Near-Duplicate Video Retrieval for Media Archives

Sebastian Binggeli[1], Alain Zanchetta[2], Julia Hartmann[1], Teresa Zulli[3], Monica Bronner[3], Henning Timcke[3], Arzu Çöltekin[2], and Susanne Suter[1]

[1] FHNW School of Computer Science, Institute of Data Science, Windisch, Switzerland
[2] FHNW School of Computer Science, Institute of Interactive Technologies, Windisch, Switzerland
[3] Werft22 - nanoo.tv, Rieden bei Baden, Switzerland

Abstract. We present a human-in-the-loop approach for curating large multimedia archives using near-duplicate video retrieval. A key contribution is a two-stage retrieval strategy that combines metadata-based candidate selection with frame-level similarity analysis, reducing computational complexity while supporting curator-centered workflows. Evaluated on a real-world archive of approximately 9,000 educational videos, the system achieved 90.2% recall and identified previously unknown near-duplicates. The approach provides transparent similarity evidence and content-difference indicators as a basis for downstream curation decisions. More broadly, the work demonstrates the potential of automated similarity analysis for efficient curation of large media collections.

Keywords: Near-duplicate video retrieval, Multimedia curation, Decision-support systems, Media archive management, Video similarity

1 Introduction

In education and similar contexts, multimedia can enhance learner engagement and conceptual understanding [1,2,3]. Consequently, multimedia use in these contexts has increased substantially. However, the volume and diversity of multimedia materials require efficient discovery, curation, and reuse. Existing platforms often fall short in such tasks due to inconsistent metadata, duplicate items, and heterogeneous formats, which hinder searchability and frustrate users. Manual curation of such repositories is prohibitively expensive and requires specialized expertise. In one of our projects, media platform specialists estimated that curating approximately 500,000 multimedia items would require several decades of effort. This motivates machine-assisted archive management. An important task in archive curation is managing duplicated content, a process where *near-duplicate video retrieval* (NDVR) can provide automated support. As there is no single agreed-upon definition of NDVR in the literature [4,5], in this work we define near-duplicate videos as videos depicting the same underlying content despite differences such as language, logo overlays, subtitles, inserted advertisements, or distribution format (e.g., DVD versus TV broadcasts). Although existing NDVR methods achieve high detection performance and scalability, they primarily focus on automated retrieval rather than supporting real-world media curation workflows. This work presents a practical decision-support system for media archive curation that integrates metadata-based retrieval with frame-level NDVR into a scalable workflow while preserving human oversight. Figure 1 summarizes the proposed workflow. The system presents duplicate candidates together with metadata, overlap information, and non-matching video segments, supporting curator assessment while retaining human control over the final curation decision.

2 Related Work

NDVR is often reliant on local feature extraction in video frames [4]. Traditionally, engineered features like the Scale Invariant Feature Transform (SIFT) [7] and Speeded Up Robust Features (SURF) [8] have been employed. These computed features are then aggregated to form global representations on a frame or video level, for example, using Bag of Visual Words (BoV) [9], Fisher vectors [10], or hashing algorithms [11]. The capability of convolutional neural networks (CNNs) to extract features from images has also been leveraged for NDVR. As early as 2016, Jiang and Wang [12] used a Siamese CNN, trained on transformed copy pairs and non-copy pairs, using the contrastive loss function on the Euclidean distance between the output neurons. They then used the activations of the penultimate fully connected layer as features, utilizing BoV to match frames and aligning the matches temporally by framing the problem as a network flow optimization process [12]. Also using a Siamese CNN, in a recent study, Fojcik et al. [13] first subsampled the video frames, selecting only frames at local maxima of the inter-frame difference curve and compared the output with the Euclidean distance, classifying them as similar or not. With this approach, Fojcik et al. [13] achieve competitive results on a much lower input resolution of 56×56 pixels compared to most previous methods. However, the method restricted the length of each video fragment to 15 extracted frames, presenting a limitation for longer videos. Rather than employing only the output of a neural network, Kordopatis-Zilos et al. [14] aggregated intermediate layers of CNNs at various depths to construct a frame representation that preserves local structure at multiple scales. The feature vectors are constructed by max-pooling the activations channel-wise for each layer, and aggregated to a frame-level histogram with BoV. However, to construct the video representation, the histograms are summed, losing temporal structure. For video similarity or near-duplicate assessment, we based our work on the subsequent work by Kordopatis-Zilos et al.'s ViSiL [6], a video similarity learning architecture that calculates video-to-video similarity from refined frame-to-frame similarity matrices to consider fine-grained spatiotemporal relations, using a CNN-based approach and Chamfer similarity [15]. As the media archives can be large, it would be computationally expensive and thus not feasible to compute the similarity of all videos. Therefore, we opted to use metadata to search for similar videos using a knowledge-distilled [16] Universal Sentence Encoder model [17] and compare video content only within a small neighborhood of similar videos based on metadata.

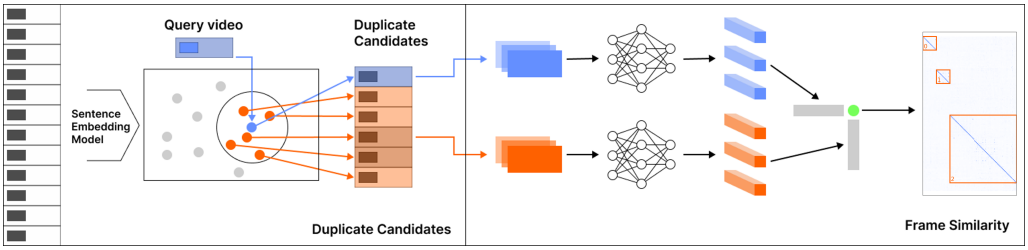


Fig. 1. Two-step NDVR process: First, we identify potential near-duplicate media items in the collection by computing a distance measure between sentence embeddings derived from their metadata. Second, representative frames are selected by a shot detector to compare video similarity, building on components of the ViSiL architecture [6]. Near-duplicate media items are identified by the presence of clear diagonal structures in the frame-to-frame similarity matrix.

3 Methods

We identify near-duplicate videos using a two-step pipeline (Fig. 1): (1) metadata embeddings retrieve *metadata candidates*, and (2) these candidate videos are compared using frame similarity. The resulting duplicate score corresponds to the percentage of matching frames.

3.1 Determine Duplicate Candidates From Metadata

Each video was represented by a sentence embedding computed from the concatenated metadata fields (*title*, *secondary title*, *description*, and *actors* in our dataset). Missing fields were omitted rather than imputed. Sentence embeddings were generated with the multilingual DistilUSE model [18,17] and indexed using ChromaDB [19]. Metadata candidate videos were retrieved by nearest-neighbor search using squared l^2 distance, returning all videos below a distance threshold.

3.2 Identify Near-Duplicates Based on Frame Similarity

In the second step, video near-duplicates from the metadata candidate list were identified using similarity calculated from video frames. Representative frames were extracted using a simple shot detector. A new frame was selected whenever at least 40% of pixels changed between consecutive frames. If no transition occurred for 60 seconds, an additional frame was extracted to ensure temporal coverage (Fig. 2). To calculate the frame similarity,



Fig. 2. Frames extracted by the shot detector from the film *Die göttliche Ordnung* [20].

we mostly built on the ViSiL architecture [6]. From an initial resolution of 480×270 pixels, the frames were resized to 256×256 and then cropped to 224×224 as the effectiveness of this approach was shown by the precursor work ViSiL. Frame features were extracted using the ViSiL feature extraction pipeline [6] based on ResNet50 pretrained on ImageNet-1K: activation maps from multiple layers were pooled, normalized, concatenated, and PCA-whitened. Unlike ViSiL, we did not use their attention mechanism here, avoiding the need for labeled training data.

To calculate the similarity between two videos, we compute the Chamfer similarity [6] for each pair of extracted frames, based on the pairwise dot-product similarities between their feature vectors. This yields a similarity matrix of size $N \times M$, where N and M denote the number of frames extracted by the shot detector from the first and second video, respectively. Near-duplicate videos produce diagonal structures in the similarity matrix. Rather than using ViSiL's CNN, we detect these diagonals with classical image-processing techniques, yielding a simpler and more interpretable pipeline. During diagonal analysis, the similarity matrix was filtered with a 3×3 diagonal kernel to suppress checkerboard artifacts while preserving diagonal structures corresponding

to near-duplicate sequences. It was then binarized using the minimum threshold algorithm [21], with Otsu's method used as a fallback when convergence failed. To group pixels belonging to the same diagonal, we used a modified connected component labeling (CCL) algorithm. Instead of the standard 4- or 8-neighborhood, the neighborhood of a pixel at position (i, j) was defined as the 4×4 region whose bottom-right corner is (i, j). Thus, all pixels $(i - \Delta i,\ j - \Delta j)$ with $\Delta i, \Delta j \in \{0, \ldots, 3\}$ were considered connected. This asymmetric neighborhood bridges gaps of up to two frames caused by inaccuracies in shot detection. Single-pixel and purely horizontal or vertical components were discarded. After identifying the diagonals, the overlap between the videos was determined using the shot detector's timestamp lookup table. Because overlap is directional, similarity is asymmetric. For example, if one video contains a movie and a modified video contains the same movie with an additional advertisement break, the original video will fully overlap with the modified one, while the modified one will not overlap with the original during the advertisement interval.

4 Experiments

4.1 Data

We evaluated our approach on the media archive of a Swiss cantonal secondary school hosted on a media-streaming platform. The archive contains recorded TV broadcasts, feature films, documentaries, and locally produced educational videos. The reference dataset comprises 8,959 unique videos: 5,970 uploaded (67%), 2,387 TV recordings (27%), 589 from online sources (7%), and two screen/camera recordings. For each video, we used *title*, *secondary title*, *description*, and *contributors* fields. Description was available for 6,016 videos (67%). They contain 107.7 ± 84.2 tokens on average [22]. Fig. 3 shows the distribution of the tokenized descriptions lengths, including peaks caused by repeated descriptions; 1,407 descriptions occur at least twice. Contributor metadata is divided into actors, creators, and directors. A total of 2,228 (25%) list at least one contributor, with averages of 4.55 ± 2.26 actors, 1.70 ± 1.34 creators, and 1.21 ± 0.49 directors. For benchmarking, an expert media curator selected 187 known near-duplicate videos, corresponding to 68 unique near-duplicate titles because some videos had multiple duplicates. Description metadata was available for 159 of these videos (85%).

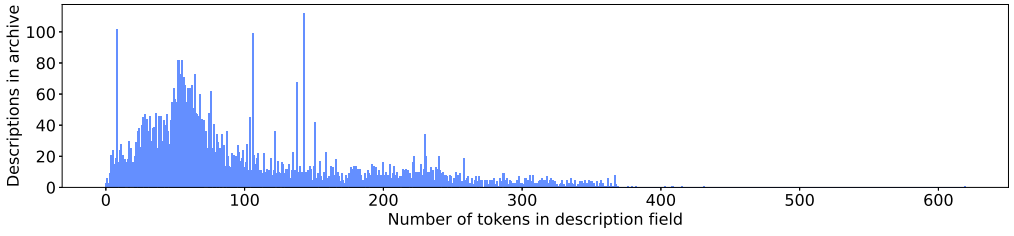


Fig. 3. Histogram of metadata description lengths in the reference archive.

4.2 Duplicate Candidates Based on Metadata

Fig. 4 shows the number of metadata candidates returned for each of the 187 labeled near-duplicate query videos at different metadata distance thresholds. Few candidates were

returned below a threshold of 0.75, whereas the number increased rapidly beyond 1.0. At a threshold of about 1.7, nearly the entire reference dataset was retrieved. Increasing the distance threshold improved recall at the expense of precision (Fig. 5). Because missed candidates cannot be recovered and false positives can be discarded during the subsequent frame-similarity stage, recall was prioritized over precision.

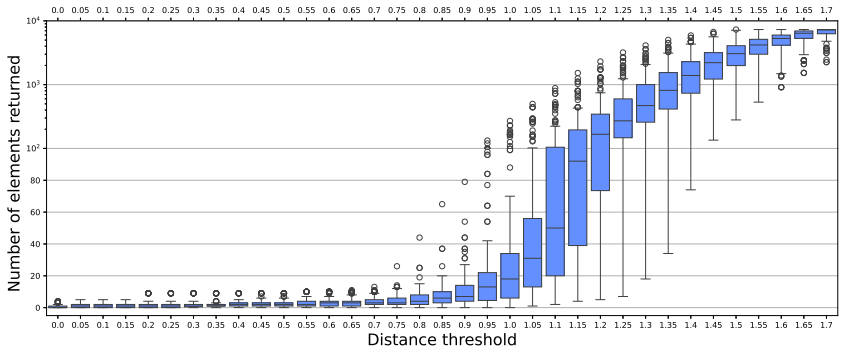


Fig. 4. Number of metadata candidates returned for the 187 labeled near-duplicate query videos at metadata embedding-distance thresholds. The results support selecting a threshold of 0.8.

Considering the precision-recall trade-off, we selected a distance threshold of 0.8, where recall begins to plateau while the number of retrieved metadata candidates remains low. At this threshold, the metadata retrieval achieved a recall of 90.2% while returning on average 5.7 ± 5.2 candidate videos per query from the combined dataset of approximately 9,000 videos. Manual inspection showed that many apparent false positives were previously unlabeled near-duplicates. After relabeling videos with at least 60% frame overlap, precision increased at low distance thresholds while recall remained largely unchanged (Fig. 5), supporting the selected threshold of 0.8. We additionally evaluated an ablation using only description metadata to determine whether semantic information alone was sufficient for candidate retrieval. Because descriptions were unavailable for some videos, this experiment used 6,167 reference videos and 159 annotated near-duplicates. This produced similar precision and recall curves, indicating that DistilUSE captures semantic similarity beyond matching titles and actors.

4.3 Frame Similarity

Frame similarity was evaluated on the 68 groups of labeled near-duplicate videos. Considering all pairwise combinations within each group, the method identified 93.6% of pairs as near-duplicates using a 60% overlap threshold. Manual inspection showed that 8 of the 12 apparent false negatives were incorrectly labeled in the ground truth. The remaining four showed overlap scores of 20%–58%, compared with below 6% for non-duplicate pairs.

4.4 Combined System

We applied frame similarity only to metadata candidates with embedding distance ≤ 0.8. Videos with a frame overlap of at least 60% were classified as near-duplicates. Each of

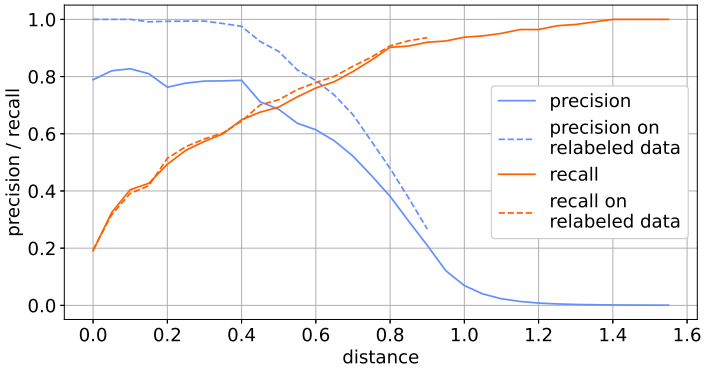


Fig. 5. Precision and recall as functions of the metadata embedding-distance threshold. Relabeling previously unknown near-duplicates improves precision at low thresholds.

the 187 annotated near-duplicates served once as the query video. The combined system correctly assigned 178 of 187 query videos to their near-duplicate group. It also discovered 13 previously unlabeled near-duplicates in the reference archive. Manual inspection confirmed that these were non-trivial near-duplicates, for example a movie with the metadata in a different language. No false positives were observed.

4.5 Recognizing Non-Original Content

Frame similarity showed also promise in localizing non-original content such as advertisements. Figure 6 shows a DVD–TV comparison, where drops in frame similarity correspond to advertisement breaks and additional broadcast material. Non-original segments were detected by thresholding the frame similarity over time.

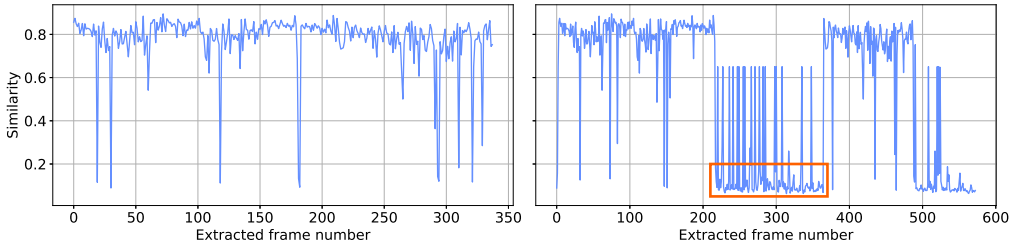


Fig. 6. Frame similarity for a duplicate movie pair. The TV version (right) shows reduced similarity during the advertisement break (orange) and after the credits.

5 Discussion

This work presents a practical architecture for near-duplicate video retrieval in media archives. Rather than performing computationally expensive frame-level comparisons

across an entire archive, the proposed two-stage workflow first retrieves duplicate candidates using metadata embeddings and subsequently verifies them through frame similarity analysis. Together with shot-based frame selection, this substantially reduced the computational requirements: metadata retrieval reduced the search space by a factor of 1,580 while frame selection reduced the number of processed frames by a factor of 280. Despite these reductions, the approach maintained high recall, making near-duplicate detection feasible for large-scale media archive curation.

While previous NDVR research has focused on improving retrieval algorithms and benchmark performance [6,13,23], this work emphasizes system integration for practical media archive curation. The proposed workflow is intended as a decision-support system rather than a fully automated curation pipeline. Duplicate videos are presented together with their metadata, estimated overlap, and detected content differences, allowing curators to determine which version should be retained or presented in the archive. The proposed user interface (Fig. 7) supports this workflow by presenting transparent similarity evidence. Besides identifying near-duplicates, the frame similarity stage can also reveal non-original content, such as advertisements or additional broadcast segments, enabling semi-automatic improvement of archive quality whenever different versions of the same content are available.

This additional curation utility is a key advantage of the proposed approach, which identifies matching and non-matching video sections, not only a single video-level similarity score. It is largely enabled by the selection of representative frames at shot boundaries. Compared with the uniform frame sampling employed by previous NDVR systems [6,23], the selected key frames produced frame-to-frame similarity matrices with more pronounced diagonal structures. This allowed near-duplicate sequences to be detected using classical image-processing techniques instead of the learned similarity network employed by ViSiL [6], thereby simplifying the architecture while maintaining effective near-duplicate detection. An attractive deployment strategy is to perform frame selection during the media platform's transcoding process, allowing representative frames to be extracted once during upload and reused for subsequent retrieval.

The first stage complements the frame analysis by narrowing the set of candidate videos using sentence embeddings derived from video metadata. Although archive metadata are often incomplete or inconsistent, the approach performed well even when using only video descriptions and successfully retrieved near-duplicates across different languages and substantially different descriptions. The metadata distance threshold is, however, a key hyperparameter that depends on the characteristics of the target archive. While a threshold of 0.8 provided a suitable balance between recall and computational cost for the evaluated media archive, its transferability to other collections remains to be investigated.

Several other limitations remain. Variations in frame rate, spatial cropping, mirroring, compression artifacts, and other real-world transformations may affect both shot detection and frame similarity computation. Although the proposed system operated effectively on frames extracted at the lowest available platform resolution (480×270, resized to 224×224), recent work suggests that NDVR can remain effective even at substantially lower resolutions [13]. Future work should investigate the robustness of the proposed workflow under such conditions and explore alternatives that operate directly on video frame embeddings instead of metadata-based candidate retrieval. Beyond these technical limitations, the human-in-the-loop aspect was evaluated only at the workflow level. Although the proposed interface was designed to support curator decisions by presenting transparent similarity evidence and content differences, we did not conduct a user

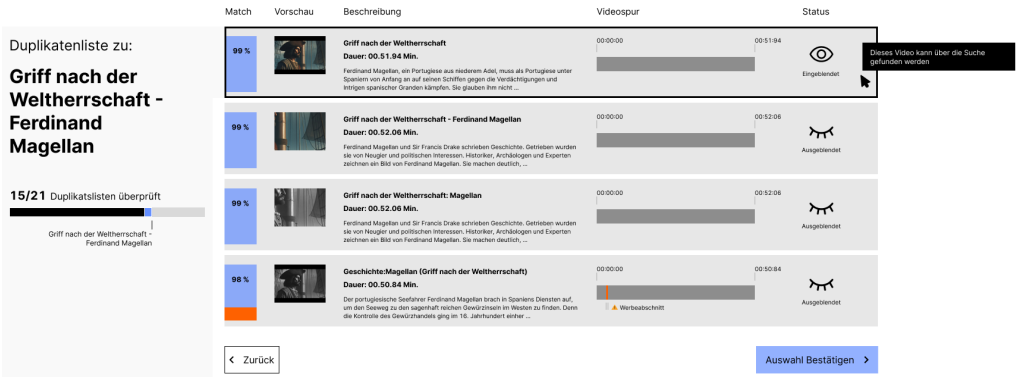


Fig. 7. Conceptual user interface for curator-assisted management of near-duplicate videos. The interface presents metadata, overlap information, and content differences to support representative selection while preserving human oversight.

study to assess its effects on usability, workload, trust, decision quality, or task completion time. Future work will therefore include controlled evaluations with media curators to assess the benefits of the proposed decision-support workflow.

6 Conclusion

This paper presents a two-stage near-duplicate video retrieval system that integrates multilingual sentence embeddings for metadata-based candidate retrieval with frame similarity analysis. Combined with shot-based frame sampling, the architecture enables scalable media archive curation while preserving human oversight. Beyond near-duplicate detection, the frame similarity stage can reveal non-original content, such as advertisements or additional broadcast segments, thereby supporting archive quality improvement.

7 Acknowledgements

This work was supported by the Innosuisse Innovation Cheque programme (INNO-ICT 61448.1), the Hightech Zentrum Aargau (Project 10.00675.03.0), and the Aargauer Forschungsfonds (Project 20230331_10_NICE). An AI-based language assistant (ChatGPT, OpenAI) was used to support editing and refinement of the text. Scientific content, interpretations, and conclusions were developed by the authors.

References

1. Brame, C.J.: Effective educational videos: Principles and guidelines for maximizing student learning from video content. CBE—Life Sciences Education **15**(4) (December 2016) es6
2. More Valencia, R.A., Nizama Reyes, M.E., Lizana Puelles, E.Y., Sandoval Valdiviezo, J.M.: Effectiveness associated with learning with video and multimedia content in engineering students' classroom sessions. **23** (December 2023)
3. Staneviciene, E., Žekienė, G.: The use of multimedia in the teaching and learning process of higher education: A systematic review. Sustainability **17**(19) (October 2025) 8859
4. Liu, J., Huang, Z., Cai, H., Shen, H.T., Ngo, C.W., Wang, W.: Near-duplicate video retrieval: Current research and future trends. ACM Computing Surveys (CSUR) **45**(4) (2013) 1–23

5. Shen, L., Hong, R., Hao, Y.: Advance on large scale near-duplicate video retrieval. Frontiers of Computer Science **14**(5) (January 2020) 145702
6. Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., Kompatsiaris, Y.: ViSiL: Fine-grained spatio-temporal video similarity learning. In: IEEE/CVF International Conference on Computer Vision (ICCV). (2019) 6350–6359
7. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision **60**(2) (November 2004) 91–110
8. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In Leonardis, A., Bischof, H., Pinz, A., eds.: Computer Vision – ECCV, Berlin, Heidelberg, Springer Berlin Heidelberg (2006) 404–417
9. Chou, C.L., Chen, H.T., Lee, S.Y.: Pattern-Based Near-Duplicate Video Retrieval and Localization on Web-Scale Videos. IEEE Transactions on Multimedia **17**(3) (March 2015) 382–395
10. Douze, M., Jégou, H., Schmid, C., Pérez, P.: Compact Video Description for Copy Detection with Precise Temporal Alignment. In Daniilidis, K., Maragos, P., Paragios, N., eds.: Computer Vision – ECCV, Berlin, Heidelberg, Springer (2010) 522–535
11. Yang, G., Chen, N., Jiang, Q.: A robust hashing algorithm based on SURF for video copy detection. Computers & Security **31**(1) (2012) 33–39
12. Jiang, Y.G., Wang, J.: Partial Copy Detection in Videos: A Benchmark and an Evaluation of Popular Methods. IEEE Transactions on Big Data **2**(1) (March 2016) 32–42
13. Fojcik, K., Syga, P., Klonowski, M.: Extremely compact video representation for efficient near-duplicates detection. Pattern Recognition **158** (February 2025) 111016
14. Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., Kompatsiaris, Y.: Near-Duplicate Video Retrieval by Aggregating Intermediate CNN Layers. In Amsaleg, L., Guomundsson, G.D., Gurrin, C., Jonsson, B.D., Satoh, S., eds.: MultiMedia Modeling, Cham, Springer International Publishing (2017) 251–263
15. Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., Wolf, H.C.: Parametric correspondence and chamfer matching: two new techniques for image matching. In: Proceedings International Joint Conference on Artificial Intelligence - Volume 2. IJCAI'77, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1977) 659–663
16. Reimers, N., Gurevych, I.: Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. arXiv preprint arXiv:2004.09813 (04 2020)
17. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (November 2019)
18. : distiluse-base-multilingual-cased-v1 (June 2021)
19. Huber, J., Troynikov, A.: ChromaDB (February 2024)
20. Volpe, P.B.: Die göttliche Ordnung (March 2017)
21. van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T., contributors, t.s.i.: scikit-image: image processing in Python. PeerJ **2** (June 2014) e453
22. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC Workshop on New Challenges for NLP Frameworks, Valletta, Malta, ELRA (May 2010) 45–50
23. Tan, W., Guo, H., Liu, R.: A Fast Partial Video Copy Detection Using KNN and Global Feature Database. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). (2022) 459–467