INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A Computational Meta-Learning Inspired Model for Sketch-based Video Retrieval

N Pavithra[1]*, **Y H Sharath Kumar**[2]

1 Research Scholar, Department of Information Science & Engineering , Maharaja Institute of Technology ,Visvesvaraya Technological University, Mysuru, India
2 Professor and Head, Department of Information Science & Engineering , Maharaja Institute of Technology Visvesvaraya Technological University, Mysuru, India

OPEN ACCESS

Received: 02-11-2022

Accepted: 25-01-2023

Published: 20-02-2023

Citation: Pavithra N, Kumar YHS (2023) A Computational Meta-Learning Inspired Model for Sketch-based Video Retrieval. Indian Journal of Science and Technology 16(7): 476-484. https://doi.org/10.17485/IJST/v16i7.2121

* **Corresponding author**.

pavithra.apr02@gmail.com

Funding: None

Competing Interests: None

Copyright: © 2023 Pavithra & Kumar. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To design and develop an efficient computing framework for sketch-based video retrieval using fine-grained intrinsic computational approach. **Methods:** The primary method of sketch-based video retrieval adopts multi-stream multi-modality of joint embedding method for improved P-SBVR from improved fine-grained KTH and TSF related dataset. It considers the potential aspects of the computation of significant visual intrinsic appearance details for sketch objects. The extracted appearance and motion-based features are used to train three different CNN baselines under strong and weak supervision. The system also implements a meta-learning model for different supervised settings to attain better performance of sketch-based video retrieval along with a relational module to overcome the problem of overfitting. **Findings:** The study derives specific sketch sequences from its formulated dataset to compute instance-level query processing for video retrieval. Further, it also addresses the limitations arising in the context of coarse-grained video retrieval models and sketch-based still image retrieval. The aggregated dataset for rich annotation assisted in the experimental simulation. The experimental evaluation with respect to the performance metric evaluates the 3D CNN baselines under strong supervision and weak-supervision where CNN BL-Type-2 attains maximum video retrieval accuracy of 99.96% for triplet grading feature under relational schema. CNN BL-Type-1 attains maximum retrieval accuracy of 97.40% considering the triplet grading features from the improved SBVR. The evaluation metric for the instance level retrieval process also considers true matching of sketches with the videos, it clearly shows that the appropriate appearance and motion based feature selection has enhanced the video retrieval accuracy up to 96.90% with 99.28% accuracy in action identification considering motion stream, 98.17% for appearance module and 98.45% for fusion module. Another important aspect of the proposed research context is that it addresses the problem of cross-modality while executing the simultaneous matching paradigm for visual appearances of the object with its movement appearing on particular video scenes. The experimental outcome shows

its comparable effectiveness relative to the existing system of CNN. **Novelty:** Unlike the conventional system of sketch analysis, which is more focused on static objects or scenes, the presented approach can efficiently compute the important visual intrinsic appearance details of the object of interest from the sketch and then activate the operations for video retrieval. The proposed CNN based learning model with improved P-SBVR dataset attains better computing time for retrieval with are approximately (200, 210 and 214) milliseconds for CNN BL-Type-1, CNN BL-Type-2, CNN BL-Type-3 and comparable with the existing deep learning based SBVR models.

Keywords: Sketch Based Video Retrieval; Intrinsic Appearance Details; Meta Learning; Sketch Dataset; Cross Modality Problem

1 Introduction

The idea behind searching video databases through query formulation has become very popular among researchers. Online video content is rising with the advancement of the Internet, social networking, and other related technological shifts. For example, YouTube uploads 72 hours of videos every minute[(1)]. Also, in most cases, online application requirements specify video repositories searching paradigm for effective content retrieval. The traditional video retrieval method of video repositories involves matching queries to keywords for each target video clip. This standard procedure also involves the assignment of these queries or keywords with the manual human intervention[(2)]. Although it is observed that the formation of keywords could be appropriate for semantic descriptors of content objects (e.g., "horse," "car," etc.), however, when describing the appearance or motion of those objects, the idea of keyword-based semantic descriptors fails. Other aspects that limit the scope of searching for video content include the level of annotation at the clip level, which does not consider the frames or even the objects located within the sequence of frames[(3,4)]. If the spatial and temporal resolutions could be calculated effectively, it might help search for accurate video content. The idea of Querying by Visual Example (QVE) came into practice to offer a potential solution to this problem. However, these QVE systems also require photorealistic queries, which are either images or videos that may not be available to the user when the query got initiated. However, the query processing, in this case, can be time-consuming as the keywords desired to describe the video content may not be efficient for effective retrieval[(5,6)]. Recently, video retrieval with sketch analysis provided convenience for retrieving a specific target video sequence. It must be noted that a sketch depicts thousands of context-oriented interpretations for the viewers. Moreover, sketch analysis (SA) has recently gained significant attention, providing an appropriate abstraction to link concepts with pixels. The SA explores the salient details and topology to provide important information about the subject context needed to be searched. However, the research on sketch-based technologies is currently flourishing across various domains, including image retrieval[(7,8)], sketch generation[(9)], hashing[(10)], abstraction[(11)], and many more. The recent studies emphasizing the aspects of SA majorly focus on illustrating static objects or visual scenes, whereas the analysis of the sketching of objects under motion has yet to be much studied.

Many studies have considered the aspect of human memory where in a normal tendency, human memory can visualize episodic events with selective effects. It can be seen that when it comes to the visual remembrance factors, subjecting the appearance and actions of the key object can be easier, e.g., a moving car, a rising sun, etc.). These remembrance factors corresponding to the specific event or the motion/appearance of an object can be combined with the sketching[(12)]. During the manual process of sketching, free-handmade drawings can easily depict the object in motion or

the appearance of the specific subject through human recollection. These important motivating factors of sketching based on human recollection factors have been adopted in sketch-based video retrieval (SBVR) research.

It is also observed that early publications related to SBVR lacks effectiveness in terms of both computation factor for timely execution and lack of computation of intrinsic appearance details from the sketch. Also, existing system datasets are not much suitable and large enough to train the learning models effectively. On the other hand, most of the objects depicted over the handmade sketches of a particular dataset do not provide significant appearance details, which restrict the sketch's practicability towards identifying the object's particular motion/visual appearance. If this way of sketch analysis/exploitation is performed, it weakens the sketch's full expressiveness and lacks practical motivation for SBVR. The important findings from reviewing the existing related literature are as follows:

1.1 Cross-Modality Problem Exploration

Very less effective studies of SBVR emphasize cross-modal matching problem.

1.2 Annotation for Instance Level Dataset

In existing studies of SBVR, very few approaches incorporate annotation for the ground truth (GT) considering the fine-grained SBVR datasets[13] which in result limits the performance aspects of appropriate video retrieval. Also, the training data scarcity and overfitting problem are also less explored.

1.3. Instance-Level Video Retrieval

Very few studies have been found to incorporate learning approaches for instance-level video retrieval considering the significant appearance attributes or features.

1.4 Motion Depiction in Sketch

No evidence of clear understanding of motion depiction in the illustrated sketch is found in the conventional sketch-analysis for video retrieval.

Taking the limitations in the existing SBVR methods where only static objects or scenes are explored, the study introduces a computationally efficient meta-learning-based SBVR framework that targets the instance-level computation for the given dataset. The framework also targets implementing simplistic execution flow while synchronizing the matching of visual appearance and motion of objects simultaneously. It also considers the appropriate and more precise intrinsic attribute computation modeling. The study's outcome shows that the proposed study understands the motion depiction in sketch and effectively retrieves the video through its formulated large dataset at instance-level timely computation. The simulation outcome on different parameter settings shows that the proposed SBVR approach outperforms the conventional baselines to a greater extent.

2 Methodology

The core idea behind this proposed study is to evolve a novel framework for efficient sketch-based video retrieval, considering the significant appearance details for instance-level computation from the sketch objects. The design of the formulated approach addresses the limitations of the static sketch cross-modal retrieval problem. It introduces instance-level video retrieval with a deeper understanding of the motion depicted within the sketch. The entire operation of execution takes place considering both appearance and motion factors associated with the object of interest.

The design of the step-by-step research method is further illustrated with details of computing. The following Figure 1 shows the preliminary steps of computing to model the framework. Figure 1 clearly shows that the research method in the preliminary design phase considers two major steps: i. Data collection, and ii. Annotation, respectively.

2.1 Data Collection

The data collection process considers skating competition videos, where each video is of approximately 6 to 56 minutes. The study considers 720P and 1080P files stored at 30 FPS. The videos also include audio narratives with channel indicators. The research further extended to extract more than 500 representative HD clips of skaters in motion where the average length is 6.8 seconds with the fine-grained dataset of SBVR. The dataset is quite larger here as it consists of approximately 500 HD figure of skating clips and more that 1400 corresponding sketches. The sketches contains both intrinsic appearance information with

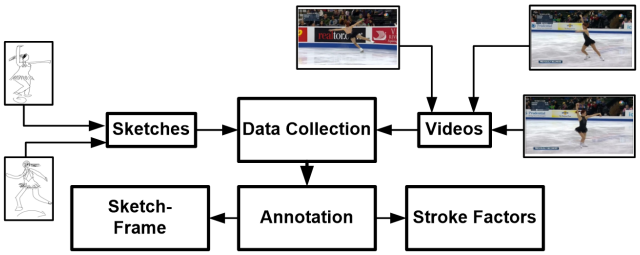

Fig 1. Stage-I Research Method for Data Collection and Annotation

fine-grained details. Here the sequence of sketches also appears to have local dynamic properties along with longer time-frame of dynamicity factors. In the case of a sketch-based dataset, the study considers hand-drawn sketches of prime subjects related to the clips. The fine-grain dataset of SVBR is larger enough as it consists of both videos and sketches. The videos are downloaded from YouTube and related to US National, European and World Championships. The dataset contains of more than 40 female figure skating video samples where each video is of 6 to 56 minutes duration. For each video 720p and 1080p files are stored at 30 FPS. From the original long videos representative clips are also constructed where more than 500 clips are considered of total duration of 3548 seconds. The corresponding sketch samples for the collected video samples are taken from the sketchy repository [14]. The sketch depicts the skater in a certain posture and the motion vector (MV) related to the key moment of the routine for the center point. The sketches consider motion and static sketches without MV to fulfill the requirement of large database systems in SBVR. The prime novelty of the dataset used in this study for sketches clearly shows that it reflects a more detailed appearance factor of the subject with a higher degree of stroke factor, which also reduces the computing effort of models in learning from the scarce data attributes.

2.2 Data preparation with Annotation

During the computation of data annotating factors, the study considers stroke factors and sketch to frame computation. It aims to execute the annotation during the data preparation process and reduce the computing efforts for the meta-learning model used in this study. The study considers stroke factor analysis for annotation, where each sketch skater and their respective motion components are annotated for the stroke factor. In the next phase of analysis, the annotation is performed on the video frames concerning the corresponding sketches.

2.3 Experimental Data Description

The study during the process of experimental validation process considers training, testing and validation dataset. Here the splitting of 520 clips considers 348 video clips for training with 920 different sketches, 36 video clips with 128 sketches for validation and also 136 video clips with 340 sketches for testing of the retrieving accuracy.

2.4 Configuration

The entire numerical implantation of the proposed system has been carried out considering PyTorch with the support of a stonger GPU.

2.5 Experimental Performance Metric

The framework basically implements its idea with the conventional baseline (BL) models of CNN including CNN BL-1 [12], CNN BL-2 [15] and CNN BL-3 [16]. The study considers instance level retrieval execution to measure the retrieval accuracy as a performance metric.

2.6 Model Design: P-SBVR

In this stage of framework modeling the study formulates the model design for proposed SBVR (P-SBVR). The study here considers that the dataset for learning dt consists of n number of sketches (s_i) along with v_i number of video objects. It can be

mathematically represented as:

$$dt = \{(s_i, v_i)\}_{i=1\to n} \tag{1}$$

Here the sketch sequence in the set of s_i consists of appearance factor (a_f) and motion factor (m_f) components for each of the m number of sketch pages which can also be shown as follows:

$$s_i = \{(a_s(j), m_s(j))\}_{j=1\to m} \tag{2}$$

Similarly in case of v_i also appearance and motion factors can be formed for individual frames (iF).

$$v_i = \{(a_v(j), m_v(j))\}_{j=1\to i\,F} \tag{3}$$

The system design basically evaluates the learning for both deep sketch factors along with the multi-modality factors of joint embedding domain. Here the Resemblance Factor (RF) between sketches-based query and the videos can be computed in the form of distance (d) for retrieval of the video object.

The following Figure 2 exhibits the research method for the implementation of the convolutional neural network (CNN) based proposed SBVR system. The CNN model is further trained with respect to the dt as computed from the above equation (1).

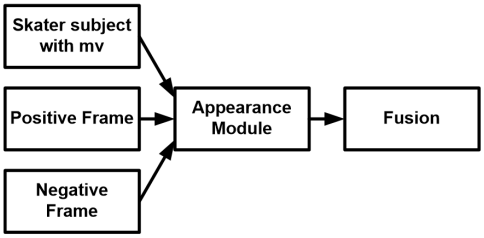

Fig 2. Stage-II Research Method of Appearance Module for the P-SBVR

The system modeling basically formulates three tuple ranking for the appearance module. Here the streaming of objects from the appearance module considers skater subject with MV, positive frame (pf) and negative frame (nf) input for the learning of CNN. The triplet constructed appearance stream is further processed to the fusion block which further also integrates two more modules of the proposed system as shown in the following Figure 3.

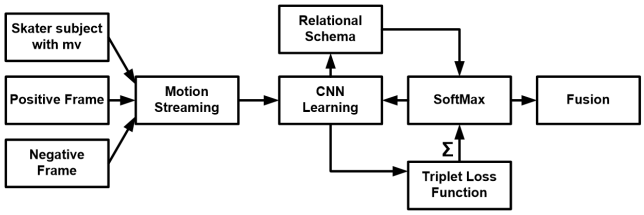

Fig 3. Stage-III Research Method for Motion Streaming and Relational Schema

The Figure 3 shows how the study have implemented motion streaming and relational schema considering the CNN based learning paradigm and further integrates the outcome of the softmax and triplet loss-based learning module with the fusion block. The system design considers optical flow (OP) of both pf and nf of frames and process the learning for OP-CNN. On the other hand, it also considers mv sequences for the sketches to learn the CNN for feature extraction. All the triplet attributes of positive, negative and skater mv further proceed to the relational schema where relation pairs with the combined effect of FC layers generates the Softmax output. The formulated relational module not only deals with the data scarcity problem but also minimizes the probability of data overfitting. The relationship module independently applied to both appearance factor and motion factor to improve the data scarcity problem. The system also processes the mini batch of triplet embedding data of appearance stream for training the CNN which is $dt^a = <a_s, a_v+, a_v->$ and further computes the loss function. The learning

process of CNN here extracts the feature attributes from the appearance stream in the form of Θa_{Fe}. Similar operations take place for the motion stream learning also considering another triplet factors of data dt^m = <m_s, m_v+, m_v->. The loss for triplet construction is denoted with L_t whereas the learning loss is denoted with L_r. However, in the relational schema total loss of each mini-batch is computed with the following mathematical equation (4).

$$T_{lo\ ss} = L_t + \kappa \times L_r \tag{4}$$

In Eq. (4) the total loss factor of the model is calculated for CNN batch processing. Here **K** represents the weight factor. After executing the relational schema for triplet grading for the two streams, the system further concatenates and perform fusion of the streams to the FC layers. The fusion approach here handles the data overfitting with grading-based fusion and feature concatenation based fusion approach

Grading Fusion Modeling**:** In grading based fusion modeling the system basically generates a series of matching videos $\{v_i\}$ for the specified sketch querys. Here it executes the Euclidian distance measuring. Here the grading g for appearance stream and motion stream of video i is computed with $g_a(j)$, $g_m(j)$. The final ranking is computed with the equation (5)

$$G_j = \kappa \times g_a(j) + (1 - \kappa) \times g_m(j) \tag{5}$$

The Euclidian distance-based grading mechanism here executed for each database video clips considers weighted averaging for appearance and motion vector streams (mv). Finally, the fusion of the feature attributes from two different streams take place.

2.7 Supervised Learning Modeling

The study further evaluates a strong process of supervised learning algorithm with multi-instance learning. Here the model is learned with more detailed annotated correspondence features of cross modal. During the learning of this model has a mapping procedure which maps a particular sketch subject in motion with different corresponding scenes of the videos. And these way the detailed correspondence factors are annotated. Here the model is trained with strong supervision considering the relationship factors between the pages of sketch query subject with the corresponding different frame instances of related video clips which is considered to be positive video candidates within the frames otherwise frames which are outside the video clips can be stated as negative.

2.8 Model Deployment

The study further performs the trained model deployment and the testing aims to match the specific videos for the sketch query generated. In order to match the sketch query with the respective video scenes, the network of the model is already trained with the grading of sketch features and the video frames. The CNN output modeling is further subject to perform the retrieval of the video sequence. The system also performs deep feature extraction procedure for sketches and videos considering both appearance and motions features. The procedure of the video retrieval includes computing of the lowest sum corresponding to the nearest neighbor correlation cost factors for each sketch. It is represented as follows:

$$\mathrm{Dist}(s, v) = \frac{1}{m} \times \sum_{j=1, k=l \to h}^{m} \min\left(d\left(s_j, v_k\right)\right) \tag{5}$$

The algorithm of the proposed P-SBVR is shown as follows:

Algorithm-1: P-SBVR Learning Algorithm

Input:$dt = \{(s_i, v_i)\}_{i=1 \to n}$

Data collection/ Preparation

Compute Appearance and Motion Factors using (2 and (3)

Formulate Appearance Module

Execute Motion Streaming and Relational Schema

dt^a = <a_s, a_v+, a_v->

dt^m = <m_s, m_v+, m_v->

Compute ΘaFe , ΘmFe

T_{loss} using (4)

Execute Grading Mechanism on Feature Ranking

$G_j = \kappa \times g_a(j) + (1 - \kappa) \times g_m(j)$

Supervised Learning with intrinsic more detailed appearance and motion features
Model Deployment
$\text{Dist}(s,v) = \frac{1}{m} \times \sum_{j=1,k=l\to h}^{m} \min\left(d\left(s_j, v_k\right)\right)$
Output: video retrieval
Performance Metric: Retrieval Accuracy (R_a), Computing Time (t).

The equation (6) shows how the model perform matching operation between the sketch and the video attributes and compute the minimum sum of matching cost to retrieve the video. The next section further shows the numerical outcome while executing this algorithm on a computing environment.

3 Results and Discussion

The experimental analysis outcome from the Figure 4 shows that the incorporation of relational schema (RS) has significantly improved the learning performance of the CNN models and which eventually the outcome of the video retrieval accuracy clearly depicts. It clearly shows the CNN BL Type-1 when trained with the prepared dataset of the proposed model achieves 98.40% retrieval accuracy under strong supervision whereas CNN-BL type-2 and type-3 achieves 99.96% and 95% accuracy respectively.

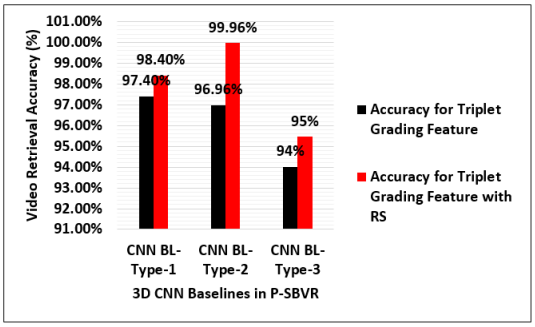

Fig 4. Video Retrieval Accuracy under Strong Supervised Learning

The study also shows that under weak-supervision learning when applied with the RS and triplet grading aspects the CNN-BL models accomplishes considerable outcome for video retrieval accuracy. The prime reason behind high accuracy of the video retrieval is that the matching schema considering the intrinsic appearance and motion analysis for both sketches and the videos. The study also further validates the performance of the models for three different types of core modules associated with the P-SBVR. It also shows how the proposed feature extraction through intrinsic appearance and motion details has improved the learning accuracy for the action identification.

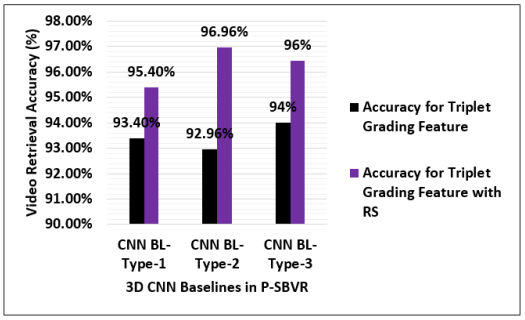

Fig 5. Video Retrieval Accuracy under Weak-Supervised Learning

The Following Figure 6 clearly shows that the motion stream analysis modules helps in reaching 99.28% accuracy for CNN BL-Type-3 whereas in the case of CNN BL-Type-1 and CNN BL-Type-2 the model accomplishes 97.5% and 98.58% accuracy of action identification from the videos.

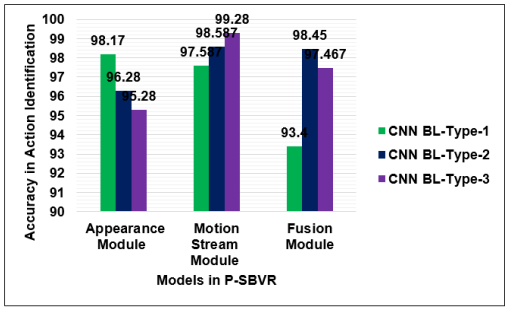

Fig 6. Accuracy (%) in Action Identification

The study further also incorporates analyzing the computing time of retrieval for different CNN baselines models when trained with the prepared dataset. It clearly shows that all the module approximately achieves 200 milliseconds of retrieval time.

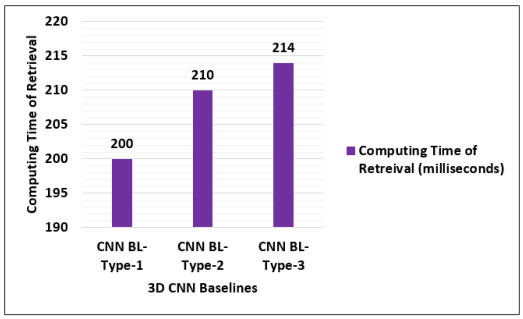

Fig 7. Computing Time for 3D CNN Baselines Models

The closer interpretation of the Figure 7 shows that CNN BL-Type-1 achieves 200 milliseconds of computing time associated with retrieval whereas CNN BL-Type-2 and type-3 achieves 210 milliseconds and 214 milliseconds respectively. The study has considered the related models of CNN, which had been highly referred for the SBVR in the recent time for comparison. It has been highlighted in the literature that CNN-based SBVR has gained much more popularity owing to its potential to learn from the features under strong supervision. As the problem of fine-grained intrinsic feature computation in SBVR is a new problem, existing methods cannot be compared directly. To maintain a fair comparison, the proposed system considers related popular video analysis models and applies the proposed approach of fine-grained intrinsic computation-based SBVR on 3D convolutional networks. The study claims that the proposed model has significantly improved the performance of 3D ResNet18 [15] compared to CNN baseline-1 [12] and CNN baseline-2 [16]. It is observed that the proposed approach of the SBVR model works efficiently with CNN baseline-2 regarding video retrieval accuracy. The prime reason is that it considers strong supervision with ranking or feature computation which has also considered complementary appearance and motion patterns of information for SBVR. Overall, it accomplishes performance of 97.6% accuracy in action identification. However, when it comes to computing time of retrieval there, the CNN BL-Type-1 has attained better performance.

4 Conclusion

This study introduces a novel SBVR system considering meta-learning inspired algorithm designs. The proposed framework when applied with the CNN different types of baselines improves the accuracy of retrieval considering a combination of both triplet formation of ranking and relational schema. The experimental validation has been conducted comparing the performance of three CNN baseline approaches towards finding the accuracy of SBVR. In the case of strong supervised learning both CNN BL-Type-1 and CNN BL-Type-2 yields considerable accuracy as compared to CNN BL-Type-3 which are approximately 98.40%, 99.96% but for very limited sketch sequence queries. On the other hand, for weak supervised learning CNN BL-Type-2 and CNN BL-Type-3 attains accuracy of 96.96% which are comparable to the existing system of SBVR. The CNN BL-Type-2 99.28% accuracy for motion stream module which is superior to the other baseline methods. The experimental

outcome clearly shows that although the proposed approach of SBVR with CNN BL-type-2 attains 99.96% of retrieval accuracy but the overall performance of retrieval accuracy for CNN BL-type-2 is found to be 97.6% on an average. The outcome also shows that the system achieves significant performance improvement in the computing time of video retrieval which is approximately 200 milliseconds and comparable with the baseline approaches. The future research work aims to evaluate the model with different parameter settings and observe the performance.

References

1) Araujo A, Girod B. Large-Scale Video Retrieval Using Image Queries. *IEEE Transactions on Circuits and Systems for Video Technology*. 2018;28(6):1406–1420. Available from: https://doi.org/10.1109/TCSVT.2017.2667710.
2) Sheng B, Li P, Gao C, Ma KL. Deep Neural Representation Guided Face Sketch Synthesis. *IEEE Transactions on Visualization and Computer Graphics*. 2019;25(12):3216–3230. Available from: https://doi.org/10.1109/TVCG.2018.2866090.
3) Xu P, Huang Y, Yuan T, Pang K, Song YZ, Xiang T, et al. SketchMate: Deep Hashing for Million-Scale Human Sketch Retrieval. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018;p. 8090–8098. Available from: https://doi.org/10.1109/CVPR.2018.00844.
4) Xu P, Yin Q, Huang Y, Song YZ, Ma Z, Wang L, et al. Cross-modal subspace learning for fine-grained sketch-based image retrieval. *Neurocomputing*. 2018. Available from: https://doi.org/10.48550/arXiv.1705.09888.
5) Muhammad UR, Yang Y, Song YZ, Xiang T, Hospedales TM. Learning Deep Sketch Abstraction. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018;p. 8014–8023. Available from: https://doi.org/10.1109/CVPR.2018.00836.
6) Madore KP, Jing HG, Schacter DL. Selective effects of specificity inductions on episodic details: evidence for an event construction account. *Memory*. 2019;27(2):250–260. Available from: https://doi.org/10.1080/09658211.2018.1502322.
7) David H, Eck D. A neural representation of sketch drawings. 2017. Available from: https://doi.org/10.48550/arXiv.1704.03477.
8) Jing T, Xia H, Hamm J, Ding Z. Augmented Multimodality Fusion for Generalized Zero-Shot Sketch-Based Visual Retrieval. *IEEE Transactions on Image Processing*. 2022;31:3657–3668. Available from: https://doi.org/10.1109/TIP.2022.3173815.
9) Sun H, Xu J, Wang J, Qi Q, Ge C, Liao J. DLI-Net: Dual Local Interaction Network for Fine-Grained Sketch-Based Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*. 2022;32(10):7177–7189. Available from: https://doi.org/10.1109/TCSVT.2022.3171972.
10) Wang L, Qian X, Zhang X, Hou X. Sketch-Based Image Retrieval With Multi-Clustering Re-Ranking. *IEEE Transactions on Circuits and Systems for Video Technology*. 2020;30(12):4929–4943. Available from: https://doi.org/10.1109/TCSVT.2019.2959875.
11) Liang S, Dai W, Wei Y. Uncertainty Learning for Noise Resistant Sketch-Based 3D Shape Retrieval. *IEEE Transactions on Image Processing*. 2021;30:8632–8643. Available from: https://doi.org/10.1109/TIP.2021.3118979.
12) Wang X, Girshick R, Gupta A, He K. Non-local Neural Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018;p. 7794–7803. Available from: https://doi.org/10.1109/CVPR.2018.00813.
13) Collomosse JP, Mcneill G, Qian Y. Storyboard sketches for Content Based Video Retrieval. *2009 IEEE 12th International Conference on Computer Vision*. 2009;p. 245–252. Available from: https://doi.org/10.1109/ICCV.2009.5459258.
14) Sangkloy P, Burnell N, Ham C, Hays J. The sketchy database: learning to retrieve badly drawn bunnies. ACM TOG. 2016. Available from: https://doi.org/10.1145/2897824.2925954.
15) Hara K, Kataoka H, Satoh Y. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018;p. 6546–6555. Available from: https://doi.org/10.1109/CVPR.2018.00685.
16) Liu K, Liu W, Ma H, Tan M, Gan C. A Real-Time Action Representation With Temporal Encoding and Deep Compression. *IEEE Transactions on Circuits and Systems for Video Technology*. 2021;31(2):647–660. Available from: https://doi.org/10.1109/TCSVT.2020.2984569.