similar frames and eliminate the remaining PTEs as noise. The key innovation is a lightweight implementation that runs on CPUs using pretrained MobileNetV3 features and requires no GPU or additional training. This enables practical deployment in resource-constrained settings. The workflow separates recording from processing, allowing batch processing and parameter adjustments without the need for re-recording. Experiments on four real-world datasets achieved perfect recall (1.0), which means that no pages were lost while maintaining a practical precision. This framework offers a cost-effective alternative for libraries and archives that operate under budgetary constraints.

Keywords: Document Digitization · Video-Based Scanning · Low-Cost Digitization.

1 Introduction

Digitizing books and loose papers keeps fragile originals safe and makes rare content accessible to all. However, specialized document scanners cost tens of thousands of dollars, and manual camera photography requires extensive labor for the positioning, focusing, and reviewing of each page.

A practical alternative is to film the entire page-turning session using an overhead camera. As illustrated in Figure 1, the software automatically detects static frames that show fully settled pages while rejecting blurry frames from the page-turn events (PTEs).

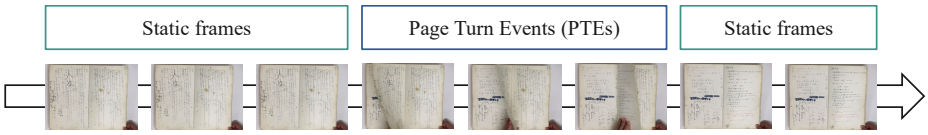


Fig. 1. Basic concept of document digitization from video. The system identifies static frames (left and right) that show stable page views and preserves them. In contrast, frames captured during Page Turn Events (center) are detected and removed.

However, existing methods have several limitations. Chakraborty et al.'s SVM-based approach fails with varying page-turning speeds [3]. Tariq and Khan's RGB difference method is illumination-sensitive and incorrectly deletes visually similar pages like blank pages [10]. Wigington's real-time CNN-LSTM system requires constant user attention and suffers from domain gap issues [11].

We propose a straightforward and robust two-stage filtering framework for PTE detection. In the first stage, anomaly detection is performed using CNN-derived features by comparing the cosine similarity between consecutive frames with a threshold to identify visual changes. The second stage applies density-based clustering [1, 6] to group similar frames, treating small clusters as noise to eliminate the remaining PTEs. This framework uses lightweight CNN features from MobileNetV3 [7], enabling efficient processing on standard laptop CPUs without GPU requirements. The approach is flexible: anomaly detection can be replaced with unsupervised methods such as isolation forest [8] or local outlier factor [2] without retraining, and alternative backbones such as FasterNet [4] can be integrated with minimal changes to the architecture of the model. The middle frame from each cluster is selected as the representative page image, minimizing motion artifacts and hand occlusions typical of overhead camera setups.

This method is effective for libraries and archives with limited funds and time. It requires a laptop and a video camera, keeping costs low. The workflow separates the capture, software processing, and final review stages. Staff members record videos daily, run batch processing overnight, and review the results the following morning. If the results are unsatisfactory, a user can adjust the software parameters and re-run the processing on the original recordings, eliminating the need for re-recording.

Our method achieved perfect recall (1.0) in four diverse test scenarios, which means that no pages were lost. This is critical for archival applications, where missing pages are far more expensive than additional images.

The main contributions of this study are as follows:

1. A lightweight two-stage filtering framework combining anomaly detection and clustering for robust PTE detection without GPU requirements.
2. A practical workflow enabling batch processing and parameter adjustment without re-recording.
3. Perfect recall (1.0) across four test scenarios, demonstrating robust performance in real-world applications.

2 Proposed Method

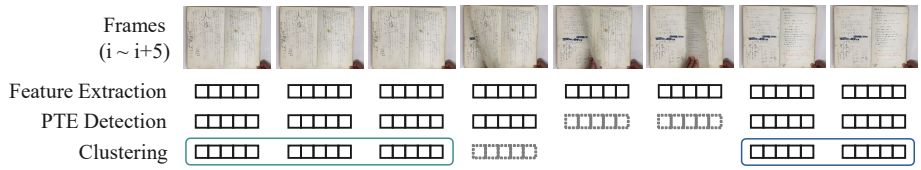


Fig. 2. Overview of the proposed two-stage filtering approach for Page Turn Event (PTE) detection. In the first stage, PTEs are detected based on changes in the cosine similarity using CNN feature vectors from consecutive frames. In the second stage, density-based clustering is employed using the same CNN features to group similar frames into clusters and identify noise frames. In this figure, each box represents a feature vector extracted from the corresponding image frame, and the dotted boxes indicate vectors that are deleted as PTEs in that step.

Our system automatically extracts high-quality page images from videos of book pages turning captured by a fixed overhead camera. The process begins with video recording, followed by automatic software steps: feature extraction, a two-stage filtering process (anomaly detection and clustering), representative frame selection, and postprocessing (hand detection and duplicate removal).

2.1 Frame Feature Extraction

As shown in the top row of Figure 2, the system first extracts frames at regular intervals from the input video. It then converts each frame into a feature vector using MobileNetV3 (Small) [7], a lightweight CNN pretrained on the ImageNet [5]. This model runs efficiently on a CPU without GPU acceleration. The system saves the resulting feature vectors with frame numbers for subsequent processing.

2.2 Anomaly Detection

In the first stage of filtering, the system detects PTEs by analyzing the temporal changes in the feature vectors. As depicted in the "PTE Detection" row of Figure 2, the system calculates the cosine similarity between consecutive frames. If the similarity falls below a predefined threshold, both frames are excluded as they likely capture a page turn. The discarded frames are represented by dotted boxes in the figure. If a page-turning action stops midway (due to a difficult-to-turn page, for example), the system will not recognize it as a PTE; therefore, the frames of that scene will be retained and sent to the next step.

2.3 Clustering of Static Page Frames

In the second stage, illustrated in the "Clustering" row of Figure 2, the system clusters the remaining frames by page using density-based algorithms such as OPTICS [1]. These algorithms group frames corresponding to the same static pages into clusters. Frames that do not meet the minimum number of samples required to form a cluster are classified as noise and removed, eliminating PTEs that escaped first-stage detection.

2.4 Representative Frame Selection

The system selects the temporally middle frame from each cluster as a representative image. This approach avoids frames near the cluster boundaries that may contain partial page transitions, ensuring the selection of stable page views.

2.5 Postprocessing

The selected representative frames are refined through two postprocessing steps: hand detection and duplicate removal.

Hand Detection We eliminate frames with significant hand obstruction. Using MediaPipe Hands [12], we identify hand regions and discard frames where this area exceeds a predefined threshold.

Duplicate Removal We remove redundant frames, which can occur if a single page is split into multiple clusters (e.g., because a hand briefly holds it steady). Using the CNN feature vectors stored during the initial extraction phase, we compute the cosine similarity between pairs of representative frames. A pair with a similarity above a threshold is considered a duplicate, and the frame that appears later is removed.

Crucially, this removal is only applied to frames within a specific temporal proximity (a frame threshold) to avoid the incorrect merging of distinct but visually similar pages, such as consecutive blank pages. This temporal constraint allows our method to preserve legitimate pages that SIFT-based approaches [9, 10] might discard.

3 Experiment

To assess the efficacy of the proposed method, we conducted tests using four different video datasets, each with unique features. As illustrated in Figure 3, these datasets reflect various real-world situations related to document digitization. These datasets vary in terms of page stabilization time, non-page-turning movements, such as using a glass rod, and document format (bound book vs. loose sheets).

(a) Book A (Video 1) **(b)** Book B (Video 2) **(c)** Book C (Video 3) **(d)** Document D (Video 4)

Fig. 3. Sample frames extracted from the four video datasets used in our experiments. Each dataset represents different capturing conditions and document types.

3.1 Datasets

We used four videos captured under different conditions featuring distinct books or documents.

- **Book A (Video 1):** A 47-page book. Each page was kept relatively still for approximately 15 seconds during capture. The total recording time was 15:08.
- **Book B (Video 2):** A 20-page book. Each page was kept still for a short duration of approximately 5 seconds. Total recording time: 3:13.
- **Book C (Video 3):** A 26-page book. Some pages were difficult to keep flat, requiring the supplemental use of a glass rod to hold the top edge of the book down. The total recording time was 6:41.
- **Document D (Video 4):** 12 non-book documents. Each sheet was captured individually by presenting it to the camera, thus involving the most significant motion between pages. The total recording time was 5:06.

3.2 Implementation Details

For clustering, we used the OPTICS algorithm [1]. The key hyperparameters for Video 1 included a frameskip of 5, a minimum of 10 samples for OPTICS, and a 300-frame threshold for duplicate detection. For Videos 2–4, these were adjusted to 1, 7, and 100, respectively. Other parameters, such as the anomaly threshold (99.9%) and duplicate threshold (98%), were kept constant across all datasets.

3.3 Evaluation Metrics

We evaluated the system performance using standard metrics: recall (proportion of correctly extracted pages from the ground truth), precision (proportion of correct pages among system outputs), F1 Score (harmonic mean of precision and recall), and total number of output images. In digital archiving, recall is the most critical metric because missing pages represent irreversible data loss, whereas extra frames can be easily removed during post-processing.

Table 1. Page extraction results for each dataset. Recall achieved 1.0 for all datasets, indicating no missing pages.

Dataset	Ground Truth Pages	Output Images	Recall	Precision	F1
Book A	47	52	1.0	0.90	0.95
Book B	20	22	1.0	0.91	0.95
Book C	26	34	1.0	0.74	0.87
Document D	12	18	1.0	0.67	0.80

3.4 Results

Table 1 summarizes the page extraction results across all datasets. The proposed method achieved perfect recall (1.0) for all datasets, successfully extracting all required pages without omission. This is critical for archival applications, where missing pages represent irreversible data loss.

The precision varied across the datasets and correlated with the amount of non-PTE motion during recording. Book A and Book B achieved high precision (0.90 and 0.91, respectively) under stable recording conditions. Book C showed lower precision (0.74) because glass rod manipulation frames were incorrectly retained as valid pages. Document D had the lowest precision (0.67) because out-of-focus frames during sheet transitions were included in the output. Despite the extra frames, perfect recall ensures complete document preservation, with unnecessary frames easily removable during post-processing.

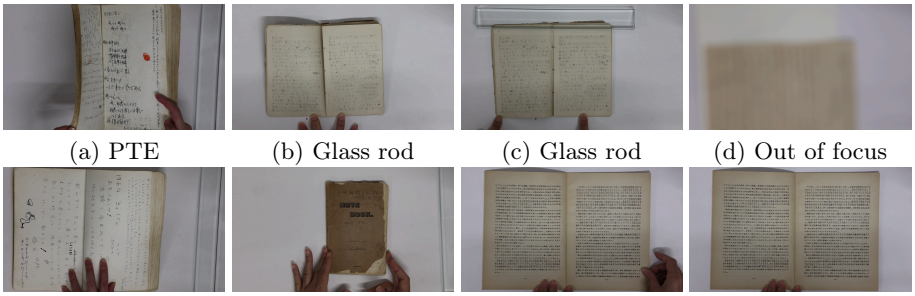
4 Discussion

As illustrated in Figure 4, the low precision stems from three main sources: residual PTEs (a), glass rod manipulation frames (b, c), and out-of-focus frames (d). Our anomaly detection relies on temporal changes in CNN features; thus, scenes with minimal movement pass through if they exceed the minimum clustering sample size.

Hand detection fails when only fingers are visible (Figure 4(e, f)) because MediaPipe cannot detect partial hands. Similarly, duplicate removal fails when identical pages contain different hand positions (g, h), causing misclassification as separate pages.

Videos with extensive non-PTE movements (Book C and Document D) exhibited lower precision. Temporary pauses, glass rod interactions, and document repositioning create frames that confuse the clustering and duplicate removal. While PTE and hand detection were performed correctly, unnecessary frames primarily consisted of duplicates and non-target pages captured during temporary stillness.

Future improvements should address these specific failure modes. Advanced hand segmentation models can detect partial hands and fingers. Blur detection modules would filter out-of-focus frames. Additionally, establishing recording guidelines—quick page turns without pauses and a consistent static page duration—would enhance clustering effectiveness.



(a) PTE (b) Glass rod (c) Glass rod (d) Out of focus

(e) Hand detection (f) Hand detection (g) Duplicate (h) Duplicate

Fig. 4. Examples of failure cases: (a–d) Unnecessary frames retained during page-turning, glass rod usage, and focus loss. (e, f) Hand detection failures with partial finger visibility. (g, h) Duplicate detection failures due to different hand positions.

5 Conclusion

This study proposes a two-stage filtering framework for the automatic extraction of static page images from overhead recordings of page turning. The framework combines temporal anomaly detection using lightweight CNN features with density-based clustering to eliminate PTEs, followed by post-processing for hand detection and duplicate removal. The workflow runs on standard laptop CPUs without GPU requirements, making it well-suited for resource-constrained libraries and archival environments.

Experiments on four real-world datasets achieved perfect recall (1.0), ensuring that no pages were missing. While precision varies with conditions, this trade-off is acceptable for archival practice, where missing pages are far more costly than extra frames are.

Future work will extend the capabilities and scope of the framework. In addition to the immediate enhancements discussed previously, promising research directions include training document-specific CNNs for higher accuracy on particular materials, adapting the system for less constrained scenarios, such as handheld cameras, and exploring multi-frame super-resolution techniques to improve the quality of extracted page images.

Acknowledgments. This study was funded by JSPS KAKENHI Grant Number JP22H00551.

References

1. Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J.: Optics: ordering points to identify the clustering structure. In: Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. p. 49–60. SIGMOD '99, Association for Computing Machinery, New York, NY, USA (1999). https://doi.org/10.1145/304182.304187

2. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. SIGMOD Rec. **29**(2), 93–104 (May 2000). https://doi.org/10.1145/335191.335388
3. Chakraborty, D., Roy, P.P., Saini, R., Alvarez, J.M., Pal, U.: Frame selection for ocr from video stream of book flipping. Multimedia Tools Appl. **77**(1), 985–1008 (Jan 2018). https://doi.org/10.1007/s11042-016-4292-3
4. Chen, J., Kao, S.H., He, H., Zhuo, W., Wen, S., Lee, C.H., Chan, S.H.G.: Run, don't walk: Chasing higher flops for faster neural networks. In: 2023 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR). pp. 12021–12031. IEEE Conference on Computer Vision and Pattern Recognition, IEEE; CVF; IEEE Comp Soc (2023). https://doi.org/10.1109/CVPR52729.2023.01157, iEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, CANADA, JUN 17-24, 2023
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
6. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. p. 226–231. KDD'96, AAAI Press (1996)
7. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for mobilenetv3. In: 2019 IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV 2019). pp. 1314–1324. IEEE International Conference on Computer Vision, IEEE; IEEE Comp Soc; CVF (2019). https://doi.org/10.1109/ICCV.2019.00140, iEEE/CVF International Conference on Computer Vision (ICCV), Seoul, SOUTH KOREA, OCT 27-NOV 02, 2019
8. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining. pp. 413–422 (2008). https://doi.org/10.1109/ICDM.2008.17
9. Lowe, D.: Distinctive image features from scale-invariant keypoints. INTERNATIONAL JOURNAL OF COMPUTER VISION **60**(2), 91–110 (NOV 2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94
10. Tariq, W., Khan, N.: Click-free, video-based document capture - methodology and evaluation. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 06, pp. 21–26 (2017). https://doi.org/10.1109/ICDAR.2017.344
11. Wigington, C.: Handheld video document scanning: A robust on-device model for multi-page document scanning. In: Proceedings of the ACM Symposium on Document Engineering 2024. DocEng '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3685650.3685662
12. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.L., Grundmann, M.: MediaPipe Hands: On-device Real-time Hand Tracking. arXiv e-prints (Jun 2020). https://doi.org/10.48550/arXiv.2006.10214