Efficient visual information indexation for supporting actionable intelligence and knowledge generation

Desale F. Nurye[a], Yagmur Aktas[b], and Jorge García[c]

[a, b, c]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián (Spain)

ABSTRACT

Visual indexing, or the ability to search and analyze visual media such as images and videos, is important for law enforcement agencies because it can speed up criminal investigations. As more and more visual media is created and shared online, the ability to effectively search and analyze this data becomes increasingly important for investigators to do their job effectively. The major challenges for video captioning include accurately recognizing the objects and activities in the image, understanding their relationships and context, generating natural and descriptive language, and ensuring the captions are relevant and useful. Near real-time processing is also required in order to facilitate agile forensic decision making and prompt triage, hand-over and reduction of the amount of data to be processed by investigators or subsequent processing tools. This paper presents a captioning-driven efficient video analytic which is able to extract accurate descriptions of images and videos files. The proposed approach includes a temporal segmentation technique providing the most relevant frames. Subsequently, an image captioning approach has been specialized to describe visual media related to counter-terrorism and cyber-crime for each relevant frame. Our proposed method achieves high consistency and correlation with human summary on SumMe dataset, outperforming previous similar methods.

Keywords: Visual indexing , temporal segmentation, key frame selection, video captioning

1. INTRODUCTION

In the midst of the ongoing digital revolution, there has been an unprecedented surge in the volume of visual content such as images and videos being disseminated online and collected from surveillance systems. This has reshaped investigative operations, notably in sectors such as counter-terrorism and cybercrime. Modern law enforcement agencies are capitalizing on this profusion of visual data, transforming it into a pivotal tool for safeguarding the public and enforcing justice. Navigating through this extensive sea of digital visual content has been crucial in accelerating investigations and improving their effectiveness. This task of filtering, categorizing, and analyzing this enormous amount of visual data is termed visual indexing. It offers an efficient way for law enforcement to handle and interpret this immense volume of visual content, thereby enhancing the speed and precision of criminal investigations and facilitating the swift identification and capture of suspects.

Yet, visual indexing [1], especially for videos, is no simple task due to the immense volume and inherent complexity of the visual content. Unlike static images, videos are frame sequences with high temporal redundancy, indicating minor scene or subject changes. Analyzing each frame can result in duplication of efforts, not a suitable approach given the copious video data often handled by law enforcement.

Content-based visual indexing uses various techniques to catalog and recall images and videos based on their visual content, thereby improving the speed and precision of database access. Conventional methods reliant on low-level visual features like spatial relationship, color, texture, shape, and object motion analysis serve as reliable content indexes. However, they might not capture the semantic richness of images or videos. Addressing this issue is the emergence of image or video captioning in visual indexing, a fundamental aspect of semantic indexing and a subfield of content-based image retrieval. Semantic indexing techniques aim to decode the inherent meaning of visual data, surpassing basic visual feature constraints. By incorporating Natural Language Processing (NLP) and Computer Vision (CV), particularly deep learning-based techniques, these systems can generate descriptive captions, reflecting the content of an image or video. These context-rich captions act as meta-tags, offering a powerful and human-like indexing system. Despite the challenges posed by semantic indexing complexity and the

Artificial Intelligence for Security and Defence Applications, edited by Henri Bouma, Judith Dijk, Radhakrishna Prabhu, Robert J. Stokes, Yitzhak Yitzhaky, Proc. of SPIE Vol. 12742, 1274207 · © 2023 SPIE · 0277-786X · doi: 10.1117/12.2680257

need for sophisticated models and large training data, caption-based semantic indexing adoption is increasing, promising a more comprehensive and contextually-aware content-based visual indexing approach [2].

Integrating caption-based semantic indexing could transform law enforcement investigative procedures. With increasing digital data, especially image and video evidence, the need for its efficient retrieval and precise analysis is crucial. Auto-generating contextual captions for images or videos significantly speeds up the search process, allowing investigators to retrieve relevant visual data using caption keywords. This saves time and improves crime investigation accuracy, as evidence can be cross-verified using diverse caption data points. Furthermore, the use of advanced machine learning techniques could reveal overlooked patterns or correlations during manual examinations. Hence, incorporating semantic indexing in law enforcement procedures has the potential to enhance both the speed and quality of criminal investigations.

Despite the considerable advancements in visual indexing methodologies, gaps remain in catering to the specific requirements of forensic analysis. Our study seeks to address these gaps by developing a tailored approach that integrates the strengths of content-based visual indexing, deep learning, video summarization, and image captioning techniques. Our method aims to extract relevant video data information by dividing it into meaningful segments and selecting key frames, representing each segment. By focusing on these key frames summarizing the video content, we can considerably reduce the computational load for video caption generation without losing crucial information. The main contributions of the proposed framework are:

- We proposed a novel approach for efficient video indexation that combines keyframe selection based video summarization and image captioning techniques.
- In the proposed framework, we introduced an integrated approach to temporal segmentation and keyframe selection to summarize videos for video captioning. Our proposed video summarizer module generate a human like video summary in different video settings(egocentric, moving and static camera). A qualitative and quantitative analysis is performed for SumMe dataset.
- We fine-tuned a grounded image captioning model on proprietary dataset of counter-terrorism and cybercrime which ensure the proposed indexation framework is able to generate captions that are relevant to forensic video indexation.

2. RELATED WORK

In a comprehensive reviews of recent advancements in the field of content-based visual indexing [1,3,4] examines various components of the CBIR process including methodologies for image and video indexing, strategies for measuring similarity among images or videos, as well as the process of search re-ranking. Content-based visual indexing using both classical and deep learning methods has been widely used to represent visual content using feature vectors. As discussed in Sec. [5] deep learning methods outperform traditional methods in content-based visual indexing tasks by concurrently learning both local and global features, and demonstrating scale, spatial, and color invariance. In contrast, traditional methods struggle with aggregating distinct features and maintaining correlation due to their requirement for separate preprocessing modules. Despite their computational and data demands, deep learning methods are increasingly viable for content-based visual indexing tasks due to the abundance of data, maintaining their advantage in performance over traditional methods. One limitation identified in [6–9] relates to the use of CNN features for image and video clip representation which usually results in large-sized feature vectors, subsequently escalating the computational cost of video matching. Instead using a textual description of the content of the video or the images is more efficient in scenarios where law enforcement agents usually look for a video or image using natural language, besides computational cost effectiveness [10].

Several researchers have attempted to address this by using video captioning for visual indexing [2, 11–16]. Video captioning, which enables the textual interpretation of visual content, can serve as a potent tool for the query and retrieval of visual data. Despite its significant benefits, its application to forensic video indexation is not without constraints. Notwithstanding their remarkable efficiency, numerous video captioning methods insist on processing every single frame of the video or uniformly sampling frames for visual encoding [17]. This process not only imposes a substantial computational burden, but it may also result in the indexation of extraneous or repetitious information. Various researchers have approached the issue from different angles, utilizing machine

learning techniques, semantic understanding, and visual attention models [18]. Despite the progress in this field, the mismatch between these techniques and forensic video indexation persists. These methods often do not cater to the specific requirements of forensic analysis, such as effective temporal segmentation and efficient representation of key details. Video summarization techniques can be utilized to reduce the computational requirements by those visual captioning methods in addition to dealing with indexation of redundant contents in videos [19].

The critical role of video content summarization in streamlining the indexing of extensive video databases is well-documented. Early approaches to video summarization primarily relied on random or uniform sampling of video frames to extract key frames. For these key frame extraction methodologies to be efficacious, it's imperative that the chosen key frames capture a comprehensive overview of the video content, inclusive of pivotal information such as individuals or objects, while evading redundancy [20–22]. Techniques such as frame difference, while simple to implement, suffer from key limitations. They are prone to false scene change signals, which may be triggered by video editing effects, rapid camera movements, and lack the capacity to group similar segments or scenes at different parts of the video [23]. Recently, there has been a surge in the development of key frame extraction algorithms, broadly classified into two main categories: segmentation-based and clustering-based techniques.

Segmentation-based algorithms primarily focus on the detection of substantial changes between successive frames. These methodologies, however, may face the issue of extracting repetitive key frames if recurring content is present in the video [24–26]. Where as clustering-based techniques involve grouping frames and selecting those nearest to the cluster centers as key frames. A variety of clustering strategies have been proposed, including unsupervised clustering [27–29], density based clustering [30], k-means clustering [31,32], and spectral clustering on spatio-temporal features [33, 34].

3. PROPOSED APPROACH

In this section, we present our proposed approach for efficient video indexation that combines temporal segmentation with keyframe selection and specialized image captioning techniques. The overview of our approach is shown in Figure 1. Our proposed methodology comprises two primary modules. The first module is a temporal segmentation module that is intricately linked with a keyframe selection sub-module. The temporal segmentation module divides a video into smaller segments based on the changes of scenes. The keyframe selection sub-module then selects representative frames that summarize the visual scenes in the video. The objective of the segmentation module is to partition a given video into smaller segments based on the underlying content. The keyframe selection sub-module aims to identify the most informative frames that can represent the content of each segment. The proposed approach is expected to provide a more precise and comprehensive summary of the video content; The second module is a specialized image captioning module that generates descriptive captions for each segment using the summary frames selected.

3.1 Temporal Segmentation and Key Frame Selection Phase

The initial stage of our suggested methodology encapsulates the processes of temporal segmentation and key frame selection. Contrary to standard methodologies that carry out these tasks in a sequential manner, our proposition introduces an amalgamated approach that concurrently manages temporal segmentation and key frame selection. This is accomplished by executing a density-based clustering algorithm on frame-level features.

In place of conventional methodologies that represent video frames utilizing rudimentary attributes such as color features (RGB or HSV histogram), HoG, and other traditional feature extraction techniques [30], we propose the use of deep learning models. These models have been demonstrated as superior encoders, converting images into feature vectors. Our methodology capitalizes on unsupervised learning to train a deep learning model that is proficient in extracting features capable of representing both the background and foreground of an image. This stands in contrast to models trained solely on object classification that predominantly focus on the image foreground.

Post-VGGNet [35], the complexity of CNN architectures has escalated to enhance model performance, resulting in heavier, more computationally demanding models. This complexity can be problematic for scenarios

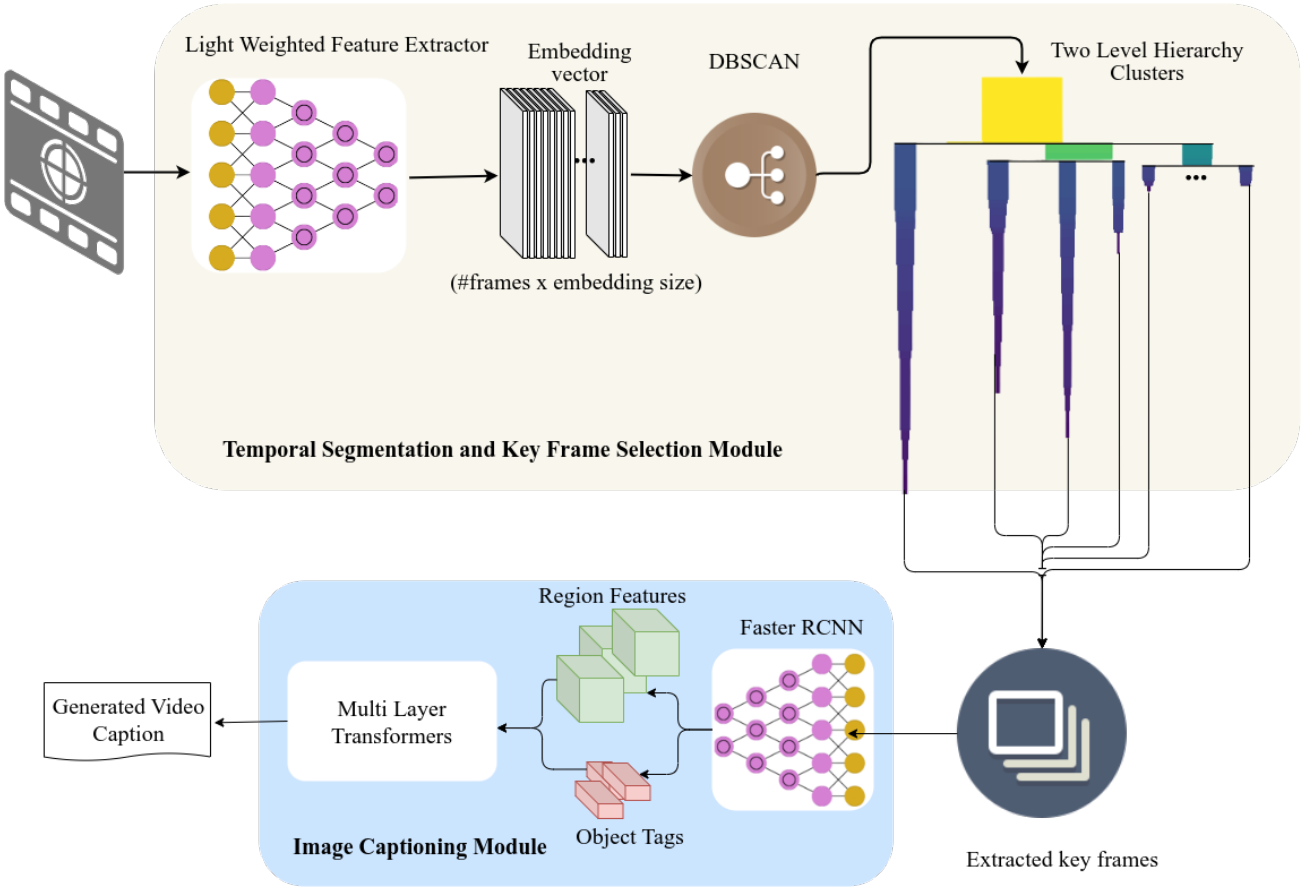

Figure 1. Proposed video indexation framework: (1)Temporal segmentation with keyframe selection which generates time stamp for each segments and representative keyframes of the segments; (2) Image caption generation module which generate grounded caption for every keyframe extracted from the previous phase.

where computational efficiency is paramount. However, the architecture proposed in [36] ingeniously implements a method to train and convert ResNet [37] into a VGG-style CNN for inference time. This makes it an optimal choice for utilization as a feature extractor in our proposed video indexation framework shown in Figure 1. It fulfills the dual criteria of being both computationally efficient and providing rapid inference time.

The extraction of key frames from videos through the application of clustering algorithms is a prevalent method [39]. Here, frames demonstrating significant similarity are grouped together, with the center of the cluster serving as a key frame for the video. The principal clustering techniques encompass partition-based, hierarchical, and density-based algorithms. Partition-based strategies such as K-means minimize the distance between data points and selected cluster centers. However, these require a predefined number of clusters, which presents a challenge for video temporal scene segmentation, as the quantity of distinct scenes in a video is typically unknown.

Hierarchical clustering, as exemplified by Agglomerative Nesting, crafts a hierarchical structure of clusters without the need for predefined clusters. Nevertheless, its application is less effective for abstract and complex data, such as videos, due to its splitting method. In contrast, density-based clustering defines clusters as areas of higher density relative to the remaining dataset. Algorithms like DBSCAN efficiently identify clusters with varying densities and can detect clusters of arbitrary shapes without the need to predefine cluster numbers. This capability makes these algorithms particularly suited for video segmenting and key frame selection, considering the variable number of different scenes in a video and the simplicity in selecting representative frames.

It's worth noting that while a video comprises a sequence of frames, these visual elements do not conform to

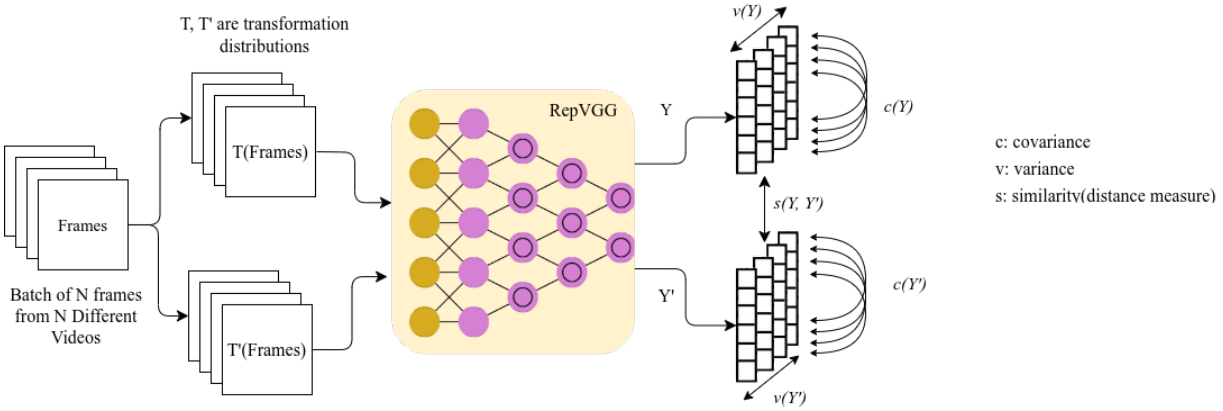

Figure 2. Light-weighted feature extractor trained using VICReg [38]. Input: N frames are selected from N videos and then transformation is applied to create two batches of frames as input. Output: two embedding vectors for the two batches feed to the RepVGG encoder network. Finally self-supervised loss will be applied which uses variance, invariance and covariance terms to prevent the encoder from collapsing problem while learning to represent images into embedding vectors(feature vecotors).

a linear arrangement. Instead, they are structured hierarchically within semantic space, as illustrated in [40,41]. Therefore, the use of a hierarchical density-based clustering algorithm is more appropriate, given the hierarchical semantic nature of videos and the efficiency of density-based clustering for segmenting videos into different scenes and selecting representative frames for each scene cluster.

Our methodology uses DBSCAN to generate clusters, demarcating scene or shot boundaries within the video. Key frames are then selected from the densest region of each cluster, ensuring these frames offer a comprehensive representation of the main content within each segment. This amalgamation of temporal segmentation and key frame selection not only streamlines the video analysis process but also generates a selective summary of the video. This summarization reduces computational demands and redundancy while enhancing the relevance of the information extracted for the subsequent stage of image captioning.

3.2 Visual Caption Generation Phase

The culmination of this module is the creation of descriptive captions for both videos and images. This stage makes use of an image captioning module based on in the architecture from [42]. This specific module is employed as opposed to a more complex, full-scale video captioning model for several reasons. First, by operating on key frames, the model significantly reduces the computational complexity inherent to processing entire videos, aligning with our broader aim of computational efficiency. Secondly, utilizing an image captioning model over a video captioning model allows our indexation framework to seamlessly index both video and image data without the need for separate models. This dual-purpose functionality streamlines the system's usability across different data types.

Our image captioning module has been further refined to the specific needs of counter-terrorism and cybercrime investigations, via fine-tuning on a proprietary dataset unique to these scenarios. The fundamental operation of our image captioning module involves two main processes. The first is the encoding of the images or selected key frames into an information-rich feature vector employing Faster R-CNN [43]. Notably, in addition to generating region features, Faster R-CNN also detects object tags within the images, serving as pivotal anchor points. These anchor points significantly ease the process of learning alignments, creating a robust link between the semantics of the input image and the generated caption. The second process employs a multi-layer Transformer, equipped with self-attention mechanisms. Using the region features and object tags derived from the Faster R-CNN, the Transformer decoder constructs a contextually enriched caption that is firmly grounded in the content of the input image.

These generated captions play a dual role. Not only do they provide a concise, text-based synopsis of the video content, but they also serve a crucial function in facilitating the search and retrieval of specific video segments

via natural language queries. This offers investigators an efficient and effective method for parsing through extensive visual data. In essence, our methodology delivers a streamlined, comprehensive, and contextually enriched text representation of image and video content that satisfies the specific demands of forensic visual information indexation, while maintaining computational efficiency.

4. EXPERIMENTAL RESULTS

4.1 Datasets

In our experimental framework, we employed the SumMe dataset [44], designed explicitly as a benchmark for video summarization tasks. This dataset encompasses 25 videos of varied lengths, from a minimum of one minute to a maximum of six minutes. Each of these videos is annotated by a minimum of 15 human annotators, yielding a total of 390 human summaries. These annotations were gathered through a crowd sourcing approach. All human-constructed summaries were mandated to maintain a length within 15% of the original video duration. The video have have three different settings namely: Egocentric (4 videos), moving camera (17 videos) and static camera (4 videos). Previous works on video summarization that reports on SumMe dataset use F1 score as their metric, thus we will also use the same metric to report the results from the temporal segmentation and keyframe selection module.

4.2 Temporal segmentation and KeyFrame Selection Module Results

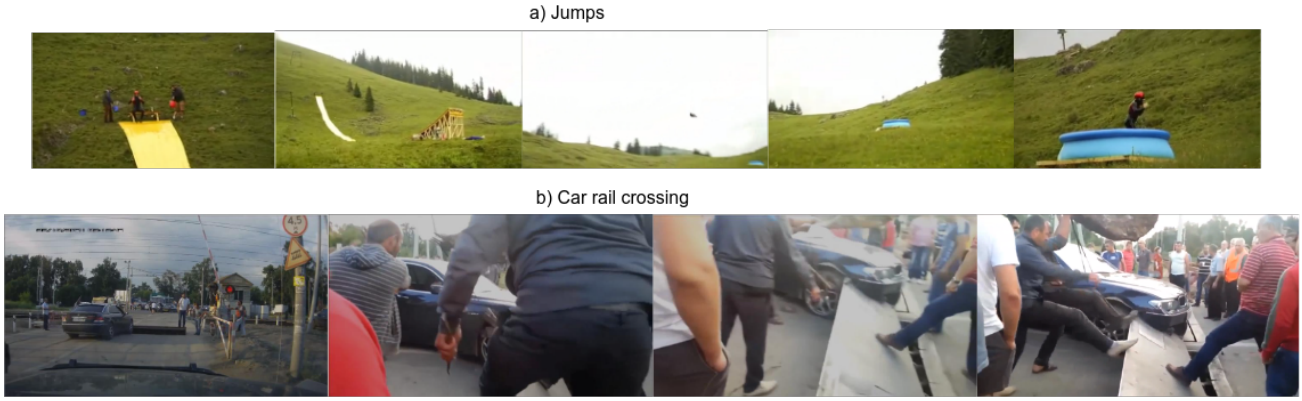

Figure 3. Two example videos from the SumMe dataset. Both a) Jumps; and b) Car rail crossing video have moving camera with dynamic background

We ran the proposed methods on the SumMe Dataset, and compared the F1-scores obtained by them (as shown in Table 1). Our main goal is to be as much close to average human summary, which we were able to obtain using the backbone network trained using self-supervised techniques discussed in Sec. 3.1. To compute the F1 score, we select frames from the cluster with high density region in the generated clusters. Those highly dense frames are assumed to be similar and representative to the cluster. Then using the selected high density frames from each cluster, evaluation script from the main work of SumMe dataset [44] is used to score the summary. Representative key frames from two illustrative videos, namely 'Jumps' and 'Car Rail Crossing', are exhibited in Figure 3.

Our video summary have a better F1 score compared to previous similar method in [45] [CNN (Gaussian)] and uniform sampling scores beside the consistency of the generated summary accross different settings. Even though our method result a lower F1 score compared to the average human score, the average human has a wider distribution as indicated by a box plot on the left of Figure 4 , indicating a larger variability in scores for different videos. As it is shown on the right side of Figure 4, our method generate a summary which is more consistent compared with human average, uniform sampling and CNN (Gaussian) [45] for videos in different settings of SumMe dataset described in Sec. 4.1.

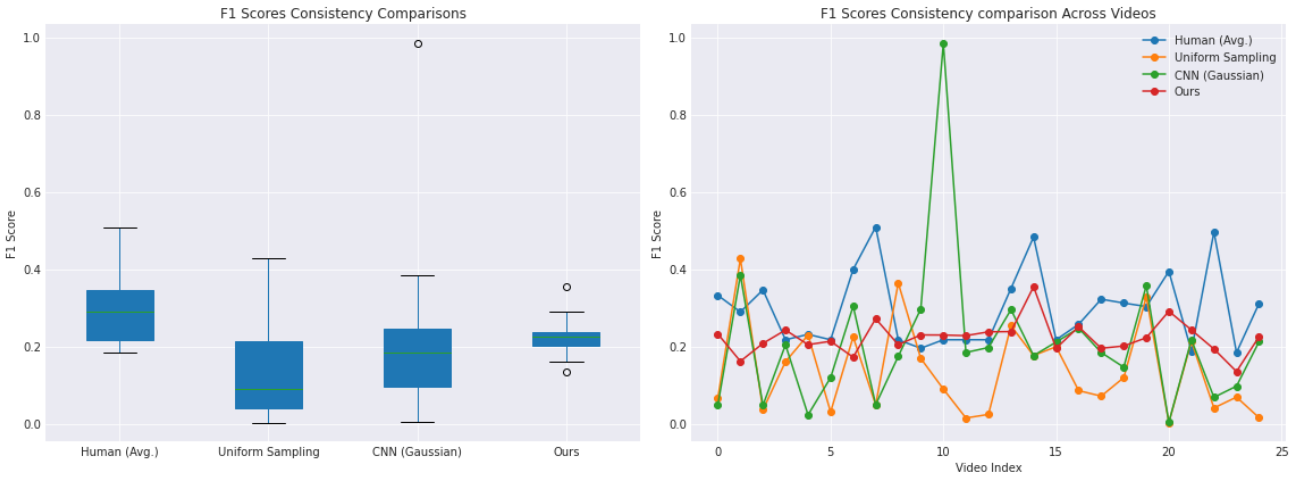

Figure 4. F1 score consistency comparison: We compare our method (red) against average human (blue), uniform sampling (orange) and CNN (Gaussian) (green) [45]

4.3 Video Captioning for Indexation

For Illustrating the image captioning module we used one of the video from SumMe dataset which have car crashing to demonstrate the generated caption for key frames in crime related video scenes. Once the temporal segmentation and keyframe selection module generate the summary keyframes, the video captioning module will generate a caption for each keyframes which crate a textual summary of the video as shown in Figure 5. Unlike the the video summary evaluation, we select only one representative frame per cluster to be captioned using the captioning module. The frame in the center of the cluster is assumed to be the most representative frame of the cluster.

Figure 5. Sample videos selected from SumMe dataset of 'Car Railcrossing' video and generated output

5. CONCLUSION

In conclusion, this paper presented an efficient and novel framework for video indexation that is particularly suitable for forensic analysis. Our approach is unique in its integrated methodology, merging temporal segmentation and key frame selection into a unified process. The use of a density-based clustering algorithm and feature representation based on the self-supervised representation learning allows for effective video summarization, in contrast to traditional frame-by-frame analysis. Subsequently, a fine-tuned image captioning module, built upon the OSCAR model, is employed to generate descriptive captions for the selected key frames. This not only facilitates efficient video summarization but also enhances the ease of video retrieval using natural language queries.

The experimental results, derived from the SumMe dataset, demonstrate the robust performance of our framework, providing high-quality video summaries that align well with human-generated summaries, considering that our pipeline haven't seen the sumMe dataset at all in the training or finetuning stages in our pipeline. The

Table 1. **Quantitative results**: We show f-measures at 15% summary length for our approach, the baselines and the human selections. We highlight the best and second best computational method. Our method consistently shows a high performance scoring higher than the worst human per video.

Setting	Video Name	Human (Avg.)	Uniform Sampling	CNN (Gaussian)	**Ours**
Ego.	Base jumping	0.257	0.085364	0.247	0.250596
	Scuba	0.217	0.0145059	0.184	0.227593
	Bike Polo	0.322	0.07112	0.184	0.19496
	Valparaiso_Downhill	0.217	0.19899	0.211	0.194968
Moving	Kids_playing_in_leaves	0.289	0.426775	0.384	0.161311
	Bearpark_climbing	0.217	0.160377	0.204	0.242144
	Notre_Dame	0.231	0.229265	0.0227	0.204327
	Bus_in_Rock_Tunnel	0.217	0.030199	0.119	0.213109
	paluma_jump	0.509	0.048565	0.049	0.271479
	Playing_on_water_slide	0.195	0.168675	0.297	0.229481
	Cockpit_Landing	0.217	0.089413	0.984	0.229277
	Car_railcrossing	0.217	0.363804	0.174	0.204315
	Cooking	0.217	0.023748	0.197	0.237521
	Uncut_Evening_Flight	0.35	0.253156	0.295	0.238098
	Jumps	0.483	0.176244	0.176	0.354185
	Eiffel Tower	0.312	0.119034	0.146	0.200886
	Excavators river crossing	0.303	0.328008	0.357	0.222269
	Saving dolphines	0.188	0.212642	0.217	0.242406
	St Maarten Landing	0.496	0.0404343	0.068	0.193098
	Statue of Liberty	0.184	0.068651	0.097	0.13479
Static	Fire Domino	0.394	0.002603	0.0035	0.290872
	Air_Force_One	0.332	0.066812	0.048	0.231993
	car_over_camera	0.346	0.035693	0.0475	0.207619
	Paintball	0.399	0.224322	0.304	0.171988
	Mean	0.311	0.0152	0.212	0.225

proposed method introduces a promising direction for the field of forensic video analysis and opens avenues for future research.

6. ACKNOWLEDGEMENTS

The work described in this paper is performed in the H2020 project STARLIGHT ("sustainable Autonomy and Resilience for LEAs using AI against High Priority Threats"). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101021797.

REFERENCES

[1] Solio, M. A. A. and Ladhake, S. A., "A review of query image in content based image retrieval," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* **2**(4), 1619–1622 (2013).

[2] Yi, X. G., "Automatic caption extraction of news video and its implementation," in [*2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE)*], **1**, 122–125 (2012).

[3] Ansari, A. and Mohammed, M. H., "Content based video retrieval systems-methods, techniques, trends and challenges," *International Journal of Computer Applications* **112**(7) (2015).

[4] Spolaôr, N., Lee, H. D., Takaki, W. S. R., Ensina, L. A., Coy, C. S. R., and Wu, F. C., "A systematic review on content-based video retrieval," *Engineering Applications of Artificial Intelligence* **90**, 103557 (2020).

[5] Ravi, A. and Nandakumar, A., "A multimodal deep learning framework for scalable content based visual media retrieval," *arXiv preprint arXiv:2105.08665* (2021).

[6] Zhang, C., Lin, Y., Zhu, L., Liu, A., Zhang, Z., and Huang, F., "Cnn-vwii: An efficient approach for large-scale video retrieval by image queries," *Pattern Recognition Letters* **123**, 82–88 (2019).

[7] Dubey, S. R., "A decade survey of content based image retrieval using deep learning," *IEEE Transactions on Circuits and Systems for Video Technology* **32**(5), 2687–2704 (2021).

[8] Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., and Kompatsiaris, Y., "Near-duplicate video retrieval by aggregating intermediate cnn layers," in [*MultiMedia Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part I 23*], 251–263, Springer (2017).

[9] Rian, Z., Christanti, V., and Hendryli, J., "Content-based image retrieval using convolutional neural networks," in [*2019 IEEE International Conference on Signals and Systems (ICSigSys)*], 1–7, IEEE (2019).

[10] Wang, Q., Zhang, Y., Zheng, Y., Pan, P., and Hua, X.-S., "Disentangled representation learning for text-video retrieval," *arXiv preprint arXiv:2203.07111* (2022).

[11] Bor-Chun Chen, Y.-Y. C. and Chen, F., "Video to text summary: Joint video summarization and captioning with recurrent neural networks," in [*Proceedings of the British Machine Vision Conference (BMVC)*], Tae-Kyun Kim, Stefanos Zafeiriou, G. B. and Mikolajczyk, K., eds., 118.1–118.14, BMVA Press (September 2017).

[12] Cho, J., Jeong, S., and Choi, B., "News video retrieval using automatic indexing of korean closed-caption," in [*Knowledge-Based Intelligent Information and Engineering Systems*], Khosla, R., Howlett, R. J., and Jain, L. C., eds., 694–703, Springer Berlin Heidelberg, Berlin, Heidelberg (2005).

[13] Su, J.-H., Huang, Y.-T., and Tseng, V. S., "Efficient content-based video retrieval by mining temporal patterns," in [*Proceedings of the 9th International Workshop on Multimedia Data Mining: Held in Conjunction with the ACM SIGKDD 2008*], *MDM '08*, 36–42, Association for Computing Machinery, New York, NY, USA (2008).

[14] Kulkarni, P., Patil, B., and Joglekar, B., "An effective content based video analysis and retrieval using pattern indexing techniques," in [*2015 International Conference on Industrial Instrumentation and Control (ICIC)*], 87–92 (2015).

[15] Tang, Z., Lei, J., and Bansal, M., "DeCEMBERT: Learning from noisy instructional videos via dense captions and entropy minimization," in [*Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*], 2415–2426, Association for Computational Linguistics, Online (June 2021).

[16] Zhang, C. and Tian, Y., "Automatic video description generation via lstm with joint two-stream encoding," *2016 23rd International Conference on Pattern Recognition (ICPR)* , 2924–2929 (2016).

[17] Zhou, L., Zhou, Y., Corso, J. J., Socher, R., and Xiong, C., "End-to-end dense video captioning with masked transformer," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 8739–8748 (2018).

[18] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., and Carlos Niebles, J., "Dense-captioning events in videos," in [*Proceedings of the IEEE international conference on computer vision*], 706–715 (2017).

[19] Chen, B.-C., Chen, Y.-Y., and Chen, F., "Video to text summary: Joint video summarization and captioning with recurrent neural networks," in [*British Machine Vision Conference*], (2017).

[20] Abdulhussain, S. H., Ramli, A. R., Saripan, M. I., Mahmmod, B. M., Al-Haddad, S. A. R., and Jassim, W. A., "Methods and challenges in shot boundary detection: a review," *Entropy* **20**(4), 214 (2018).

[21] Xue, L., Li, C., Li, H., and Xiong, Z., "A general method for shot boundary detection," in [*Proceedings of the International Conference on Multimedia and Ubiquitous Engineering*], 394–397 (2008).

[22] Yuan, J. et al., "A formal study of shot boundary detection," *IEEE Transactions on Circuits and Systems for Video Technology* **17**(2), 168–186 (2007).

[23] Gygli, M., Grabner, H., Riemenschneider, H., and Gool, L. V., "Creating summaries from user videos," in [*European Conference on Computer Vision*], (2014).

[24] Guan, G., Wang, Z., Lu, S., Deng, J. D., and Feng, D. D., "Keypoint-based keyframe selection," *IEEE Transactions on Circuits and Systems for Video Technology* **23**(4), 684–694 (2013).

[25] Ejaz, N., Tariq, T. B., and Baik, S. W., "Adaptive key frame extraction for video summarization using an aggregation mechanism," *Journal of Visual Communication and Image Representation* **23**(7), 1031–1040 (2012).

[26] Widiarto, W., Yuniarno, E. M., and Hariadi, M., "Video summarization using a key frame selection based on shot segmentation," in [*2015 International Conference on Science in Information Technology (ICSITech)*], 207–212 (2015).

[27] Zhuang, Y., Rui, Y., Huang, T., and Mehrotra, S., "Adaptive key frame extraction using unsupervised clustering," in [*Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*], **1**, 866–870 vol.1 (1998).

[28] Tang, H., Ding, L., Wu, S., Ren, B., Sebe, N., and Rota, P., "Deep unsupervised key frame extraction for efficient video classification," *ACM Transactions on Multimedia Computing, Communications and Applications* **19**(3), 1–17 (2023).

[29] Zhuang, Y., Rui, Y., Huang, T., and Mehrotra, S., "Adaptive key frame extraction using unsupervised clustering," in [*IEEE International Conference on Image Processing*], **1**, 866–870, IEEE Comp Soc (1998). Proceedings of the 1998 International Conference on Image Processing, ICIP. Part 2 (of 3) ; Conference date: 04-10-1998 Through 07-10-1998.

[30] Zhang, X., Li, Y., Li, X., Zhang, X., and Li, Z., "A clustering algorithm for key frame extraction based on density peak," *Journal of Computer and Communications* **6**(12), 1–9 (2018).

[31] Muhammad, B., Sadiq, B., Umoh, I., and Bello-Salau, H., "A k-means clustering approach for extraction of keyframes in fast-moving videos," *International Journal of Information Processing and Communication (IJIPC)* **9**(1&2), 147–157 (2020).

[32] Avila, S. E. F. D., Lopes, A. P. B., Luz, A. d., and Araújo, A. d. A., "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," *Elsevier PRL* **32**(1), 56–68 (2011).

[33] Vázquez-Martín, R. and Bandera, A., "Spatio-temporal feature-based keyframe detection from video shots using spectral clustering," *Pattern Recognition Letters* **34**(7), 770–779 (2013).

[34] Ioannidis, A. I., Chasanis, V. T., and Likas, A. C., "Key-frame extraction using weighted multi-view convex mixture models and spectral clustering," in [*2014 22nd International Conference on Pattern Recognition*], 3463–3468 (2014).

[35] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).

[36] Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., and Sun, J., "Repvgg: Making vgg-style convnets great again," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 13733–13742 (2021).

[37] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[38] Bardes, A., Ponce, J., and LeCun, Y., "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906* (2021).

[39] John, A. A., Nair, B. B., and Kumar, P. N., "Application of clustering techniques for video summarization – an empirical study," in [*Artificial Intelligence Trends in Intelligent Systems*], Silhavy, R., Senkerik, R., Kominkova Oplatkova, Z., Prokopova, Z., and Silhavy, P., eds., 494–506, Springer International Publishing, Cham (2017).

[40] Sarfraz, S., Murray, N., Sharma, V., Diba, A., Van Gool, L., and Stiefelhagen, R., "Temporally-weighted hierarchical clustering for unsupervised action segmentation," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 11225–11234 (2021).

[41] Xiao, J., Yao, A., Liu, Z., Li, Y., Ji, W., and Chua, T.-S., "Video as conditional graph hierarchy for multi-granular question answering," in [*Proceedings of the AAAI Conference on Artificial Intelligence*], **36**(3), 2804–2812 (2022).

[42] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al., "Oscar: Object-semantics aligned pre-training for vision-language tasks," in [*Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*], 121–137, Springer (2020).

[43] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J., "Vinvl: Revisiting visual representations in vision-language models," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 5579–5588 (2021).

[44] Gygli, M., Grabner, H., Riemenschneider, H., and Van Gool, L., "Creating summaries from user videos," in [*Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*], 505–520, Springer (2014).

[45] Jadon, S. and Jasim, M., "Unsupervised video summarization framework using keyframe extraction and video skimming," in [*2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*], IEEE (oct 2020).