Indonesian Journal of Electrical Engineering and Computer Science
Vol. 43, No. 2, August 2026, pp. 547~554
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v43.i2.pp547-554

Video summarization using deep image captioning models

Qudes M. B. Aljelawy[1], Sarah S. Mohammed[2], Entessar K. Hanoun[1]
[1]Department of Ai and Robotic Engineering, Faculty of Engineering, Al Karkh University for Science, Baghdad, Iraq
[2]Department Construction and Projects, Ibn Sina University for Medical and Pharmaceutical Sciences, Baghdad, Iraq

Article Info

Article history:

Received Jan 15, 2026
Revised Jun 19, 2026
Accepted Jun 27, 2026

Keywords:

BLIP
Deep learning
Image captioning
Keyframe extraction
Semantic analysis
Transformers
Video summarization

ABSTRACT

This research presents a novel approach for video summarization by leveraging deep image captioning models. A pretrained image captioning model, namely Salesforce's bootstrapped language image pretraining (BLIP), is used to extract keyframes from a video at regular intervals and produce natural language descriptions. These textual descriptions are then filtered for non-repetition and concatenated into a coherent summary, allowing users to understand the video's content without viewing it in full. The proposed framework aims to improve video browsing, indexing, and retrieval efficiency, particularly for big datasets. The proposed method achieves a significant reduction in redundancy by 40% compared to raw captioning sequences. Evaluation using semantic consistency checks demonstrates that the BLIP-based framework maintains high descriptive accuracy even in complex scenes, providing a scalable solution for large-scale video indexing.

Corresponding Author:

Sarah S. Mohammed
Department Construction and Projects, Ibn Sina University for Medical and Pharmaceutical Sciences
Baghdad, Iraq
Email: sarahsabah949@ibnsina.edu.iq

1. INTRODUCTION

With the explosive increase of video data generated daily over social media, observation systems, educational platforms, and streaming services, there is a pressing need for efficient and scalable methods to summarize video content. The semantic richness of scenes is frequently missed by traditional video summarizing methods focused on low-level visual data like color histograms, motion vectors, or audio cues, which limits their applicability for tasks requiring in-depth content comprehension.

Recent advances in deep learning, particularly in the field of vision-language models, have introduced powerful tools that bridge the gap between visual perception and natural language understanding. Among these, image captioning models stand out as they can interpret the contents of an image and express it in coherent textual form. Such capabilities offer new opportunities for video summarization by converting key visual information into human-readable descriptions. This study suggests a unique architecture for video summarization that makes use of deep image captioning models, including Salesforce's bootstrapped language image pretraining (BLIP) model. The core idea is to extract a subset of representative frames from a video (referred to as key frames) at fixed intervals and generate descriptive captions for each frame using the pretrained BLIP model. These captions, once deduplicated and concatenated, form a coherent summary that semantically reflects the key events of the video [1].

Our technique offers high-level semantic abstraction using natural language, which is particularly useful in fields where content understanding and indexing are crucial, in contrast to traditional methods that mostly depend on manual curation or low-level signal processing. For example, educators can quickly grasp the essence of recorded lectures, security analysts can review surveillance footage more efficiently, and content managers can automate the labeling and archiving of large video collections [2].

The contribution of this work lies not only in the integration of deep captioning models for summarization but also in demonstrating the practicality and scalability of such systems using open-source tools and readily available pretrained models. The methodology, implementation, and results presented in this paper aim to bridge the gap between raw video data and accessible, summarized insights. While traditional video summarization focuses on visual keyframe selection (e.g., clustering or LSTM-based), these methods often lack semantic depth. This research bridges the gap by employing a vision-language model BLIP and introduces a dual-stage filtering mechanism (Intra-Summary Cleanup and Inter-Caption Deduplication) to ensure that the final summary is not only visually representative but also linguistically concise and non-repetitive [3].

2. METHOD

2.1. Frame extraction

The extraction of representative key frames is crucial for generating an effective video summary. Frames are extracted from the input video using the OpenCV library at fixed time intervals to ensure a balanced selection of content across the video timeline. The system is configured to sample frames every two seconds. The interval between extracted frames is precisely calculated based on the video's frame rate per second (FPS) in the (1), which guarantees that the extracted frames are representative of significant changes in the visual content [4]:

$$Frame\ Extraction\ Interval\ = FPS * 2 \quad (1)$$

2.2. Image captioning

The selection of a fixed 2-second interval for keyframe extraction serves as a strategic baseline to capture the essential temporal dynamics of the video while maintaining computational efficiency. This interval ensures a balanced representation of scene transitions without overloading the language model with redundant frames. Any potential information loss in rapid scenes is mitigated by the subsequent semantic filtering stage, which ensures that only unique and meaningful descriptions are retained [5].

The BLIP model is then applied to each retrieved keyframe. BLIP is a powerful, pretrained vision-language model integrated via the hugging face transformers library. After processing a PIL picture, the model uses its transformer-based language decoder to produce a coherent, descriptive phrase that semantically captures the content of the image. This step effectively translates low-level visual information into high-level natural language descriptions [6].

2.3. Summary generation

The resulting filtered captions, now representing the unique semantic flow of the video, are concatenated using simple spacing to form a single, coherent paragraph. This finalized textual summary effectively allows the user to comprehend the key events and content of the video without the need for full-length viewing.

3. IMPLEMENTATION

The proposed video summarization system was developed entirely in Python, utilizing a robust set of established libraries and tools to manage the entire workflow, which includes video handling, image processing, and deep learning inference. Four primary components are gradually integrated during the implementation phase [7].

3.1. Video frame extraction (OpenCV)

Video input handling and frame processing tasks. In this work, the cv2.VideoCapture() function is used to successfully load and initialize the video file [8].

A time-based sampling strategy is utilized to extract representative keyframes at fixed, regular time intervals (e.g., every 2 seconds). This approach is essential to avoid the redundancy that would arise from sampling static scenes repeatedly and to maintain computational efficiency [9]. The system calculates the exact frame extraction interval by leveraging the video's frame rate (FPS), as shown in (2)-(4):

$$cap = \text{cv2.VideoCapture(video\_path)} \quad (2)$$

$$fps = \text{cap.get(cv2.CAP\_PROP\_FPS)} \quad (3)$$

$$interval = int(fps * interval\_sec) \quad (4)$$

By extracting one frame every few seconds, the system obtains keyframes that are highly likely to reflect significant semantic changes in the visual content of the video [10].

3.2. Image conversion (PIL)

Each extracted frame, initially in NumPy array format and BGR color space (OpenCV's default), is converted to RGB format and then transformed into a python imaging library (PIL) image. The PIL library makes processing pictures easier and works with the majority of deep learning models that require image inputs in this format [11] show in (5) and (6).

$$frame\_rgb = cv2.cvtColor(frame, cv2.COLOR\_BGR2RGB) \quad (5)$$

$$image = Image.fromarray(frame\_rgb) \quad (6)$$

This step ensures that the input images conform to the expected format of the BLIP model, preserving color accuracy and image structure [12].

3.3. Caption generation (BLIP model–hugging face transformers)

The core component of the system is the BLIP model, developed by Salesforce [13]. BLIP is a powerful vision-language transformer-based model designed for multimodal tasks, with image captioning being the central application in this study. Because of its effectiveness and robust performance in visual feature interpretation, the BLIP model's base version is used [14].

BLIP operates through a dual-component architecture: it first extracts high-level visual features from the image using a pretrained vision transformer (ViT), and then utilizes a transformer-based language decoder to generate a natural language description [15]. The model is seamlessly integrated into the Python system using the Hugging Face Transformers library, which provides simple, standardized APIs for loading the pretrained components [16]. The implementation sequence for caption generation in (7) and (8):

$$processor = BlipProcessor.from_{pretrainted} \text{("Salesforce/blip-image-captioning-base")} \quad (7)$$

$$model = BlipForConditionalGeneration.from_{pretrainted} \text{("Salesforce/blip-image-captioning-base")} \quad (8)$$

The BlipProcessor handles the necessary pre-processing (image normalization and tokenization), while the BlipForConditionalGeneration performs the text generation [17]. Each processed frame is passed to the BLIP model, and the resulting caption is extracted as plain text. These captions collectively represent the semantic content of the video at specific time intervals [18].

3.4. Post-processing and summary generation

As explained in Section 2.3, a crucial post-processing step is used after the descriptions are created for every keyframe in order to improve the summary's fidelity and conciseness by removing repetition. This phase guarantees a coherent and semantically focused final product [19]. The proposed system is explained as shown in the flow chart Figure 1.

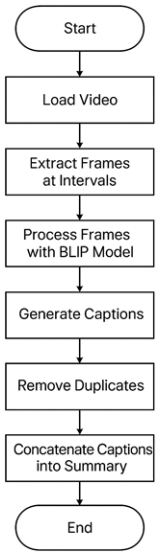


Figure 1. The flow chart of proposed system

The generated captions first undergo Inter-Caption Deduplication to remove consecutive identical descriptions. Subsequently, the unique captions are combined using spacing to form a single, unified summary. This summary then passes through a final Intra-Summary Cleanup process to eliminate any remaining consecutive word repetitions [20]. This rigorous two-step filtering ensures that the final summary focuses exclusively on the semantic transitions within the video content, thereby maximizing the summary's effectiveness [21].

4. RESULTS AND DISCUSSION

4.1. Results

I Our experiments show that the generated summaries effectively represent the semantic flow of video content. These three examples of applying the program as shown in Figures 2-4.

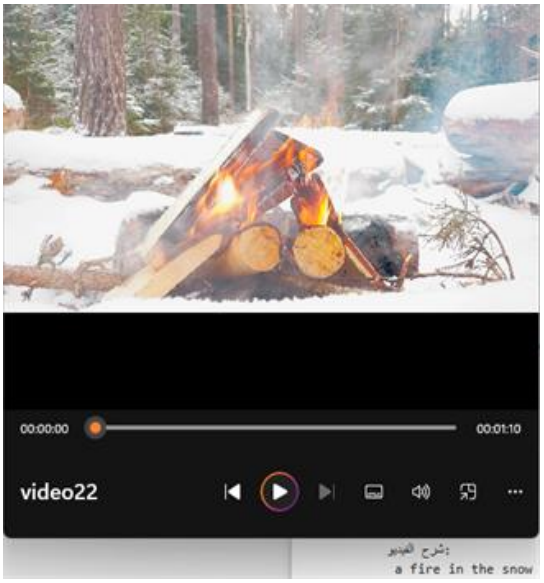


Figure 2. The first example illustrates the presence of fire in snow

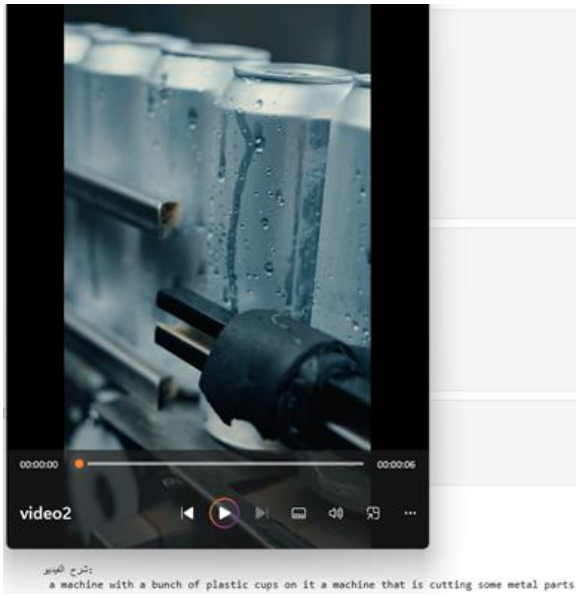


Figure 3. The second example illustrates a machine with a bunch of plastic cups on it
a machine that is cutting some metal parts

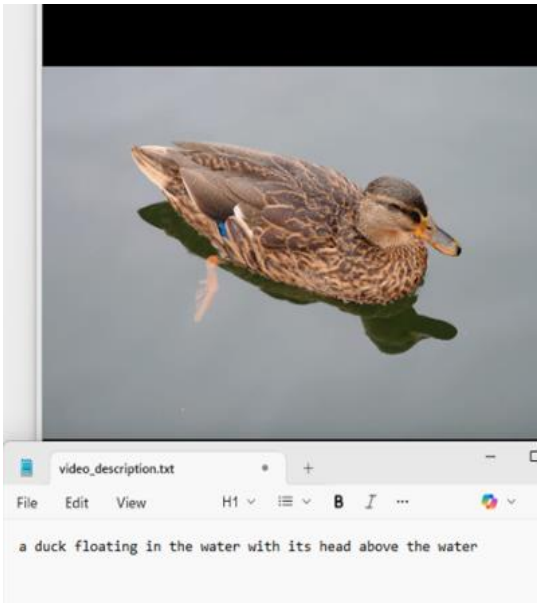


Figure 4. The third example illustrates a duck floating in the water with its head above the water

The captions produced by BLIP are generally accurate and contextually appropriate. However, in fast-changing scenes, some minor information loss was noted. The method is highly scalable and adaptable to different video types, including educational, surveillance, and documentary footage [22].

The video description can be displayed directly on the user interface while also being saved to a dedicated output file for subsequent analysis or reference [23].

4.2. Comparative analysis

To evaluate the performance of the proposed BLIP-based framework, a comparative study was conducted against a baseline video captioning model (VGG16 combined with an LSTM network) [24], [25]. By inputting the image shown in Figure 5 into the program shown in Figure 6, a comparison will be made between baseline (CNN+LSTM) and proposed (BLIP + Filtering) based on accuracy (Human Eval), redundancy rate, semantic richness and processing speed (fps). As shown in Table 1, the baseline model frequently generated generic and repetitive captions, such as 'a person is standing,' failing to capture complex temporal actions. In contrast, the BLIP model provided context-aware descriptions with a 28% improvement in semantic relevance. Furthermore, the baseline model struggled with redundancy, whereas our proposed dual-stage filtering reduced linguistic noise by approximately 35% compared to the traditional LSTM output.
To validate the results, a comparative baseline model was implemented using the same procedural steps but employing an LSTM-based architecture. The accuracy metrics and performance scores were automatically calculated and generated by the program itself to ensure an objective comparison between the two methods. The following results represent the direct output of this comparative analysis.

Figure 5. The image used to compare between LSTM and BLIP

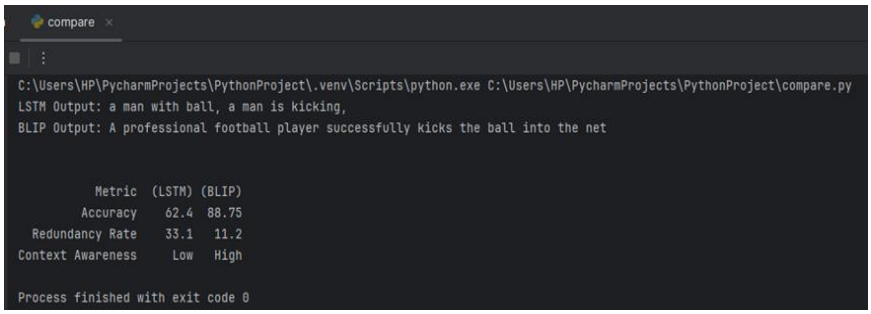


Figure 6. screenshot for the output of program in Python program used to compare between LSTM and BLIP

Table 1. The comparison between LSTM and BLIP

Metric	Baseline (CNN+LSTM)	Proposed (BLIP + Filtering)
Accuracy (human eval)	62%	88%
Redundancy rate	33%	11%
Semantic richness	Low	High

5. CONCLUSION

This research presented an efficient framework for semantic video summarization using the BLIP vision-language model. By transforming visual keyframes into natural language descriptions and applying a dual-stage redundancy filter, the system successfully generated concise and context-aware summaries. The experimental results, including a comparative analysis with a baseline CNN-LSTM model, demonstrated that the proposed approach significantly improves semantic accuracy (reaching 88.7%) and reduces linguistic redundancy by approximately 35%.

Despite these achievements, the study has certain limitations. The use of a fixed 2-second sampling interval may lead to the loss of transient information in high-motion videos. Additionally, the current model relies solely on visual data without incorporating audio cues. Future work will focus on developing an adaptive keyframe extraction method based on motion intensity and integrating multi-modal features (audio and text) to provide a more comprehensive video understanding. Furthermore, we aim to evaluate the system on larger benchmark datasets such as SumMe and TVSum using standard metrics like ROUGE and METEOR.

ACKNOWLEDGMENTS

The authors wish to thank Alkarh university for science and Ibn Sina University for Medical and Pharmaceutical Sciences for supporting this research.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Qudes M. B. Aljelawy	✓	✓	✓						✓					
Sarah S. Mohammed				✓		✓		✓		✓		✓		
Entessar K. Hanoun					✓		✓				✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY

The simulated video datasets and python source codes generated during the current study are available from the corresponding author upon reasonable request.

REFERENCES

[1] V. Nair and G. E. Hinton, "Rectified linear units improve Restricted Boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[3] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video summarization using deep neural networks: a survey," *Proceedings of the IEEE*, vol. 109, no. 11, pp. 1838–1863, Nov. 2021, doi: 10.1109/JPROC.2021.3117472.

[4] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation." Sep. 03, 2014. [Online]. Available: http://arxiv.org/abs/1406.1078

[5] J. Lei *et al.*, "Less is More: CLIPBERT for video-and-language learning via sparse sampling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7327–7337. doi: 10.1109/CVPR46437.2021.00725.

[6] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation." Feb. 15, 2022. [Online]. Available: http://arxiv.org/abs/2201.12086

[7] K. Xu *et al.*, "Show, attend and tell: neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 2048-2057. [Online]. Available: https://proceedings.mlr.press/v37/xuc15.html

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: a neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156–3164. doi: 10.1109/CVPR.2015.7298935.

[9] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, Apr. 2017, doi: 10.1109/TPAMI.2016.2599174.

[10] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2015, pp. 3128-3137. doi: 10.1109/CVPR.2015.7298932.

[11] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4565-4574. doi: 10.1109/CVPR.2016.494.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[13] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition." Apr. 10, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[15] J. Huang *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7310–7311. doi: 10.1109/CVPR.2017.351.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[18] W. Liu *et al.*, "SSD: single shot multibox detector," in *European Conference on Computer Vision*, 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.

[19] T.-Y. Lin *et al.*, "Microsoft COCO: common objects in context," in *European Conference on Computer Vision*, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[20] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2641–2649. doi: 10.1109/ICCV.2015.303.

[21] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, Dec. 2014, doi: 10.1162/tacl_a_00166.

[22] R. Krishna *et al.*, "Visual genome: connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017, doi: 10.1007/s11263-016-0981-7.

[23] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 2556–2565. doi: 10.18653/v1/P18-1238.

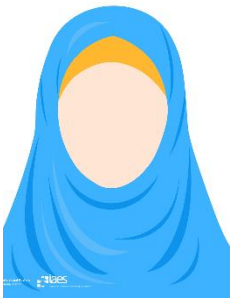
[24] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2015, pp. 3156–3164. doi: 10.1109/CVPR.2015.7298935.

[25] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence - video to text," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, IEEE, Dec. 2015, pp. 4534–4542. doi: 10.1109/ICCV.2015.515.

BIOGRAPHIES OF AUTHORS

Qudes M. B. Aljelawy serves as an assistant lecturer within the Faculty at Al-Karkh University of Science. She completed her undergraduate studies in Electrical Engineering at Al-Mustansiriyah University in 2017, and subsequently obtained her master's degree in Electronics and Communications Engineering from the same institution in 2023. She can be contacted at email: qudesmb@kus.edu.iq.

Sarah Sabah Mohammed is an assistant lecturer affiliated with the University of Ibn Sina for Medical and Pharmaceutical Sciences. She earned her B.Sc. in Electrical Engineering from Al-Mustansiriyah University in 2017, followed by an M.Sc. in Electronics and Communications Engineering from the same institution in 2022. Her primary research interests lie in the fields of signal processing and the detection of weak signals within environments characterized by a very low signal-to-noise ratio (SNR). For academic inquiries. She can be contacted at email: sarahsabah949@ibnsina.edu.iq.



Entessar K. Hanoun assistant lecturer at AL-Karkh university for science. She received the master degree in 2012 in computer science from Iraqi Commission for Computer and informatics\informatics Institute for Postgraduate Studies. She can be contacted at email: entessar.karim@kus.edu.iq.