MoReSo

A DNN Framework Expediting Content-based Video Image Retrieval (CBVIR)

Li, Sinian; Profeta, Doruk Barokas; Dauwels, Justin

DOI
10.23919/EUSIPCO63174.2024.10715173

Publication date
2024

Document Version
Final published version

Published in
32nd European Signal Processing Conference, EUSIPCO 2024 - Proceedings

MoReSo: A DNN Framework Expediting Content-based Video Image Retrieval (CBVIR)

Sinian Li, Doruk Barokas Profeta, and Justin Dauwels
Signal Processing Systems, Dept. of Microelectronics
Delft University of Technology, Delft, Netherlands
J.H.G.Dauwels@tudelft.nl

***Abstract*—With the exponential growth of video data, individuals, particularly scholars in the fields of history and sociology, are increasingly reliant on video materials. However, the task of locating specific frames within videos remains a laborious and time-consuming endeavor. Advanced machine learning-assisted video processing techniques have emerged, including text-based video searches, video summarization, real-time object detection, and person re-identification. However, distinct from these, the main challenge of retrieving video frames based on given visual content is how to efficiently and accurately pinpoint the instance occurrences. To expedite the process while maintaining retrieval performance, we propose a two-stage approach, combining KeyFrame Extraction (KFE) and Content-based Image Retrieval (CBIR), underpinned a DNN-empowered framework called MoReSo. Our innovations include 1) the integration of improved statistical features with dynamic clustering in the KFE stage and 2) the development of the MoReSo framework, which consists of MobileNet and ResNet backbones with SOA layer to jointly represent video frames, achieving 2.67x increase in efficiency compared to existing solutions. Our framework is evaluated on two datasets: the annotated EHM Historical Database provided by digital history researchers and the widely-used image retrieval benchmark datasets, the Oxford and Paris datasets. The experimental results showcase that the proposed framework and scheme excel among other models in the CBVIR task. We make our code available for further exploration through our GitHub repository. This repository contains the implementation of our model and CBVIR system with a GUI prototype.**

***Index Terms*—Content-Based Video Image Retrieval, Content-Based Image Retrieval, Key Frame Extraction, Image Retrieval from Video**

I. Introduction

Content-based Video Image Retrieval (CBVIR), which aims to pinpoint target visual content within long videos, has been increasingly drawing attention because the advent of streaming and video has sparked a revolutionary shift in the presentation and usage of digital materials across various fields, such as history, sociology, art, and media [1]. But compared with Content-based Image Retrieval (CBIR), which focuses on how to search the given query from an image gallery, CBVIR is a relatively new topic. Most recent progress in enhancing visual content retrieval or searching for a specific instance revolves around CBIR capabilities [2], [3]. The academic literature on video image retrieval remains limited.

CBIR techniques have transitioned from traditional methods like SIFT [4], ORB [5] and AKAZE [6] to more complex architectures of neural networks: VGG [7] and ResNet [8]. Its shift to a learning-based, data-driven approach marked a new level of retrieval accuracy. As deep learning models in image representation expand in complexity, there arises an inherent trade-off between computational time and accuracy [3].

Apart from the problems CBIR faces, video image retrieval faces another challenge: the searching gallery has a great number of repetitive or redundant frames, requiring higher computation power and impeding the processing speed. With the advances in image and pattern matching, automatically searching among video frames is not a daunting task. The simplest idea is to extract the video into a time-indexed image library given a specific frame rate [9].

Though the literature on solely visual content-based video frame retrieval remains limited, two reviews in CBVIR [9], [10] highlighted the need for shot boundary detection and keyframe selection in managing extensive video frames. Because with the essence of the video, the search in the candidate gallery would be greatly simplified. KFE is an important topic in video summarization and movie highlight generation. These studies focus on employing multimodal algorithms to discern the narrative structure, aiming to produce a synopsis of video [11]. While the efficiency of this module can only be achieved when it scans through frames with relatively straightforward feature representations [12]. VSUMM uses HSV color features to represent frames and k-means clustering algorithm to generate video summaries [13]. Similarly, other researchers exploit a rapid wavelet histogram technique to select another set of keyframes. Two sets are combined using mutual information to produce the final selection of keyframes [14]. Conversely, instead of dealing with visual features, a framework trains the network to associate visual information to textual inputs, allowing for the generation of keyframes in response to textual queries at the expense of efficiency. [15]. The main challenge in KFE for our task lies in balancing low-level and high-level features effectively.

VISIONE is the latest video image retrieval framework for large-scale video search [16]. This paper includes keyword-based search using the Caffe framework and object detection by integrating YOLOv3. For visual search, it extracts visual features and transforms them into textual representations for text search engines. This proposed framework heavily relies on text-based indexing, which loses visual nuances and could potentially struggle with complex content understanding and retrieval. Moreover, for content-based visual media retrieval, a framework leverages 2D and 3D Convolutional Neural

Networks, and Long Short-Term Memory networks (LSTMs) to process images and videos jointly for feature extraction and classification [17]. This approach improves video understanding through LSTM but compromises efficiency by reducing the entire video to a single feature, sacrificing temporal details. Accurate frame capture demands high-level image representations, whereas efficient processing needs simpler, low-level image analysis. The current framework predominantly emphasizes accuracy, using complex but large Convolutional Neural Networks for in-depth video analysis.

In summary, this paper is written with objectives:

- First, to complement and enhance the research in CBVIR and address practical needs, we propose and establish a comprehensive framework that takes a video and a query image as input and retrieves relevant frames from the video with high efficiency and accuracy.
- Second, to address the problem of redundant frames impeding the processing speed, we incorporate and improve the state-of-the-art statistical feature-based KeyFrame Extraction algorithm to facilitate the processing of the overall scheme, emphasizing the balance of features.
- Third, to enhance the efficiency of the CBIR module, we build a novel framework tailored to efficient visual content retrieval from video.

The rest of the paper is organized as follows: Section II introduces our CBVIR scheme, the methods we use in KFE, and our proposed framework, MoReSo, for CBIR. Section III presents experimental settings, results, comparisons, and analyses. Section IV concludes this paper.

II. Methodologies

In this paper, we address the need for a fast and reliable retrieval system by introducing a new CBVIR approach, featuring an enhanced KFE algorithm and the MoReSo framework to boost CBIR efficiency.

A. A novel CBVIR scheme

The proposed CBVIR scheme is the structured pipeline in Fig. 1. The key steps are highlighted in blue boxes. Once video is processed and keyframes are extracted by KFE, MoReSo module will conduct retrieval process. The retrieval process is decomposed into two stages, denoted as the green box and the orange box: 1) fast image representation and shortlisting, filtering out completely unrelated keyframes; 2) refined feature extraction that captures features out of the most relevant candidates so that the ensuing search algorithm can quickly find the target frames.

The idea behind our proposed scheme is that statistical features serve as a foundational yet efficient mechanism for filtering indexes that could encapsulate the long video (containing thousands of frames) into a compact keyframe set. Even though this first step eases the burden on subsequent searches, the retrieval gallery is still sizable. Therefore, the ensuing MoReSo framework is built to mitigate the challenges of intensive and time-consuming feature extraction by taking advantage of compact neural networks without sacrificing accuracy. Shortlisting helps to remove target-irrelevant images from the gallery set. And then the framework leverages the deep network to refine the retrieval results. It ensures that the subsequent searching algorithm rapidly identifies accurate retrieval results. This strategic framework makes the most of low-level features and high-level image representation techniques in video processing.

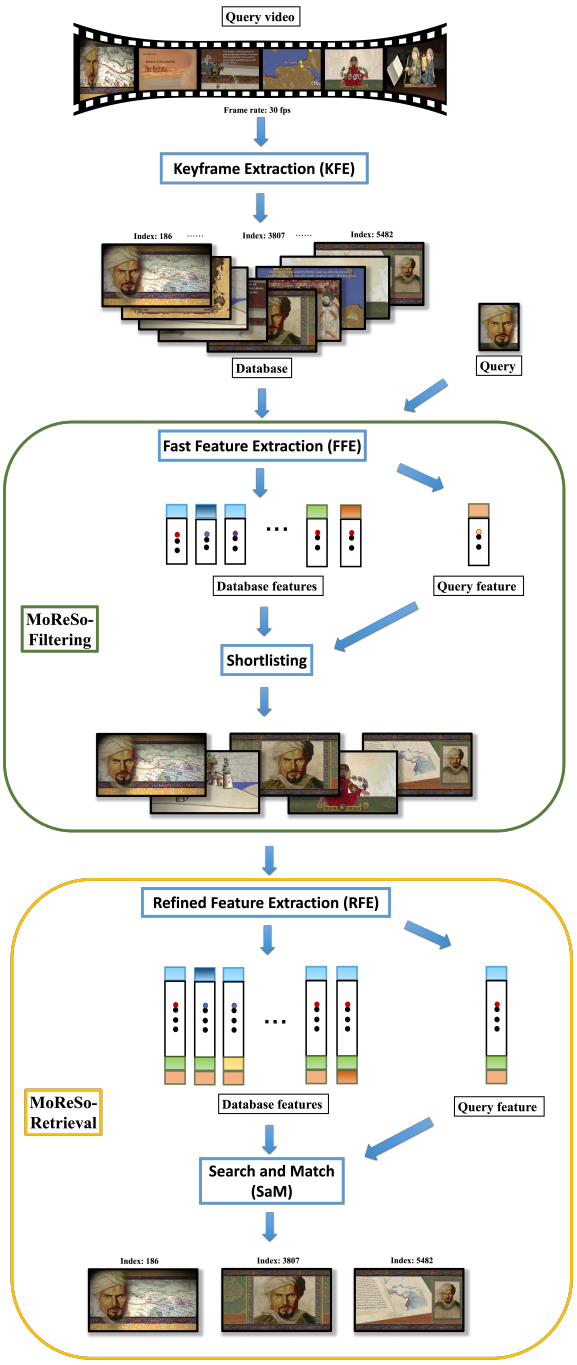

Fig. 1. The pipeline of our proposed CBVIR scheme. The main steps are indicated as the blue boxes.

B. Improved Keyframe Extraction algorithm

In the KFE module, extracted features later serve as cluster references for the system to form keyframe galleries from the video. Both soft and hard transitions in scenes can cause noticeable color statistical changes, which are sufficient to discern the shots or scenes. This appears to be the most economical KFE feature. In this realm, Video SUMMarization (VSUMM) based on HSV color histograms and K-means clustering [13] is the state-of-the-art system. As in Fig. 2, we improved efficiency by further dimension reduction using SubMatrix Selection Singular Value Decomposition and accuracy by dynamic clustering, which does not require predefined hyperparameters and can automatically adjust the cluster assignments as a new feature becomes available.

Given that the color span is divided into L color bins, the color histogram of an image block located at p^{th} row and q^{th} column with $N_{(p,q)}$ pixels can be presented as:

$$h_{(p,q)}(k) = \frac{\eta_{(p,q)}(k)}{N_{(p,q)}}, \qquad k \in \{1, 2, \cdots, L\}, \tag{1}$$

where $\eta_{(p,q)}(k)$ is the $(p,q)^{th}, p \in \{1, 2, \cdots, P\}, q \in \{1, 2, \cdots, Q\}$ total number of pixels in the k^{th} color bin. Therefore, the image feature matrix can be formed as $\mathbf{H}(I)$. And for a more compact expression that can easily be used in later feature comparison, we can vectorize $\mathbf{H}(I)$ into a feature vector $\mathbf{h}(I)$. Moreover, a video consists of many frames, which can be denoted as $\mathbf{I}$. Therefore, we can express the color-based feature matrix as $\mathbf{H}(\mathbf{I})$.

As shown in Fig. 2, after forming the feature matrix, Submatrix SVD is performed to reduce the feature dimension. And by taking the new feature matrix, we implement the dynamic clustering algorithm to cluster the new feature space $\mathbf{V}(\mathbf{I})$. Fig. 3 illustrates the dynamic clustering, keeping and leveraging the temporal information inherent in video frames.

Each feature vector $\mathbf{v}_i$ is examined by a clustering algorithm. Algorithm 1 showcases the implementation of dynamic clustering in grouping frames and detecting keyframes.

The threshold is a decision boundary for how similar two centroid vectors should be. We apply cosine similarity as a measure of similarity:

$$\cos(\theta) = \frac{\boldsymbol{c_1} \cdot \boldsymbol{c_2}}{\|\boldsymbol{c_1}\| \, \|\boldsymbol{c_2}\|} = \frac{\sum_{i=1}^{n} c_1^i c_2^i}{\sqrt{\sum_{i=1}^{n} \left(c_1^i\right)^2} \sqrt{\sum_{i=1}^{n} \left(c_2^i\right)^2}}. \tag{2}$$

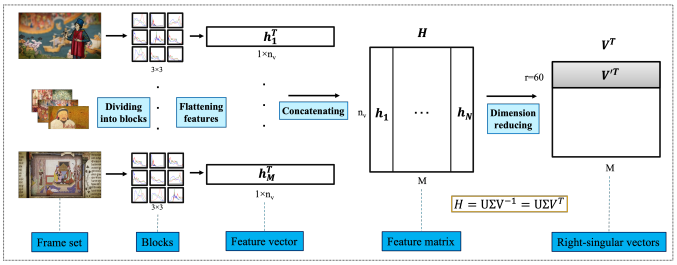

Fig. 2. The pipeline of color-based feature matrix formation

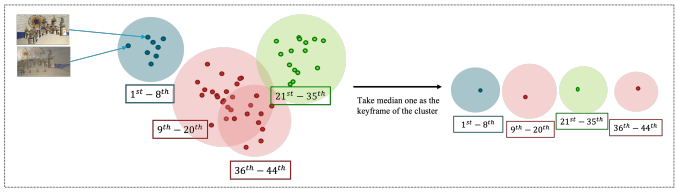

Fig. 3. Illustration of dynamic clustering

Inputs: Feature set: $\mathbf{V}$, Num. of features: N, Threshold: thr
Result: Final centroids: $\mathbf{C}$, Keyframe indices: $\mathbf{I}$
Two initial features: $\mathbf{v}_1, \mathbf{v}_2$;
One initial centroid: $C_0 = \frac{\mathbf{v}_1 + \mathbf{v}_2}{2}$;
$i \leftarrow 2$;
$n \leftarrow 2$;
while $i \leq N$ **do**
  $\mathbf{C}_k^{i+1} \leftarrow \frac{n \cdot \mathbf{C}_k^i + \mathbf{v}i+1}{n+1}$
  if $\cos(\mathbf{C}_k^i, \mathbf{C}_k^{i+1}) \leq thr$ **then**
    $\mathbf{C}_k \leftarrow \mathbf{C}_k^{i+1}$ $n \leftarrow n+1$;
    $i \leftarrow i+1$;
  else
    if $\cos(\mathbf{C}_k^i, \mathbf{C}_k^{i+1}) \geq thr$ **then**
      $\mathbf{C}_k \leftarrow \mathbf{C}_k^i$ $I_k \leftarrow i$ $n \leftarrow 2$;
      $i \leftarrow i+2$;
      $k \leftarrow k+1$;
      $\mathbf{C}_k^i \leftarrow \frac{\mathbf{v}_{i-2} + \mathbf{v}_{i-1}}{2}$
    end
  end
end

Algorithm 1: Dynamic clustering in KFE

C. MoReSo framework

MoReSo is an integrated framework that leverages two networks: MobileNetV2 [18] and ResNet101 with Second Order Attention (SOA) layer [19]. It also includes fast nearest neighbor search technique, ANNOY. Table I introduces the structure of this framework.

After KFE, the keyframe set is processed through pretrained MobileNetV2 to generate the feature maps. MobileNetV2, renowned for its compact structure, outperforms other state-of-the-art networks in terms of computation time, making it the fastest option for video frame search in a CBIR module [20]. These features will be used to build Approximate Nearest Neighbor (ANN) search indexes and to be shortlisted based on the query feature. With the shortlisted set, the Second Order Loss and Attention ResNet101 (SOLAR) architecture leverages second-order information through spatial attention and descriptor similarity to improve large-scale image retrieval.

The global descriptor $\mathbf{d}$ of an input image is obtained by conducting GeM pooling on the feature maps $\mathbf{f}$:

$$\mathbf{d} = \mathrm{GeM}(\mathbf{f}, p) = \left(\frac{1}{N} \sum_{i=0}^{N} f_i^p \right)^{\frac{1}{p}}, \tag{3}$$

TABLE I
THE STRUCTURE AND SPECIFICATION OF MORESO FRAMEWORK

Stage	Method	Structure
Filtering	MobileNet features	Conv2d (224 × 224 × 3) Bottlenecks (112 × 112 × 32 - 7 × 7 × 160) Conv2d (7 × 7 × 320) Avg Pool (7 × 7 × 1280) FC (1280)
	Dynamic Shortlisting	Euclidean distance
Retrieval	ResNet features	Conv2d (224 × 224 × 3) Max Pool (112 × 112 × 64) Residual Blocks (56 × 56 × 64 - 14 × 14 × 1024) Avg Pool (7 × 7 × 2048) FC (2048)
	ANNOY	Cosine distance measure

where p is the pooling parameter. The global descriptor $\mathbf{d}$ has a finite receptive region, resulting in a limited contribution from each local feature. To incorporate spatial information into feature pooling, the SOA layer creates its feature map $\mathbf{z}$ by generating three projections from the feature maps $\mathbf{f}$, corresponding to query $\mathbf{q}$, key $\mathbf{k}$, and value $\mathbf{v}$. This SOA map can be expressed as:

$$\mathbf{z} = softmax(\alpha \cdot \mathbf{q}^\top \mathbf{k}), \tag{4}$$

where α is the scaling factor. This way, feature map $\mathbf{f}$ is able to combine all the local features across the entire map. In details, the employment of the first-order attention (an additional 1×1 convolution) denoted as ϕ supervises the influence of the attention, after which we can get the SOA feature map by:

$$\mathbf{f}^{so} = \mathbf{f} + \Phi(\mathbf{z} \times \mathbf{v}). \tag{5}$$

Within the feature map $\mathbf{f}^{so}$, a new feature is refined as a function g, mapping local features from different positions $\mathbf{z}_{i,j}$ within $\mathbf{f}$ to $f^{so}_{i,j}$, and it is expressed as:

$$f^{so}_{i,j} = g\left(\mathbf{z}_{ij} \odot \mathbf{f}\right). \tag{6}$$

Then, an extended GeM pooling is :

$$\mathrm{GeM}\left(\mathbf{f}^{so}, p\right) = \left(\frac{1}{N}\sum_{i=0}^{N} f_i^{so^p}\right)^{\frac{1}{p}}. \tag{7}$$

This module includes fine-tuning the SOAs and the whitening layer using the Google Landmark 18 [21] dataset, training with the triplet loss, and utilizing SOSNet [22] for local descriptor learning. The utilization of this model aims to guarantee accuracy in the refinement stage.

III. EXPERIMENTS

In this section, we apply our individual module, framework, and overall system to three datasets for evaluation, i.e., EHM Historical Dataset, Oxford5k and Paris6k [23].

Datasets: The experiments of the proposed scheme and the comparison with baseline methods are based on three visual databases: EHM Historical Dataset, Oxford5k and Paris6k.

Due to the limitations of the CBVIR benchmark dataset, we cooperated with a history professor and scholars from NTU, Singapore, and curated a database of videos and query images. These videos were carefully selected to encompass a diverse range of historical figures and events, and the query images are varied in targets. It includes 51 videos, 397 query images, and 5435 extracted keyframes (for the ablation test). Though our primary focus is to make efficient video image retrieval possible, it is essential to concurrently consider and evaluate the ablation performance of MoReSo framework as an individual module. Therefore, we also use the well-known image databases Oxford5k and Paris6k.

Experimental Setting: Tests were executed using the server equipped with a CPU (64 processors, 32 cores, AMD Ryzen Threadripper PRO 3975WX CPU @ 2.20GHz, 125 GiB RAM) and a GPU (1x NVIDIA RTX A6000 48 GiB RAM).

Evaluation Metrics: For KFE evaluation on EHM dataset, accuracy, redundancy, and efficiency ratio are considered. MoReSo framework is evaluated using mean Average Precision (mAP) and Feature Extraction (FE) efficiency ratio in CBIR tests on Oxford and Paris datasets and overall performance tests on EHM dataset.

A. Evaluation on Historical Dataset

The EHM Historical Dataset is used to evaluate the improved KFE and the proposed CBVIR approach. KFE performance comparison is reported in Table II. Our improved KFE method improves the state-of-the-art statistical feature-based KFE accuracy from 87.9% to 96.9%, and its efficiency ratio ranked the highest with 49.5 among the methods tested. Compared with the deep feature models, though the proposed KFE only ranked third in redundancy with 34.9%, which is less competitive, it is still encouraging to see that our approach significantly outperforms them in efficiency even without using GPU resources.

TABLE II
PERFORMANCE COMPARISON ON DIFFERENT IMAGE REPRESENTATION METHODS IN KFE TASK

Methods	Efficiency ratio	Accuracy	Redundancy
VSUMM	34.22	0.879	0.42
VGG16	1.96/18.06*	0.962	**0.195**
ResNet18	3.45/33.82*	0.949	0.258
Raw color	2.88	0.76	0.35
Proposed KFE	**49.5**	**0.969**	0.349

* Values were obtained using GPU: NVIDIA RTX A6000.

Due to the limitations of open CBVIR systems, the most advanced CBIR system, i.e., SOLAR, is tested, and it is concatenated together with our proposed KFE for fair comparison. Table III shows the performances on accuracy and efficiency of our scheme and the competing state-of-the-art system. Our analysis reveals a substantial increase in efficiency between the proposed method and the existing approach. According

to this table, the proposed CBVIR system with MoReSo module achieved an Efficiency Ratio of 17.68, significantly outstripping the competing method's Efficiency Ratio of 6.62 and being nearly threefold higher.

TABLE III
PERFORMANCE COMPARISON OF THE CBVIR SYSTEMS

Model	mAP(%)	Efficiency ratio
MoReSo	70.38	**17.68**
State-of-the-art	**71.71**	6.62

B. Evaluation on Oxford and Paris Datasets

To gain more insights on the performance of the individual MoReSo framework, we conducted an ablation test by evaluating its efficacy as a standalone CBIR module on the benchmark datasets. We compare it with the state-of-the-art backbones with advanced pooling methods. The efficiency is measured by the inverse of the mean feature extraction time of the test database to execute one query search. In this context, the shortlisting threshold is set as 10% the dataset volume. As indicated in Table IV, the MoReSo framework achieved the highest feature extraction efficiency of 32.74 with only a small compromise in mAP scores, which are 86.92% and 92.58% for Oxford and Paris dataset, respectively.

TABLE IV
PERFORMANCE COMPARISON OF OUR PROPOSED FRAMEWORK WITH STATE-OF-THE-ART CBIR MODELS

Model	Oxford5k mAP(%)	Paris6k mAP(%)	FE efficiency
MoReSo	86.92	92.58	**32.74**
ResNet101 + SOLAR	88.27	**95.04**	4.65
ResNet101 + GeM	**88.74**	94.52	8.77
VGGNet	80.3	83.8	18.87
ResNet101	85.2	88.8	18.52

IV. CONCLUSION

In this study, we focus on the design of a CBVIR model and the improvement of constituent modules. We proposed an effective framework that retrieves visual information from videos efficiently. Significant enhancements were made in KFE module to distill the video content, thereby contributing to the overall efficiency. A novel framework, MoReSo, demonstrates remarkable efficiency and robust accuracy compared with state-of-the-art models across various datasets. Looking forward, our research will revolve around the end-to-end, unified DNN-based framework for CBVIR.

ACKNOWLEDGMENT

We would like to express our sincere appreciation to Dr. Andrea Nanetti for his support and the meticulous construction of the EHM Historical Dataset.

REFERENCES

[1] H. Salmi, What is digital history? John Wiley & Sons, 2020.
[2] L. Zheng, Y. Yang, and Q. Tian, 'SIFT Meets CNN: A Decade Survey of Instance Retrieval', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, pp. 1224–1244, 05 2018.
[3] W. Chen et al., 'Deep Learning for Instance Retrieval: A Survey', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 6, pp. 7270–7292, 2023.
[4] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94.
[5] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, 'ORB: an efficient alternative to SIFT or SURF', 11 2011, pp. 2564–2571.
[6] P. Fernández Alcantarilla, 'Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces', 09 2013.
[7] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', arXiv [cs.CV]. 2015.
[8] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep residual learning for image recognition', in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[9] V. A. Wankhede and P. S. Mohod, 'Content-based image retrieval from videos using CBIR and ABIR algorithm', in 2015 Global Conference on Communication Technologies (GCCT), 2015, pp. 767–771.
[10] P. N. Chatur and R. M. S. Ranjit.M.Shende, 'A Simple Review On Content Based Video Images Retrieval', International journal of engineering research and technology, vol. 2, 03 2013.
[11] P. Papalampidi, F. Keller, and M. Lapata, 'Movie summarization via sparse graph construction', in Proceedings of the AAAI Conference on Artificial Intelligence, 2021, vol. 35, pp. 13631–13639.
[12] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, 'Video summarization using deep neural networks: A survey', Proceedings of the IEEE, vol. 109, no. 11, pp. 1838–1863, 2021.
[13] S. E. F. De Avila, A. P. B. Lopes, A. da Luz Jr, and A. de Albuquerque Araújo, 'VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method', Pattern recognition letters, vol. 32, no. 1, pp. 56–68, 2011.
[14] Z. Zong and Q. Gong, 'Key frame extraction based on dynamic color histogram and fast wavelet histogram', in 2017 IEEE International Conference on Information and Automation (ICIA), 2017, pp. 183–188.
[15] J.-H. Huang and M. Worring, 'Query-Controllable Video Summarization', in Proceedings of the 2020 International Conference on Multimedia Retrieval, Dublin, Ireland, 2020, pp. 242–250.
[16] G. Amato et al., 'VISIONE At Video Browser Showdown 2022', in MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part II, Phu Quoc, Vietnam, 2022, pp. 543–548.
[17] A. Ravi and A. Nandakumar, 'A multimodal deep learning framework for scalable content based visual media retrieval', arXiv preprint arXiv:2105.08665, 2021, unpulished.
[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, 'Mobilenetv2: Inverted residuals and linear bottlenecks', in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.
[19] T. Ng, V. Balntas, Y. Tian, and K. Mikolajczyk, 'SOLAR: second-order loss and attention for image retrieval', in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16, 2020, pp. 253–270.
[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, 'MobileNetV2: Inverted residuals and linear bottlenecks', 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
[21] T. Weyand, A. Araujo, B. Cao, and J. Sim, 'Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval', in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2575–2584.
[22] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas, 'Sosnet: Second order similarity regularization for local descriptor learning', in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 11016–11025.
[23] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, 'Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases', in IEEE Conference on Computer Vision and Pattern Recognition, 2008.