ISSN: 2710-0588

Keyframe Extraction Approaches for Videos under Multimodal Scenarios: A Survey

Chunlei Zhao, Ziqin Ye, Yuanyuan Cao, Zhan Zhang, and Liwei Wu*

North China University of Science and Technology, Tangshan 063210, China

*Corresponding Author

Abstract

Video keyframe extraction technology, as a core component of video summarization, retrieval, and content analysis, continues to receive widespread attention in the field of computer vision. With the rapid development of diverse applications such as smart cities, autonomous driving, and human-computer interaction, the sources and scenarios of video data have become increasingly complex, placing higher demands on keyframe extraction techniques. In dynamic scenes, factors such as intense environmental interference and highly variable motion of subjects often cause traditional extraction methods to face challenges like insufficient robustness and limited extraction accuracy. To address these issues, numerous intelligent extraction algorithms based on deep learning have emerged in recent years, significantly enhancing extraction performance in dynamic environments. By systematically reviewing keyframe extraction methods suitable for various dynamic scenes including surveillance, sports, and gesture analysis, and summarizing mainstream algorithmic models based on attention mechanisms, temporal modeling, and reinforcement learning, this paper analyzes their core concepts, advantages, and limitations. It concludes with experimental findings and offers personal insights, aiming to provide a clear reference framework and development direction for future research on efficient and robust video parsing technologies in dynamic scenes.

Keywords

Keyframes, Multimodal, Attention Mechanism.

1. Introduction

Video keyframe extraction has traditionally relied on methods such as frame difference, color histogram, and shot boundary detection. However, these conventional approaches often fall short in meeting the demands for accuracy, robustness, and semantic completeness in complex dynamic scenarios. For instance, in specialized applications like abnormal behavior identification in security surveillance, action analysis in sports, and road condition perception in autonomous driving, traditional keyframe extraction techniques struggle to effectively address multiple challenges, including sudden illumination changes, cluttered backgrounds, highly dynamic object motions, and semantic redundancy. To tackle these issues, researchers worldwide have continuously pursued in-depth studies. Over time, intelligent keyframe extraction techniques based on deep learning have emerged accordingly, gradually becoming a core research direction in the fields of computer vision and video analysis.

2. Key Frame Extraction in Surveillance Videos

Domestic and international scholars have investigated the impact of image preprocessing, motion object detection, global-local similarity fusion, and adaptive threshold setting on extraction effectiveness and storage efficiency in surveillance video key frame extraction. This

ISSN: 2710-0588

is achieved by combining PSNR and SURF features with an adaptive extraction strategy based on representativeness and independence.

Zhang Qiaoqiao et al.[1]proposed a spatiotemporal graph representation and visual attention-based method for extracting key frames from surveillance videos. The approach constructs a spatiotemporal graph where moving objects in the video serve as nodes and the feature distances between objects form the edge weights. It then employs a normalized graph cut algorithm to segment the video into visually consistent sub-shots. Finally, by integrating three visual attention constraints-content completeness, compactness of object distribution, and uniformity-the method selects the most representative key frames. Results demonstrate that this method effectively suppresses interference from static backgrounds in surveillance videos and addresses issues of key frame redundancy and omission of critical information inherent in traditional approaches that overlook inter-object relationships. Experiments on both self-built and public datasets show that the key frames extracted by this method outperform several baseline methods in terms of completeness and visual satisfaction. This research provides a reliable technical framework for video summarization and efficient browsing under fixed-camera scenarios.

Yuan Ting[2]proposed an intelligent security video surveillance system for smart airports based on key frame extraction. The system constructs comprehensive feature vectors by integrating color features and local binary pattern (LBP) texture features of video frames, uses these to partition video segments and extract key frames, and subsequently performs anomaly detection and risk warning in airport monitoring areas by calculating the feature vector distance between key frames and subsequent frames. Results demonstrate that across multiple monitoring scenarios of varying complexity-such as airport entrances and runways-the system significantly outperforms comparative systems based on optical character recognition and digital twin technologies in terms of the number of key objects captured, yielding results closer to manual statistical standards. This approach effectively enhances the utility and early warning capability of surveillance videos. The study provides a practical engineering solution for achieving comprehensive and efficient security monitoring in smart airports.

Zhang Jiayu[3]proposed a surveillance video key frame extraction framework based on time-frequency domain analysis, incorporating multiple transform domain methods and evaluation criteria. The approach addresses issues such as local detail loss, compromised image integrity, and illumination sensitivity in key frame extraction by employing techniques including fractional Fourier transform, multi-feature fusion with quaternion Fourier transform, and contourlet transform. Results demonstrate that the proposed method effectively improves the precision, recall, and F1-score of key frame extraction on public datasets such as VISOR and UCF-Crime. In complex scenarios, compared to traditional methods, it not only significantly enhances the capture capability of global and local motion states of targets but also strengthens robustness to illumination variations. Additionally, the study introduces a trajectory reconstruction evaluation criterion aligned with human visual perception. This research provides a systematic solution for efficient surveillance video summarization and performance evaluation.

Zhou Hanxing et al.[4]proposed a surveillance video keyframe extraction method combining PSNR and SURF features, further introducing an adaptive selection strategy based on representativeness and independence. The method first applies grayscale conversion and median filtering to the surveillance video images to reduce noise and illumination effects. It segments the video using a motion object detection algorithm integrating the Vibe algorithm and frame difference method, obtaining key sequences containing moving objects. To address the high computational complexity of traditional global similarity calculations, the method introduces normalized PSNR to describe global similarity and combines it with SURF features for local similarity calculation, fusing both to derive image similarity. The mean similarity value

ISSN: 2710-0588

serves as the threshold for keyframe extraction. Experimental results demonstrate that on a self-built surveillance video dataset, compared to methods based solely on SURF features or shot segmentation, this approach effectively improves the recall and precision rates of keyframe extraction. In system implementation, it significantly reduces the memory space required for video storage. This research provides an effective solution for efficient keyframe extraction and storage optimization in video surveillance systems.

In summary, the experimental results of the authors demonstrate strong consistency. The findings indicate that, compared to SURF-based and shot segmentation-based methods, the proposed approach effectively improves the recall and precision rates of keyframe extraction on a self-constructed surveillance video dataset while significantly reducing memory storage requirements. The quality of keyframe extraction is closely related to the representativeness and independence of video frames, and the introduction of an adaptive penalty term optimizes the compactness and content completeness of the keyframe set. This provides a reliable solution for efficient data processing and storage optimization in video surveillance systems.

3. Key Frame Extraction in Sports Videos

Researchers in the field of sports video keyframe extraction have dedicated efforts to enhancing the accuracy, semantic completeness, and real-time performance of keyframe extraction through approaches such as multimodal information fusion, action semantic analysis, and skeletal point feature optimization. Significant progress has been made in reducing redundancy, improving model generalization capability, and adapting to complex motion scenarios.

Yu Yixuan et al.[5]proposed a multi-modal hierarchical keyframe extraction method for continuous composite motions. Based on the spatial, temporal, and rhythmic characteristics of motion keyframes, the method designs a two-layer processing architecture from global to local. At the global level, it utilizes background music beats and joint spatiotemporal information to segment the motion sequence multi-modally, ensuring consistent spatial variation within segments. At the local level, parameter-adaptive spatial feature clustering is performed on each segment, combined with temporal segmentation to handle repetitive postures. Experimental results using broadcast gymnastics as an example demonstrate that this method significantly outperforms curve simplification, K-means clustering, and GKEN in terms of recall, precision, and F_1-score, effectively addressing the issues of large spatial variation and high redundancy in keyframes for continuous composite motions. This research provides a semantically consistent and hierarchically clear keyframe extraction framework for the automated evaluation of rhythmic continuous composite motions.

Zeng Zhonghui et al.[6]addressed issues such as high redundancy and lack of action semantics in production line surveillance videos by proposing a key frame extraction method based on action semantics. The approach involves extracting ORB local features from video frames, training an action semantic dictionary, generating global features via VLAD encoding, and finally employing K-means clustering to achieve precise and controllable key frame extraction. Experimental results demonstrate that even at a low compression ratio of 3.33%, the method fully preserves video action semantics, improving the F_1-score by 52.16% compared to baseline methods. It also outperforms optical flow analysis, VSUMM, and ResNet151V2-based clustering methods in terms of runtime and redundancy metrics. By constructing an action semantic space and aggregating global features, the study significantly enhances the efficiency and semantic fidelity of key frame extraction in repetitive structured scenarios.

Gao Xuexue et al.[7]developed a sports video keyframe extraction model based on skeletal keypoint features. The method utilizes OpenPose to extract human skeletal keypoints and constructs a motion feature matrix comprising limb angles, first-order differences, and dynamic peaks. An improved perceptual hash algorithm is introduced to calculate inter-frame visual

ISSN: 2710-0588

distance features, and a Random Forest model is ultimately employed for classification prediction. Experiments conducted on aerobics and cheerleading videos demonstrate that the model outperforms traditional methods based on SIFT-HD and STC in terms of both recall and precision, achieving a video compression rate of approximately 30% while maintaining the complete expression of motion content. By integrating motion features with ensemble learning, this research effectively enhances the accuracy and robustness of keyframe extraction, providing a practical solution for motion comparison and video retrieval.

In summary, the aforementioned studies have achieved consistent progress in sports keyframe extraction: The introduction of strategies such as multi-modal segmentation, action semantic dictionaries, and skeletal keypoint feature matrices has significantly enhanced the semantic representativeness and extraction accuracy of keyframes. Simultaneously, these methods demonstrate strong adaptability and efficiency in handling repetitive motions, complex movement sequences, and highly redundant video content. These approaches provide reliable technical support for applications such as motion analysis, industrial monitoring, and video summarization.

4. Key Frame Extraction in Surveillance Video

Researchers worldwide have continuously explored and optimized key performance aspects of video keyframe extraction-such as action representativeness, redundancy elimination, cross-scenario adaptability, and computational efficiency-by developing diverse pose estimation and feature extraction models.

Cai Guanlan[8]proposed a key frame extraction method for table tennis action videos by integrating flexible pose estimation and spatiotemporal features. The approach begins with non-uniform segmentation of video clips based on dense optical flow, detects spatiotemporal interest points using separable linear filters, and constructs spatiotemporal cubes to extract pixel-level features. Spatiotemporal feature edges are incorporated into pose estimation to ensure temporal and spatial continuity, while pose similarity is calculated via histogram intersection. Key frames are extracted based on a Hog vector difference matrix and threshold judgment. Experimental results show that the method achieves both fidelity and compression ratios exceeding 0.7 across diverse scenarios. Subjective evaluations confirm its effectiveness in extracting reasonable key frames, particularly maintaining high performance in complex scenes with rapid camera switching. This study provides a reliable technical pathway for accurately capturing key actions in dynamic sports videos.

Zhang Hongli et al.[9]designed a multi-feature fusion-based keyframe extraction system for high-dynamic dance videos. The system integrates color, texture, and shape feature vectors through weighted allocation to form a combined feature, which is then processed with a clustering algorithm to achieve keyframe extraction. The system adopts a B/S three-tier architecture, integrating video preprocessing, feature extraction, and keyframe selection modules. Experimental results demonstrate that in ballet video tests, the extracted keyframes completely matched the actual keyframes, with zero missed and false detections, achieving an accuracy of 100%. This validates the effectiveness of multi-feature fusion in enhancing the completeness and accuracy of keyframe extraction. The study provides a practical tool for learning and demonstrating high-dynamic dance videos, though its adaptability in low-quality, real-world video environments requires further improvement.

Jia Donglin[10]investigated a pose recognition-based method for extracting key frames from equine videos. The approach employs BlendMask for instance segmentation, uses the Canny operator to extract horse contours, constructs pose vectors using Fourier descriptors and head-to-body area ratios, and finally applies an SVM classifier to categorize three poses (standing, walking, and head-turning) for selecting key frames suitable for body measurement.

ISSN: 2710-0588

Experimental results show that the SVM classifier achieves 96.43% accuracy in equine pose recognition, with standing-pose key frame extraction accuracy nearing 97% and a Macro-F1 score of 0.9789 in multi-class evaluation. This study provides an automated, low-data-dependent solution for large animal video measurement, significantly reducing computational resource consumption.

Zhu Jianan[11]proposed a gait cycle clustering-based keyframe extraction algorithm for dance videos. The method first extracts the distance sequence between the dancer's feet using the OpenPose model and identifies the minimum points of the gait cycle to partition effective motion intervals. It then integrates color and shape features (histogram of oriented boundaries) and employs an improved adaptive K-means clustering algorithm to extract keyframes. Experimental results demonstrated that the algorithm successfully extracted 1,336 keyframes from a 2,052-frame dance video sequence with no redundancy or missed detection, achieving a compression rate exceeding 99.3% and a fidelity of approximately 0.8, outperforming all comparative algorithms. By combining cycle partitioning and multi-feature clustering, this study effectively addresses the issues of motion redundancy and insufficient representativeness in dance video analysis.

In summary, the experimental results from various authors show strong consistency. Studies demonstrate that by integrating spatiotemporal features with flexible pose estimation, employing weighted fusion of multiple visual features, utilizing contour analysis with machine learning classification, or applying gait cycle clustering, the accuracy, compression rate, and visual representativeness of video keyframe extraction can be significantly enhanced. Consistent progress has also been made in improving adaptability to complex scenarios, reducing redundancy, and lowering computational costs.

5. Conclusion

In summary, a review of existing literature reveals that although significant progress has been made in multimodal video keyframe extraction technology, several challenges remain to be thoroughly addressed when dealing with complex dynamic scenes. Current models still lack a deep understanding of high-level semantic information in videos under strong interference conditions such as sudden illumination changes and rapidly moving targets, which directly affects the semantic representativeness and completeness of selected keyframes. Furthermore, existing methods generally exhibit weak cross-scene adaptability, with limited generalization capabilities across diverse scenarios such as surveillance, sports, and film production. The absence of a unified and comprehensive evaluation system also hinders effective technical comparison and iterative optimization. On the practical application front, balancing real-time processing efficiency and lightweight model design while ensuring keyframe extraction accuracy and integrity remains a critical bottleneck for large-scale deployment in resource-constrained edge devices or real-time streaming scenarios. Future research could focus on cross-scene adaptive learning, lightweight network architecture design, and the establishment of a unified evaluation standard incorporating semantic importance.

Acknowledgments

Basic scientific research business cost project of provincial universities of North China University of Science and Technology(JJC2024067).

Study on the Behavior and Design Method of Grouted Sleeve Connections for Steel Tubes under Combined Compression, Bending, and Torsion(S202510081040).

ISSN: 2710-0588

References

[1] ZHANG Qiaoqiao. Research on Key Frame Extraction from Surveillance Videos Based on Spatiotemporal Graph Representation[D]. Anhui University, 2018.

[2] YUAN Ting. Intelligent Airport Security Video Surveillance System Based on Key Frame Extraction[J]. Information and Computer(Theoretical Edition), 2022, 34(24): 59-61.

[3] ZHANG Jiayu. Research on Key Frame Extraction from Surveillance Videos Based on Time-Frequency Domain Analysis[D]. Shijiazhuang Tiedao University, 2023. DOI:10.27334/d.cnki.gstdy.2023.000798.

[4] ZHOU Hanxing. Research and System Implementation of Key Frame Extraction Technology in Video Surveillance[D]. Chongqing University of Posts and Telecommunications, 2018. DOI:10.27675/d.cnki.gcydx.2018.000115.

[5] YU Yixuan, YANG Geng, GENG Hua. Multimodal Hierarchical Key Frame Extraction Method for Continuous Compound Motion[J]. Journal of Shandong University(Engineering Science), 2023, 53(02): 42-50.

[6] ZENG Z H, XIANG H, LIN Z C, et al. A keyframe extraction method for production line videos based on action semantics[J]. Manufacturing Technology & Machine Tool, 2025, (05): 172-180. DOI:10.19287/j.mtmt.1005-2402.2025.05.023.

[7] GAO X X, GU L. A sports video keyframe extraction model based on skeletal keypoint features[J]. Foreign Electronic Measurement Technology, 2022, 41(09): 88-94. DOI:10.19652/j.cnki.femt.2203974.

[8] CAI G L. Keyframe extraction for table tennis action video clips combining flexible pose estimation and spatiotemporal features[J]. Science Technology and Engineering, 2019, 19(25): 268-272.

[9] ZHANG H L. A keyframe extraction system for high-dynamic dance videos based on multi-feature fusion[J]. Techniques of Automation & Applications, 2022, 41(06): 91-94+116. DOI:10.20033/j.1003-7241.(2022)06-0091-05.

[10] JIA Donglin, ZHANG Jingjing, LI Quansheng, et al. Pose Recognition-Based Key Frame Extraction Method for Equine Videos[J]. Computer and Digital Engineering, 2025, 53(01): 263-268.

[11] ZHU Jianan. Dance Video Key Frame Extraction Algorithm Based on Gait Cycle Clustering[J]. Information Technology, 2025, (06): 101-106. DOI:10.13274/j.cnki.hdzj.2025.06.017.