Jornadas de Automática

Human-Centric Video Summarization via Identity-Aware Tracking

Mirjalili, Milad[a,*], Alegre, Enrique[a], Fidalgo, Eduardo[a], González-Castro, Víctor[a], Tanveer, Waqar [a]

[a] *Department of Electrical, Systems and Automation, Universidad de León, León, España.*

To cite this article: Mirjalili, Milad, Alegre, Enrique, Fidalgo, Eduardo, González-Castro, Víctor, Tanveer, Waqar. 2025. Human-Centric Video Summarization via Identity-Aware Tracking. Jornadas de Automática, 46. https://doi.org/10.17979/ja-cea.2025.46.12249

Resumen

En este artículo, presentamos un enfoque para el resumen de videos en base a la presencia e identidad de las personas a lo largo de los fotogramas. El enfoque propuesto combina puntos de referencia de la pose, representaciones faciales detalladas y características visuales del cuerpo para construir una representación robusta de cada persona detectada. Estas características se agrupan de forma offline para realizar un seguimiento consistente de los individuos a lo largo del video. Nuestro método no requiere datos etiquetados, lo que lo hace adecuado para procesar colecciones de video a gran escala sin necesidad de anotaciones. Al seleccionar fotogramas representativos donde los individuos clave aparecen con mayor frecuencia, el sistema genera resúmenes concisos y conscientes de la identidad que reflejan la dinámica de la presencia humana a lo largo del tiempo. Ejecutamos experimentos en diversas secuencias de video y logramos una puntuación F1 promedio del 99.4% para el seguimiento consistente de identidades. Esta estrategia centrada en la persona ofrece una solución escalable y generalizable para resumir videos en dominios donde comprender la actividad humana es esencial.

Palabras clave: Visión por computadora, Interacción humano-máquina (HMI), Diseño centrado en el ser humano, Aprendizaje profundo, Inteligencia artificial (IA)

Human-Centric Video Summarization via Identity-Aware Tracking

Abstract

In this paper, we present an approach to video summarization that focuses on the presence and identity of people across video frames. The proposed framework combines pose landmarks, rich facial embeddings, and visual appearance features of the body to build a robust representation for each detected person. These features are clustered offline to enable consistent tracking of individuals throughout the video. Our method does not require labeled data, making it suitable for processing large-scale video collections without the need for annotations. By selecting representative frames in which key individuals appear most frequently, the system generates concise and identity-aware summaries that reflect the dynamics of human presence over time. We conducted experiments on diverse video sequences and achieved an average F1 score of 99.4% for consistent identity tracking. This person-centric strategy offers a scalable and generalizable solution for summarizing videos in domains where understanding human activity is essential.

Keywords: Computer vision, Human-machine interaction (HMI), Human-centric design, Deep learning, Artificial intelligence (AI)

1. Introduction

The growing volume of video content, from social media clips to surveillance footage and instructional recordings, has increased the need for effective video summarization (VS) techniques (Meena et al., 2023). These methods aim to reduce visual redundancy while preserving the most informative and relevant segments of a video. In many real-world scenarios,

*Autor para correspondencia: mmir@unileon.es the presence, activity, and interaction of people form the central narrative of a video, such as security monitoring and activity-centered recordings (Yang et al., 2016).

Traditional VS methods often emphasize low-level features like motion or scene transitions. While effective in certain cases, they tend to overlook the importance of human presence and interactions, particularly when individuals are not visually prominent or appear under challenging conditions such as varying poses and partial occlusions (Biswas et al., 2021).

In this work, we propose a human-focused, offline VS pipeline that detects, represents, and tracks individuals throughout a video to create concise and interpretable summaries. Our strategy integrates several modern computer vision components, including body pose landmarks (Varghese and M., 2024), facial embeddings (Deng et al., 2019), and spatially localized appearance features of the body (Varghese and M., 2024). These signals are fused into a robust representation for each person, which is then clustered offline to assign consistent identity labels across the entire video.

By quantifying how often and where individuals appear, the method selects representative frames that capture key people and moments of interaction. This enables identity-aware summarization without relying on scene transitions or predefined activity labels. Importantly, the system operates without any labeled data, making it scalable for large and diverse video datasets.

We demonstrate that our method generates interpretable and compact summaries aligned with human presence and activity. This method is especially effective for multi-person scenes where people may appear at different times, in varying poses or partial occlusions, and where face-only detection would fail.

Therefore, our main contributions are:

- A pipeline that combines body pose, face embeddings, and spatial body features to form robust person descriptors.
- An offline identity clustering approach for consistent person tracking.
- A summary generation method that highlights key individuals and interactions, even under imperfect tracking conditions.

The rest of the paper is organized as follows: Section 2 reviews related work in VS and human-centric analysis. Section 3 describes the proposed algorithm in detail. Section 4 presents experimental results and qualitative evaluations. Section 5 concludes with final observations and outlines future directions.

2. **Related work**

Generating video summaries is an active research area, as evidenced by the increasing number of research papers published each year (Alaa et al., 2024), with various approaches being proposed. This section reviews related works in VS, person detection and tracking, and person-centric analysis.

2.1. General VS

Conventional VS methods can broadly be categorized into different learning-based approaches, depending on the availability of annotated datasets (Meena et al., 2023; Tiwari and Bhatnagar, 2021):

Supervised methods often rely on training data annotated with important segments or frames to learn spatiotemporal relationships between video frames (Gygli et al., 2014; Yale Song et al., 2015). For example, some approaches use recurrent neural networks (RNNs) or Transformers to predict frame importance based on learned representations (Fajtl et al., 2019; Hsu et al., 2023; Ke Zhang et al., 2016).

Unsupervised methods, on the other hand, identify patterns without requiring labeled data. Early works often focused on visual saliency, motion, or audio cues to detect important parts of a video (Paul and Musfequs Salehin, 2019). More recently, deep learning-based methods have utilized rich representations extracted from video data. These approaches typically employ techniques like clustering to group frames with similar features (Basavarajaiah and Sharma, 2021; Zhao et al., 2015), or generative models such as Generative Adversarial Networks (GANs) to learn the distribution of important frames (Apostolidis et al., 2021; Li et al., 2024).

Weakly supervised learning uses imperfect annotations, such as video titles or tags, to guide summary generation(Argaw et al., 2024; Ramos et al., 2023). The goal is to train models simultaneously on multimodal data (e.g., text and video) to align them in a shared semantic space.

Reinforcement learning (RL) has also been used to train agents that optimize summary generation by maximizing a reward function based on criteria such as representativeness and diversity of the selected frames (Liu et al., 2022).

2.2. Person detection and tracking

Robust person detection and tracking form the foundation of our methodology. Object detection models, such as the YOLOv8 (Varghese and M., 2024), have revolutionized real-time object detection due to their speed and accuracy. Similarly, pose estimation models such as OpenPose (Cao et al., 2021) and BlazePose (Bazarevsky et al., 2020) provide detailed body keypoints, enabling a richer understanding of human actions and postures. For face detection, models like MTCNN (Kaipeng Zhang et al., 2016) and BlazeFace (Bazarevsky et al., 2019) offer efficient and accurate localization, which is crucial for extracting facial features. Multi-object tracking (MOT) algorithms such as Deep SORT (Wojke et al., 2017) aim to maintain consistent identities of objects across frames.

2.3. Person-centric VS

This subcategory generates summaries with a focus on humans and can be categorized into several types:

Detection-based summarization: Some research focuses on identifying and tracking specific individuals of interest throughout a video and then summarizing their appearances. For example, (Biswas et al., 2021) use face detection to summarize frames based on the presence of faces. The resulting summaries can be further used in downstream tasks like law enforcement applications (Chaves et al., 2020; Gangwar et al., 2017). While effective in some domains like interviews or vlogs, face-only methods suffer from limitations when people are turned away, occluded, or partially out of frame—leading to missed detections or fragmented presence tracking.

Interaction-based summarization: Some methods prioritize frames or segments where multiple people interact. This can involve detecting specific social actions (e.g., handshake, conversation), or simply detecting when people appear close together in space or time (Yang et al., 2016).

Event-driven summarization: In applications such as surveillance or sports analysis, summaries are created by detecting specific events involving people—like a goal in soccer or suspicious behavior. These methods typically combine person detection with action recognition to identify and select the most relevant frames (U. and Kovoor, 2021).

Query-based summarization: More advanced systems allow users to input a query, generating summaries that align with the user's preferences. These networks often employ multimodal transformers to combine different modalities, producing aligned text-image summaries (Radford et al., 2021; Wu et al., 2022).

3. **Methodology**

Our proposed method distinguishes itself by combining detailed body, pose, and face representations to enable robust person identification via offline clustering. This comprehensive approach, combined with a flexible summary frame selection strategy driven by appearance duration, frequency, and interaction, offers a robust framework for generating interpretable summaries centered around human activity. A summary of the model is shown in Figure 1.

3.1. *Person Detection and Integrated Feature Extraction*

Each frame is processed by YOLO8s-Pose (Jocher et al., 2023) to detect and localize person bounding boxes. Once persons are detected, we extract a comprehensive feature vector for each individual, consisting of:

Pooled Body Features: Extracted from three intermediate YOLO feature maps, these capture important body appearance cues. After alignment, pooling, and concatenation, the final tensor has shape [N, 1152].

Pose Keypoints: Estimated from YOLO's output, providing 17 body landmark coordinates per person along with confidence scores. The resulting tensor shape is [N, 51]

Face embeddings: If the confidence scores for detected face keypoints indicate a face is present, we use BlazeFace (Bazarevsky et al., 2019) to obtain precise facial bounding boxes. Then, a 512-dimensional ArcFace embedding (Deng et al., 2019) is computed. In cases where a face is not detected or is unreliable, the embedding is filled with zeros, and a binary flag (1.0 for detected face, 0.0 for no face) is appended, extending the embedding dimension to 513. We also introduce a weighting factor to control the influence of this embedding on the final feature representation. The resulting tensor is in the shape of [N, 513].

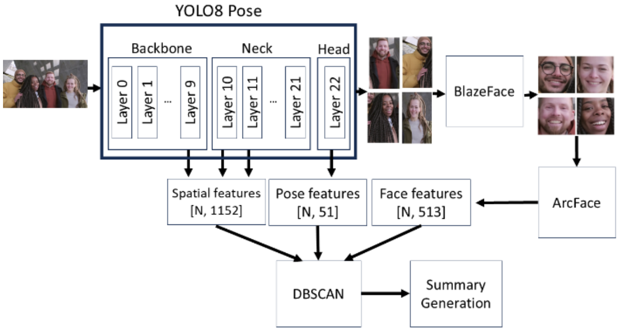

Figure 1: Overview of the proposed person-centric VS pipeline. Each frame undergoes integrated body, pose, and face feature extraction, followed by offline identity clustering using DBSCAN, and selection of a representative summary based on presence frequency, interaction density, and temporal diversity.

The final concatenated feature vector for each person detection is a 1716-dimensional vector.

3.2. *Offline Clustering and Identity Assignment*

Features from all frames are collected into a single global dataset. This enables an offline clustering approach, which provides greater robustness than online tracking by considering the entire temporal context, though it requires more memory. The high-dimensional feature vectors are then used for identity assignment. We employ Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996) for clustering, as it effectively identifies density-based clusters and handles noise. The key DBSCAN parameters are epsilon, which defines the maximum distance between two samples for them to be considered neighbors, and min_samples, which specifies the minimum number of samples required to form a cluster. Each resulting cluster corresponds to a unique person identity.

3.3. *Summary frame selection*

Although our system uses rich person descriptors and offline clustering for identity assignment, achieving perfect consistency is difficult in cases of occlusion or abrupt environmental changes. To address these limitations, our summarization strategy is designed to remain effective even when identity assignments are imperfect. Instead of relying solely on exact identity tracking, the system generates concise visual summaries based on patterns of human presence and interaction.

We compute a score for each frame based on the following complementary criteria:

Presence Frequency: We measure how often each person appears by counting the unique frames they occupy. Frames showing frequently appearing individuals are given higher scores, highlighting the main characters or subjects in the video.

Interaction Density: Frames where multiple people appear together are identified as potential key moments of interaction or important events. Such frames receive higher scores, emphasizing social or contextually rich segments.

Temporal Diversity: To cover the entire video and avoid repeating similar content, we sample top-scoring frames evenly throughout the timeline. This ensures that the summary reflects the different stages, contexts, and activities over time.

Final summaries are generated by selecting these representative frames, balancing individual presence, social interactions, and coverage over time.

4. Experiments and results

We evaluate our method on a diverse set of videos featuring people in varied contexts, including different numbers of individuals, movement patterns, and scene dynamics. These videos present challenges such as occlusions, varying lighting conditions, and frequent camera transitions.

4.1. Experimental setup

We performed experiments on a workstation equipped with an NVIDIA RTX 2080 Super GPU, running Windows 11, with an Intel Core i7-10750H CPU operating at a base frequency of 2.60 GHz, and 32 GB of RAM. The pipeline was implemented in Python 3.11, using PyTorch 2.7 with CUDA support for model development and processing.

We selected five videos for evaluation, downloaded from YouTube, featuring adults from diverse ethnic backgrounds. Table 1 summarizes the key video properties, including total processed frames, duration, resolution, number of unique individuals, and the availability of pose and face data. We processed each video at a frame rate of 8 frames per second (FPS), resizing frames to 640 × 640 pixels as input to the YOLO model.

Table 1: Characteristics of the videos used for evaluation. Each row summarizes key properties, including total frames processed, duration, resolution, number of unique individuals, and availability of pose and face data.

Name	#Frames	Video duration (s)	Video resolution (px)	#Unique persons	Person availability (%)	Face availability (%)
Test 1	166	20	320 × 240	1	39.76	39.76
Test 2	341	42	426 × 226	2	100	99.85
Test 3	77	9	3840 × 2160	4	100	96.42
Test 4	414	51	1280 × 720	6	76.81	75.36
Test 5	695	86	1280 × 720	6	55.89	49.05

We evaluate the performance of our offline person tracking system by measuring how consistently it detects and assigns unique identities across frames. The evaluation is conducted based on manually verified identity assignments, obtained by visually inspecting the annotated outputs. Table 2 reports key metrics computed for each video: the number of clusters formed (including noise), the number of unique individuals detected, as well as precision, recall, F1 score, and silhouette score. Precision, recall, and F1 score estimate how accurately individuals are identified, while the silhouette score reflects the separation quality of the formed clusters. These metrics provide both quantitative and qualitative insights into the reliability of the tracking results.

Table 2: Quantitative evaluation of person identity assignment across videos. The table reports the number of clusters formed (including noise), the number of unique individuals detected, and performance metrics including precision, recall, F1 score, and silhouette score.

Name	#Clusters formed	#Unique persons detected	Precision (%)	Recall (%)	F1 Score (%)	Silhouette Score
Test 1	1	1	100	100	100	N/A
Test 2	2	2	100	100	100	0.83
Test 3	4	4	99.7	99.7	99.7	0.58
Test 4	**7**	**7**	97.6	97.6	97.6	0.45
Test 5	**7**	6	100	99.5	99.7	0.63

4.2. Quantitative results

We observe near-perfect performance in Test 1 and Test 2, where only one and two individuals appear, respectively. These simple cases yield ideal clustering with 100% precision and recall, with the number of clusters perfectly matching the number of ground-truth identities. In these scenarios, the model consistently assigns identities without confusion.

As the number of individuals increases (Test 3–Test 5), the system maintains high accuracy, achieving F1 scores above 97%, which demonstrates its robustness in more complex scenarios. For example, In Test 5, the system forms more clusters (7) than the actual number of individuals (6) but still achieves 100% precision and 99.5% recall, leading to a high F1 score of 99.7%.

The silhouette scores show a decrease in inter-cluster separability as the number of individuals increases, which is expected due to higher visual similarity and occlusions. Nonetheless, a silhouette score of 0.63 in Test 5 suggests that the clusters remain well-separated even in moderately crowded scenes.

Overall, the results validate the accuracy and scalability of our identification and clustering approach across different scenarios.

4.3. Qualitative analysis

To complement our quantitative results, we now present qualitative examples illustrating the model's performance under various scene conditions. In Test 2, we observe two unique individuals in a setting with minimal scene dynamics and a stationary camera. Although in some frames the model cannot rely on face embeddings due to occlusion, the remaining features still manage to bring the new observations close to previous ones, resulting in consistent identity assignments.

In another video (Test 3), there are four unique individuals, and the scene involves camera movement along with partial face occlusions. To visualize how the high-dimensional person descriptors are organized, we apply Principal Component Analysis (PCA) to project them into a two-dimensional space. As shown in Figure 3, we can see that the features form distinct groups, giving insight into how the model separates different identities, even under more dynamic and challenging conditions.

To investigate the slight performance drop observed in Test 4, we conduct a frame-by-frame analysis. We find that in the initial frames, one person's face is heavily occluded, with limited visibility of body parts and shadowing that affects the quality of extracted features. However, as the video progresses, the model can gather more distinctive features, eventually forming a clear identity separation.

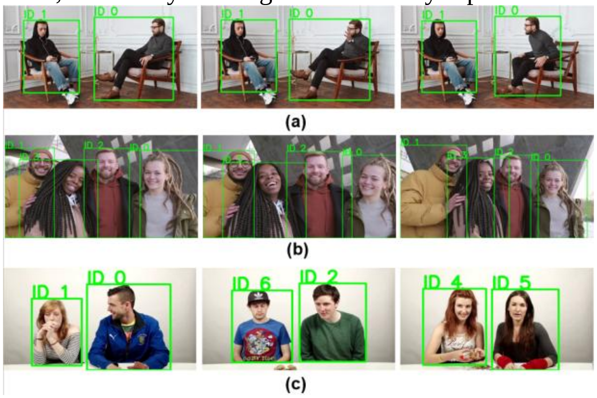

Figure 2: Three test videos: (a) Test 2, (b) Test 3, and (c) Test 4.

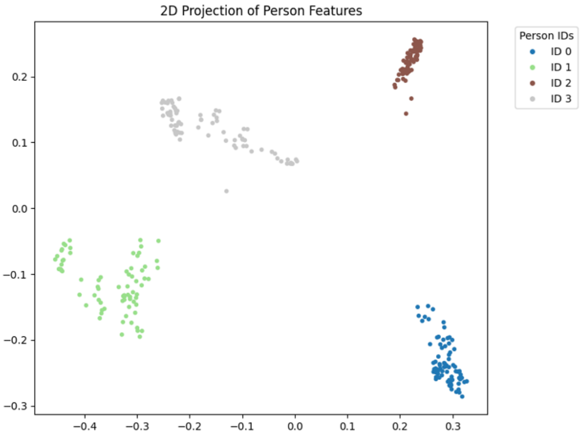

Figure 3: 2D projection of features for sequence Test 3, showing four distinct clusters corresponding to four unique individuals.

The final video we analyze corresponds to the summary output generated for Test 5. In this sequence, two main individuals appear prominently in the foreground, while a third person, positioned in the background and facing away from the camera, occupies only a small portion of the frame. Additionally, the video includes multiple scene transitions, and there are three segments where only a hand is visible.

Despite these challenges, our system effectively incorporates such subtle visual cues into the final summary. This highlights its ability to focus on subtle cues and partial appearances when forming person-aware summaries.

In our summary generation, we focus on frames where at least a part of a person's body is visible. We also assign higher weights to frames where more than two individuals are detected, as these often indicate potential interactions. To ensure a diverse selection of frames and avoid redundancy, we divide the video into time spans and choose high-scoring frames from each segment accordingly.

Two examples of the resulting summaries are shown in Figure 4, showing how our method emphasizes relevance, interaction, and temporal variety.

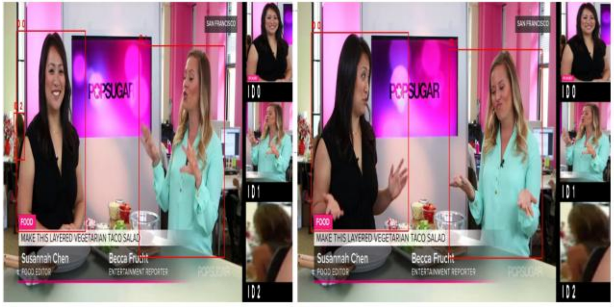

Figure 4: Two examples of generated video summaries, focusing on frame selection based on person-centric scoring. A legend indicates the first appearance of each person's identity for visual reference.

5. Conclusion

In this work, we present a novel approach to VS that centers on the identities of individuals appearing in the video. By using person detection followed by integrated feature representation and identity clustering, our method achieves robust and accurate person tracking across diverse and challenging scenarios. For future work, we plan to use temporal models such as transformers to also accommodate the temporal correlation of images existing in video frames. We also have plans to work on an online tracking algorithm, which will run in real-time, assigning identities frame by frame.

Acknowledgements

This work has been funded by the Recovery, Transformation, and Resilience Plan, financed by the European Union (Next Generation), thanks to the LUCIA project (Fight against Cybercrime by applying Artificial Intelligence) granted by INCIBE to the University of León.

References

Alaa, T., Mongy, A., Bakr, A., Diab, M., Gomaa, W., 2024. Video Summarization Techniques: A Comprehensive Review. https://doi.org/10.48550/ARXIV.2410.04449

Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I., 2021. AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization. IEEE Trans. Circuits Syst. Video Technol. 31, 3278–3292. https://doi.org/10.1109/TCSVT.2020.3037883

Argaw, D.M., Yoon, S., Heilbron, F.C., Deilamsalehy, H., Bui, T., Wang, Z., Dernoncourt, F., Chung, J.S., 2024. Scaling Up Video Summarization Pretraining with Large Language Models, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition

(CVPR), IEEE, Seattle, WA, USA, pp. 8332–8341. https://doi.org/10.1109/CVPR52733.2024.00796

Basavarajaiah, M., Sharma, P., 2021. GVSUM: generic video summarization using deep visual features. Multimed. Tools Appl. 80, 14459–14476. https://doi.org/10.1007/s11042-020-10460-0

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T.L., Zhang, F., Grundmann, M., 2020. BlazePose: On-device Real-time Body Pose tracking. ArXiv abs/2006.10204.

Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., Grundmann, M., 2019. BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. https://doi.org/10.48550/ARXIV.1907.05047

Biswas, R., Chaves, D., Fernández-Robles, L., Fidalgo, E., Alegre, E., 2021. A Video Summarization Approach to Speed-up the Analysis of Child Sexual Exploitation Material, in: XLII JORNADAS DE AUTOMÁTICA : LIBRO DE ACTAS. Servizo de Publicacións da UDC, pp. 648–654. https://doi.org/10.17979/spudc.9788497498043.648

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y., 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. IEEE Trans. Pattern Anal. Mach. Intell. 43, 172–186. https://doi.org/10.1109/TPAMI.2019.2929257

Chaves, D., Fidalgo, E., Alegre, E., Alaiz-Rodríguez, R., Jáñez-Martino, F., Azzopardi, G., 2020. Assessment and Estimation of Face Detection Performance Based on Deep Learning for Forensic Applications. Sensors 20, 4491. https://doi.org/10.3390/s20164491

Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, pp. 4685–4694. https://doi.org/10.1109/CVPR.2019.00482

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96. AAAI Press, Portland, Oregon, pp. 226–231.

Fajtl, J., Sokeh, H.S., Argyriou, V., Monekosso, D., Remagnino, P., 2019. Summarizing Videos with Attention, in: Carneiro, G., You, S. (Eds.), Computer Vision – ACCV 2018 Workshops, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 39–54. https://doi.org/10.1007/978-3-030-21074-8_4

Gangwar, A., Fidalgo, E., Alegre, E., González-Castro, V., 2017. Pornography and child sexual abuse detection in image and video: a comparative evaluation, in: 8th International Conference on Imaging for Crime Detection and Prevention (ICDP 2017). Presented at the 8th International Conference on Imaging for Crime Detection and Prevention (ICDP 2017), Institution of Engineering and Technology, Madrid, Spain, pp. 37–42. https://doi.org/10.1049/ic.2017.0046

Gygli, M., Grabner, H., Riemenschneider, H., Van Gool, L., 2014. Creating Summaries from User Videos, in: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 505–520. https://doi.org/10.1007/978-3-319-10584-0_33

Hsu, T.-C., Liao, Y.-S., Huang, C.-R., 2023. Video Summarization With Spatiotemporal Vision Transformer. IEEE Trans. Image Process. 32, 3013–3026. https://doi.org/10.1109/TIP.2023.3275069

Jocher, G., Qiu, J., Chaurasia, A., 2023. Ultralytics YOLO.

Li, H., Klabjan, D., Utke, J., 2024. Unsupervised Video Summarization via Iterative Training and Simplified GAN, in: Proceedings of the Asian Conference on Computer Vision (ACCV). pp. 1585–1601.

Liu, T., Meng, Q., Huang, J.-J., Vlontzos, A., Rueckert, D., Kainz, B., 2022. Video Summarization Through Reinforcement Learning With a 3D Spatio-Temporal U-Net. IEEE Trans. Image Process. 31, 1573–1586. https://doi.org/10.1109/TIP.2022.3143699

Meena, P., Kumar, H., Kumar Yadav, S., 2023. A review on video summarization techniques. Eng. Appl. Artif. Intell. 118, 105667. https://doi.org/10.1016/j.engappai.2022.105667

Paul, M., Musfequs Salehin, Md., 2019. Spatial and Motion Saliency Prediction Method Using Eye Tracker Data for Video Summarization. IEEE Trans. Circuits Syst. Video Technol. 29, 1856–1867. https://doi.org/10.1109/TCSVT.2018.2844780

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., others, 2021. Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning. PmLR, pp. 8748–8763.

Ramos, W., Silva, M., Araujo, E., Moura, V., Oliveira, K., Marcolino, L.S., Nascimento, E.R., 2023. Text-Driven Video Acceleration: A Weakly-Supervised Reinforcement Learning Method. IEEE Trans. Pattern Anal. Mach. Intell. 45, 2492–2504. https://doi.org/10.1109/TPAMI.2022.3157198

Tiwari, V., Bhatnagar, C., 2021. A survey of recent work on video summarization: approaches and techniques. Multimed. Tools Appl. 80, 27187–27221. https://doi.org/10.1007/s11042-021-10977-y

U., S.M., Kovoor, B.C., 2021. An aggregated deep convolutional recurrent model for event based surveillance video summarisation: A supervised approach. IET Comput. Vis. 15, 297–311. https://doi.org/10.1049/cvi2.12044

Varghese, R., M., S., 2024. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness, in: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). Presented at the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), IEEE, Chennai, India, pp. 1–6. https://doi.org/10.1109/ADICS58448.2024.10533619

Wojke, N., Bewley, A., Paulus, D., 2017. Simple online and realtime tracking with a deep association metric, in: 2017 IEEE International Conference on Image Processing (ICIP). Presented at the 2017 IEEE International Conference on Image Processing (ICIP), IEEE, Beijing, pp. 3645–3649. https://doi.org/10.1109/ICIP.2017.8296962

Wu, G., Lin, J., Silva, C.T., 2022. Intentvizor: Towards generic query guided interactive video summarization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10503–10512.

Yale Song, Vallmitjana, J., Stent, A., Jaimes, A., 2015. TVSum: Summarizing web videos using titles, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA, pp. 5179–5187. https://doi.org/10.1109/CVPR.2015.7299154

Yang, J.-A., Lee, C.-H., Yang, S.-W., Somayazulu, V.S., Chen, Y.-K., Chien, S.-Y., 2016. Wearable social camera: Egocentric video summarization for social interaction, in: 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Presented at the 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, Seattle, WA, USA, pp. 1–6. https://doi.org/10.1109/ICMEW.2016.7574681

Zhang, Ke, Chao, W.-L., Sha, F., Grauman, K., 2016. Video Summarization with Long Short-Term Memory, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 766–782. https://doi.org/10.1007/978-3-319-46478-7_47

Zhang, Kaipeng, Zhang, Z., Li, Z., Qiao, Y., 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Process. Lett. 23, 1499–1503.

Zhao, Y., Lv, G., Ma, T., Ji, H., Zheng, H., 2015. A novel method of surveillance video Summarization based On clustering and background subtraction, in: 2015 8th International Congress on Image and Signal Processing (CISP). Presented at the 2015 8th International Congress on Image and Signal Processing (CISP), IEEE, Shenyang, China, pp. 131–136. https://doi.org/10.1109/CISP.2015.7407863