Journal of Computer and Communications, 2024, 12, 26-36
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

CLIP4Video-Sampling: Global Semantics-Guided Multi-Granularity Frame Sampling for Video-Text Retrieval

Tao Zhang, Yu Zhang*

School of Computer Science and Engineering, The Key Lab of Computer Network and Information Integration (Ministry of Education), Southeast University, Nanjing, China
Email: *zhang_yu@seu.edu.cn

How to cite this paper: Zhang, T. and Zhang, Y. (2024) CLIP4Video-Sampling: Global Semantics-Guided Multi-Granularity Frame Sampling for Video-Text Retrieval. *Journal of Computer and Communications*, **12**, 26-36.
https://doi.org/10.4236/jcc.2024.1211002

Received: October 17, 2024
Accepted: November 3, 2024
Published: November 6, 2024

Abstract

Video-text retrieval (VTR) is an essential task in multimodal learning, aiming to bridge the semantic gap between visual and textual data. Effective video frame sampling plays a crucial role in improving retrieval performance, as it determines the quality of the visual content representation. Traditional sampling methods, such as uniform sampling and optical flow-based techniques, often fail to capture the full semantic range of videos, leading to redundancy and inefficiencies. In this work, we propose CLIP4Video-Sampling: Global Semantics-Guided Multi-Granularity Frame Sampling for Video-Text Retrieval, a global semantics-guided multi-granularity frame sampling strategy designed to optimize both computational efficiency and retrieval accuracy. By integrating multi-scale global and local temporal sampling and leveraging the CLIP (Contrastive Language-Image Pre-training) model's powerful feature extraction capabilities, our method significantly outperforms existing approaches in both zero-shot and fine-tuned video-text retrieval tasks on popular datasets. CLIP4Video-Sampling reduces redundancy, ensures keyframe coverage, and serves as an adaptable pre-processing module for multimodal models.

Keywords

Video Sampling, Multimodal Large Language Model, Text-Video Retrieval, CLIP Model

1. Introduction

Video-text retrieval (VTR) has emerged as a critical research area in multimodal learning, aiming to bridge the semantic gap between text descriptions and video

data. It involves retrieving the most relevant video for a given text query, or vice versa, using deep learning models to understand both visual and textual semantics.

Traditional methods for video sampling often include naive strategies like uniform sampling (selecting frames at regular intervals) or optical flow-based sampling to focus on dynamic changes [1] [2]. While effective to some extent, these approaches fail to fully utilize semantic content, often resulting in redundancy or missing critical information. The recent rise of transformer-based models like Vision Transformer (ViT) [3] has improved feature extraction capabilities by better capturing global semantics across frame images, compared to traditional CNN-based methods, which are more suited to local spatial feature extraction [4], which limits their ability to capture high-level global semantics. ViT-based models have proven effective in representing both high-level global features and specific details across frames, which is crucial for video-text retrieval tasks [5] [6].

Pre-trained models such as CLIP (Contrastive Language-Image Pre-training) [7] have shown remarkable potential in bridging the gap between visual and textual modalities through shared feature spaces. CLIP uses contrastive learning to embed text and images into a unified high-dimensional space, making it a suitable backbone for VTR task. However, directly applying CLIP to videos introduces challenges, such as how to effectively capture both temporal and spatial aspects of videos and avoid overwhelming computational demands. To address these issues, we propose CLIP4Video-Sampling: a global semantics-guided multi-granularity frame sampling strategy for video-text retrieval. This novel method combines global semantic understanding with a multi-granularity sampling approach to effectively capture key frames, enhancing the overall efficiency and performance of video-text retrieval. Our method integrates coarse-grained and fine-grained frame sampling to capture essential video content, leveraging the CLIP model for semantic feature extraction. This approach aims to reduce redundancy and computational costs while maintaining high retrieval accuracy. Our contributions include a new keyframe selection method that is based on the similarity of global semantics between frames, providing a more flexible and resource-efficient preprocessing strategy for multimodal models.

The task of video-text retrieval has gained significant attention due to the rapid growth of online video content and the need for efficient search mechanisms. Traditional approaches relied on simple frame sampling techniques, such as selecting frames at fixed intervals or using optical flow to identify moments of significant motion [8]. These methods were computationally simple but struggled with semantic nuance and redundancy, especially in scenes with static backgrounds or repetitive actions.

Recent advances aimed to address these limitations by integrating more sophisticated techniques. For example, Streamflow proposed a multi-frame optical flow estimation using a streamlined batch-in multi-frame (SIM) pipeline to improve efficiency and reduce redundant computations [9]. Other improvements in optical flow included the use of segmented smoothing and anisotropic regularization

for edge preservation and noise resistance [10], localized dense optical flow combined with YOLOv3 for behavior recognition [11], and integration with terrain data to enhance motion parameter estimation [12]. Multi-stream networks applied optical flow to extract key frames from RGB and flow streams, enhancing action recognition [13], while Lagrange interpolation-based key frame selection further improved sampling efficiency [14].

Deep learning has introduced sophisticated CNN-based approaches to better capture spatial and temporal features in video content. Spatiotemporal attention mechanisms, such as those introduced by CSTA [15] and TAFCN-SUM [16], have been effective in modeling inter-frame and intra-frame relationships, thereby enhancing key frame selection. Adaptive sampling approaches, like mmSampler, use lightweight policy networks to dynamically select salient frames for video retrieval [4].

CLIP-based methods have also seen substantial adoption in video-text retrieval. Approaches like Dixit Player with Open CLIP [17], X-CLIP [18], CLIP4Clip [6], and InternVideo [19] have extended CLIP's capabilities by introducing mechanisms such as Cross-Frame Attention, multiple similarity metrics, and multi-level feature fusion to enhance global and fine-grained semantic understanding for diverse retrieval tasks.

Despite these advancements, earlier research has limitations in video frame sampling methods. Traditional optical flow techniques are limited to local features and struggle to capture complex temporal relationships, while average sampling designs are overly simplistic. These approaches fundamentally lack compatibility with modern Transformer-based multimodal models, which require richer temporal context and more sophisticated feature extraction to fully leverage their capabilities.

Our method, CLIP4Video-Sampling, builds on previous advancements by integrating multi-scale global and local temporal features, using parameter-free plug-and-play modules, and leveraging ViT-based CLIP models for feature extraction. We optimize video-text retrieval performance and efficiency by combining global semantics-guided sampling with flexible keyframe selection.

2. Method

In this section, we will describe in detail CLIP4Video-Sampling: global semantics-guided multi-granularity combined frame sampling strategy. First, we provide an overview of our global semantics-guided multi-granularity sampling method. Then, we detail the extraction method of global semantics for each frame image. Finally, we elaborate on the method (Sampler) that uses the similarity between the global semantics of frames to guide multi-granularity combined sampling.

2.1. Overview

Multimodal video-text models typically require pre-processing the input text and video to better extract information for inference. Among them, video as a visual

modality input is usually composed of a series of densely captured frames. Therefore, the vast majority of multimodal models adopt a frame sampling process when processing video data, converting videos into frame image sets. Considering temporal redundancy and limited computational resources, the pre-processing step does not usually sample all frames of a video but selects keyframes through a specific strategy to improve computational efficiency. Our proposed global semantics-guided multi-granularity combined sampling is an efficient sampling strategy. We conducted video-text retrieval tasks on multiple public datasets. Compared to other sampling methods, our approach achieved results superior to the SOTA (Table 2, Table 3) while sampling an equal or fewer number of frames (Table 1). Compared to other sampling methods, our method is more versatile and can serve as a pre-processing module suitable for various multimodal models (e.g., CLIP4Clip, InternVideo1 [19], etc.). It is more flexible and can sample different numbers of video frames based on computational resource constraints and precision requirements.

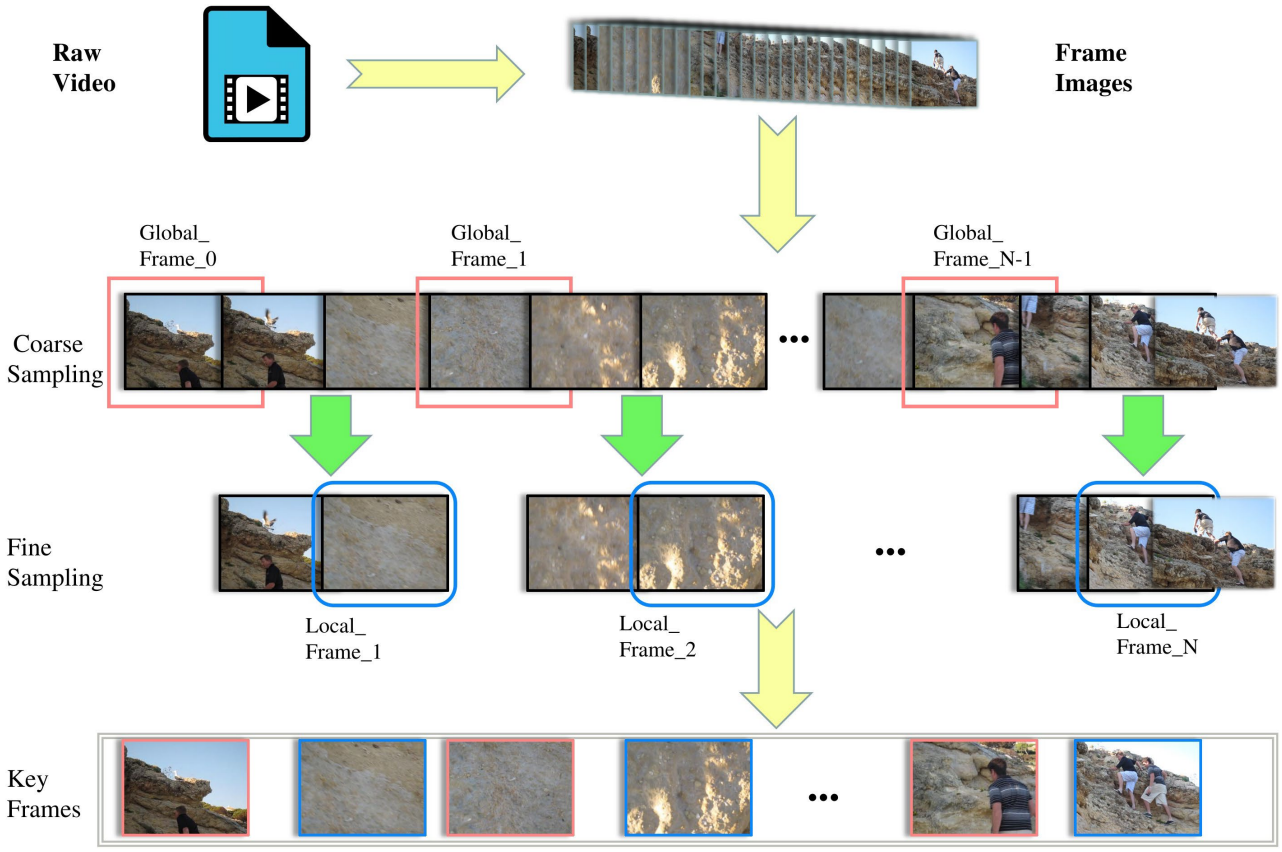

Figure 1. The pipeline of CLIP4Video Sampling.

The core of our sampling method is global semantics understanding and similarity calculation between adjacent frames, as well as a multi-granularity keyframe selection strategy, as shown in Figure 1. The keyframe selection strategy involves calculating the similarity between all adjacent frames in the video. A lower similarity

indicates greater differences. We combined multiple temporal segmentation scales and selected the second frame from pairs with low similarity as the sampled keyframe. For global semantics understanding, we use the ViT (Vision Transformer)-based vision model pre-trained with large-scale image-text contrastive loss to extract global semantics from video frames. For similarity calculation, we adopt the cosine similarity method used during the model's contrastive loss training. Based on these foundations, we design a keyframe selection strategy, including coarse-grained global timespan frame extraction and fine-grained local timespan frame extraction. By combining the results from both coarse-grained and fine-grained frame sampling, we obtain a set of keyframes that represent the entire video. In the following sections, we will define and describe in detail the semantic understanding of frames, the similarity calculation between frames, and the multi-granularity sampling process.

2.2. Global Semantics Extraction and Similarity Calculation

To reduce computational complexity while retaining sufficient information, we first adjust the frame rate of the video to a lower value (e.g., 5 to 10 frames per second). Then, we extract each frame sequentially, resulting in the entire frame set F of the video:

$$F = \{f_1, f_2, \cdots, f_n\}, \tag{1}$$

where f_i represents the i-th frame. For each frame f_i, we use the CLIP model with a ViT-based Vision-Tower to extract the high-dimensional feature vector v_i, capturing the global visual semantics v_i of each frame:

$$v_i = \text{CLIP\_Extract}(f_i), \tag{2}$$

We chose CLIP because the model learns image-text embeddings in a shared high-dimensional feature space through large-scale contrastive training, making it particularly suitable for video content analysis tasks. CLIP captures not only visual information but also embeds both visual and textual semantics in close proximity using cosine similarity, enabling it to effectively understand the relationship between visual content and specific language descriptions. This capability is challenging for traditional methods, such as optical flow, which can handle dynamic information but often overlook the global semantic content of images.

Next, we need to compare whether each pair of consecutive frames undergoes significant global semantic change. Here, we use cosine similarity for the calculation, as it aligns with the contrastive loss used in training the CLIP model and effectively reflects global information differences between images. For the feature vectors (v_i, v_{i+1}) of adjacent frames (f_i, f_{i+1}), the similarity $s_{i,i+1}$ is calculated as:

$$s_{i,i+1} = \frac{v_i \cdot v_{i+1}}{\|v_i\| \|v_{i+1}\|}, \tag{3}$$

where $s_{i,i+1}$ is the similarity result. A higher $s_{i,i+1}$ indicates that frames (f_i, f_{i+1}) are similar in content and semantics, whereas a lower $s_{i,i+1}$ indicates a greater semantic difference between frames (f_i, f_{i+1}), implying significant new semantic content, which serves as the basis for selecting keyframes.

2.3. Keyframe Selection with Multi-Granularity Combined Sampling Strategy

For convenience, we assume that the frame before the first frame of the video is a matrix filled with zeros. Thus, the first frame f_0 of the video is always sampled as a keyframe, initializing the start of the video. Subsequently, we calculate the similarity between each frame and its predecessor, starting from the second frame of the video, to obtain the similarity score set S for all frame pairs, containing $n-1$ scores (where n is the total number of frames):

$$S = \{s_{1,2}, s_{2,3}, \cdots, s_{n-1,n}\}, \tag{4}$$

In a frame pair (f_i, f_{i+1}), the lower the similarity $s_{i,i+1}$, the greater the change in global semantics from f_i to f_{i+1}. Therefore, f_{i+1} represents the beginning of new visual information, marking the transition from a previous state to a new one. By applying sampling operations of different temporal granularities to the similarity score set S, we can select the corresponding frames to obtain the desired keyframes. To ensure our sampling strategy captures both global content and local details of the video, we combine two sampling strategies, as follows.

1) Coarse Sampling: To ensure that our sampling strategy first focuses on the global content of the video, we conduct coarse-grained sampling on the entire frame set. We sort the similarity scores S in ascending order:

$$S' = \text{sort}(S), \tag{5}$$

$$S' = \{s'_1, s'_2, \cdots, s'_{n-1}\},\ s'_1 \le s'_2 \le \cdots \le s'_{n-1}, \tag{6}$$

For the sorted scores S', we select each last frame of the top $N-1$ pairs, along with the first frame f_0 of the video, to form the coarse sampling result set G, consisting of N frames. Arranging these frames in temporal order yields the coarse sampling result set G:

$$G = \{g_1, g_2, \cdots, g_N\}, \tag{7}$$

The coarse sampling results quickly capture the moments of greatest content change, allowing the keyframes to cover the entire video while maintaining temporal continuity.

2) Fine Sampling: However, coarse sampling alone may not cover all video content and might introduce redundancy. To further capture the details, we design a fine-grained temporal sampling strategy. For each keyframe obtained from coarse sampling, we treat it as a temporal node, and, together with the end frame of the video, partition the original video into N segments. In each segment, we identify the frame pair with the lowest similarity score and select the second frame of this pair as the keyframe for fine sampling, yielding the fine-grained local sampling

set L with N sampled frames:

$$L = \{l_1, l_2, \cdots, l_N\}, \tag{8}$$

2.4. Frame Sampling Results

By combining the coarse and fine sampling results, we obtain the final frame set:

$$\text{Final\_Samples} = \{f_0\} \cup G \cup L, \tag{9}$$

which contains the initial frame of the video, $N-1$ coarse-grained sampled frames, and N fine-grained sampled frames, with a total of $2N$ frames. This sampling strategy effectively captures the main variations of the video with a small number of frames, adapting to different video content analysis requirements. Both the number of coarse-grained frames $N-1$ and the number of iterations in fine-grained sampling can be flexibly adjusted based on precision requirements and available computational resources.

3. Experiment

To verify the advanced performance and practical significance of our sampling method, we conducted experiments to compare it with state-of-the-art (SOTA) models and traditional sampling methods on multiple datasets.

Dataset We selected commonly used datasets for Video-Text Retrieval tasks, including MSR-VTT [20], DiDeMo [21], and MSVD [22]. MSR-VTT is a large dataset for open-domain video captioning consisting of 10,000 video clips from 20 categories, with 20 English sentences annotated per clip, and about 29,000 unique words in total. The standard split uses 7k clips for trainval and 3k for testing [20]. DiDeMo is one of the largest and most diverse datasets for temporally localizing events in videos using natural language descriptions. These videos were collected from Flickr, with a maximum duration of 30 seconds. It uses 9k clips for trainval and 1k for testing [21]. MSVD consists of 2k video clips and 12k sentence annotations containing bilingual descriptions. It uses 1300 clips for trainval and 640 clips for testing [22].

Set We followed InternVideo1's [19] pre-processing and center-cropped all videos in the datasets to 224 × 224, with a frame rate of 30 fps. It is worth mentioning that although InternVideo1 achieved SOTA performance on the Video-Text Retrieval tasks for the three datasets, it used different maximum sampled frame numbers for each dataset: 8 for MSRVTT, 12 for MSVD, and 32 for DiDeMo. For comparison purposes, we used the same or fewer sampled frames than InternVideo1 across various datasets. This demonstrates that the performance improvement of our method is not due to an increased number of sampled frames, as reported in **Table 1**. We selected the EVA02-CLIP [23] vision-tower as the global feature extractor for frames, replacing the random keyframe sampling strategy in InternVideo1 to perform the Video-Text Retrieval task.

Result After using our sampling method, we tested the performance of InternVideo1 in both finetuned and zero-shot scenarios. The results show that the model

achieved higher accuracy compared to SOTA in both finetuned and zero-shot Video-Text Retrieval tasks, as reported in Table 2 and Table 3. From these tables, we derive the following conclusions: First, comparing finetuned and zero-shot results, we find that our method provides significant improvements for zero-shot performance across multiple datasets, with improvements greater than those for the finetuned model (Table 2, Table 3). We believe the reason is that multimodal models might develop semantic biases after fine-tuning datasets using their respective sampling strategies (e.g., random or average temporal sampling). In contrast, the zero-shot model, which was not finetuned or post-trained, exhibited significant improvements using our superior sampling strategy, demonstrating the efficacy of our method and its potential for further training optimization. Second, we observed that performance improvements varied across datasets, with the most significant improvement seen in the MSRV-TT dataset: 3.5% and 6.4% for T2V and V2T, respectively, in zero-shot experiments, as reported in Table 3. Upon analyzing the dataset structure, we found that MSRV-TT's video clips are more content-rich, with each clip potentially containing multiple semantic events. In comparison, MSVD and DiDeMo consist of shorter clips with single semantic events and minimal scene changes. Therefore, we conclude that our method is advantageous in handling longer videos with multiple scenes and semantic events.

Table 1. Number of frame sampling operations for MSRV-TT, MSVD, and DiDeMo.

Dataset	Coarse-grained sampling	Fine-grained sampling	Sampling result
MSRVTT	3	4	8
MSVD	5	6	12
DiDeMo	10	11	22

Table 2. Fine-tuned model performance on Video-Text retrieval tasks for MSRVTT, MSVD, and DiDeMo datasets (T2V: Text to Video, V2T: Video to Text, reporting R@1: the percentage of correct matches in the top 1 retrieval result).

Method	MSRV-TT		MSVD		DiDeMo	
	T2V	V2T	T2V	V2T	T2V	V2T
CLIP4Clip [6]	45.6	45.9	45.2	48.4	43.0	43.6
TS2Net [24]	49.4	46.6	-		41.8	-
X-CLIP [18]	49.3	48.9	50.4	66.8	47.8	47.8
InternVideo1 [19]	55.2	57.9	58.4	76.3	57.9	59.1
Ours	**55.4**	**58.1**	**58.7**	**76.4**	**58.9**	**57.4**

Table 3. Zero-shot model performance on Video-Text retrieval tasks for MSRVTT, MSVD, and DiDeMo datasets (T2V: Text to Video, V2T: Video to Text, reporting R@1: the percentage of correct matches in the top 1 retrieval result).

Method	MSRV-TT		MSVD		DiDeMo	
	T2V	V2T	T2V	V2T	T2V	V2T
CLIP [7]	35.0	32.3	39.2	63.3	29.8	24.3
CLIP4Clip [6]	30.6	-	36.2	-	-	-
InternVideo1 [19]	40.7	39.6	43.4	**67.6**	31.5	33.5
Ours	**44.3**	**46.0**	**49.6**	62.2	**33.3**	**33.6**

4. Conclusions

Our proposed method, CLIP4Video-Sampling, introduces a novel, multi-granularity frame sampling strategy for video-text retrieval. By integrating multi-scale global and local temporal features, our approach provides a flexible keyframe selection mechanism that captures the semantic nuances of video content. The key advantages of CLIP4Video-Sampling are its parameter-free plug-and-play design, the use of global semantics-guided sampling, and the capability to improve retrieval performance while maintaining efficiency. Compared to traditional sampling techniques and existing multimodal methods, our approach effectively balances computational cost with retrieval accuracy, achieving state-of-the-art results in both zero-shot and fine-tuned scenarios.

Despite its strengths, CLIP4Video-Sampling has limitations. The method has not yet been trained end-to-end within a complete multimodal learning framework, which may leave untapped performance potential. Future work could focus on integrating this sampling strategy into trainable modules, thereby improving the learning process for multimodal video understanding.

The contributions of this work are expected to positively impact dataset optimization for multimodal tasks and facilitate the use of efficient data pre-processing in training other multimodal models. In the future, integrating CLIP4Video-Sampling into trainable models can further enhance the understanding and retrieval of complex video content.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Bain, M., Nagrani, A., Varol, G. and Zisserman, A. (2021) Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 1728-1738. https://doi.org/10.1109/iccv48922.2021.00175

[2] Chen, Y., Wang, S., Zhang, W. and Huang, Q. (2018) Less Is More: Picking Informative

Frames for Video Captioning. *Computer Vision—ECCV* 2018 15*th European Conference*, Munich, 8-14 September 2018, 367-384. https://doi.org/10.1007/978-3-030-01261-8_22

[3] Dosovitskiy, A. (2020) An Image Is Worth 16 x 16 Words: Transformers for Image Recognition at Scale.

[4] Hu, Z.M., Ye, N. and Mohomed, I. (2022) mmSampler: Efficient Frame Sampler for Multimodal Video Retrieval. Proceedings of Machine Learning and Systems, 4, 153-171.

[5] Fang, H., Xiong, P.F., Xu, L.H. and Chen, Y. (2021) Clip2video: Mastering Video-Text Retrieval via Image Clip.

[6] Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., et al. (2022) Clip4clip: An Empirical Study of CLIP for End to End Video Clip Retrieval and Captioning. *Neurocomputing*, **508**, 293-304. https://doi.org/10.1016/j.neucom.2022.07.028

[7] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., *et al.* (2021) Learning Transferable Visual Models from Natural Language Supervision. *International Conference on Machine Learning*, 18-24 July 2021, 8748-8763.

[8] Tamgade, S.N. and Bora, V.R. (2009) Notice of Violation of IEEE Publication Principles: Motion Vector Estimation of Video Image by Pyramidal Implementation of Lucas Kanade Optical Flow. 2009 2*nd International Conference on Emerging Trends in Engineering & Technology*, Nagpur, 16-18 December 2009, 914-917. https://doi.org/10.1109/icetet.2009.154

[9] Sun, S., Liu, J.M., Li, T.H., Li, H.X., Liu, G.Q. and Gao, W. (2023) Streamflow: Streamlined Multi-Frame Optical Flow Estimation for Video Sequences.

[10] Wei, H., Qian, Z., Bo, Q. and Yang, Y. (2018) Research on HS Optical Flow Algorithm Based on Motion Estimation Optimization. *Journal of Computer and Communications*, **6**, 171-184. https://doi.org/10.4236/jcc.2018.611017

[11] Zheng, H., Liu, J. and Liao, M. (2021) Study on Local Optical Flow Method Based on Yolov3 in Human Behavior Recognition. *Journal of Computer and Communications*, **9**, 10-18. https://doi.org/10.4236/jcc.2021.91002

[12] Kupervasser, O.Y. and Rubinstein, A.A. (2013) Correction of Inertial Navigation System's Errors by the Help of Video-Based Navigator Based on Digital Terrarium Map. *Positioning*, **4**, 89-108. https://doi.org/10.4236/pos.2013.41010

[13] Xia, L.M. and Wen, X. (2024) Multi-Stream Network with Key Frame Sampling for Human Action Recognition. *The Journal of Supercomputing*, **80**, 11958-11988.

[14] Yuan, M., Long, Y. and Li, X. (2024) Real-Time Mosaic Method of Aerial Video Based on Two-Stage Key Frame Selection Method. *Open Journal of Applied Sciences*, **14**, 1008-1021. https://doi.org/10.4236/ojapps.2024.144067

[15] Son, J., Park, J. and Kim, K. (2024) CSTA: CNN-Based Spatiotemporal Attention for Video Summarization. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 18847-18856. https://doi.org/10.1109/cvpr52733.2024.01783

[16] Gupta, D. and Sharma, A. (2023) A Two-Stage Attention Augmented Fully Convolutional Network-Based Dynamic Video Summarization. *Multimedia Systems*, **29**, 3685-3701. https://doi.org/10.1007/s00530-023-01154-2

[17] Wei, R. (2023) Dixit Player with Open Clip. *Journal of Data Analysis and Information Processing*, **11**, 536-547. https://doi.org/10.4236/jdaip.2023.114027

[18] Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., *et al.* (2022) Expanding Language-Image Pretrained Models for General Video Recognition. *Computer Vision—ECCV* 2022 17*th European Conference*, Tel Aviv, 23-27 October 2022, 1-18.

https://doi.org/10.1007/978-3-031-19772-7_1

[19] Wang, Y., Li, K.C., Li, Y.Z., He, Y.N., Huang, B.K., Zhao, Z.Y., *et al.* (2022) Internvideo: General Video Foundation Models via Generative and Discriminative Learning.

[20] Xu, J., Mei, T., Yao, T. and Rui, Y. (2016) MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 5288-5296. https://doi.org/10.1109/cvpr.2016.571

[21] Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T. and Russell, B. (2017) Localizing Moments in Video with Natural Language. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 5803-5812. https://doi.org/10.1109/iccv.2017.618

[22] Chen, D. and Dolan, W.B. (2011) Collecting Highly Parallel Data for Paraphrase Evaluation. *Proceedings of the* 49*th Annual Meeting of the Association for Computa*tional Linguistics: Human Language Technologies, Portland, 19-24 June 2011, 190-200.

[23] Sun, Q., Fang, Y.X., Wu, L., Wang, X.L. and Cao, Y. (2023) Eva-Clip: Improved Training Techniques for Clip at Scale.

[24] Liu, Y., Xiong, P., Xu, L., Cao, S. and Jin, Q. (2022) TS2-Net: Token Shift and Selection Transformer for Text-Video Retrieval. *Computer Vision—ECCV* 2022 17*th European Conference*, Tel Aviv, 23-27 October 2022, 319-335. https://doi.org/10.1007/978-3-031-19781-9_19