Research Article GU J Sci, Part A, 13(1): 229-241 (2026) 10.54287/gujsa.1839173

Gazi University
Journal of Science
PART A: ENGINEERING AND INNOVATION
http://dergipark.org.tr/gujsa

# Transforming Frame-Based Video Descriptions into Text Using Large Language Models

Özge SOĞUKPINAR GÜL[1*] Ahmet SAYAR[2]

[1] Kocaeli University, Graduate School of Natural and Applied Sciences, Kocaeli, Türkiye
[2] Kocaeli University, Institute of Science, Kocaeli, Türkiye

**Keywords**

Video Processing
Llm Video Description
Video Analysis With LLM
Large Language Models
Turkish NLP

**Abstract**

This study proposes a novel framework that integrates frame-based video description techniques with transformer-based large language models (LLMs) to automatically generate coherent and contextually accurate narratives from video content. In the first stage, each frame of a video is processed using a vision-language model, specifically BLIP, to produce preliminary captions that describe the visual elements present in individual frames. In the second stage, these frame-level captions are sequentially provided to a fine-tuned LLM, FLAN-T5, which is capable of leveraging temporal coherence and contextual information across frames to generate a fluent and semantically consistent text. The proposed method is applied to Turkish video datasets, highlighting its contribution to low-resource language processing within the field of Natural Language Processing (NLP). Experimental results demonstrate that the system achieves a BLEU-4 score of 0.67 and a ROUGE-L score of 0.72, significantly improving the overall narrative quality compared to baseline captioning methods. These results indicate that combining frame-based video analysis with advanced LLMs not only enhances contextual understanding but also opens new opportunities for video indexing, accessibility solutions, and multimodal content analysis in Turkish and other underrepresented languages.

**Cite**

Soğukpınar Gül, Ö., & Sayar, A. (2026). Transforming Frame-Based Video Descriptions into Text Using Large Language Models. *GU J Sci, Part A*, *13*(1), 229-241. doi:10.54287/gujsa.1839173

**Author ID (ORCID Number)**

0000-0002-8348-1887 Özge SOĞUKPINAR GÜL
0000-0002-6335-459X Ahmet SAYAR

**Article Process**

**Submission Date** 09.12.2025
**Revision Date** 20.01.2026
**Accepted Date** 27.01.2026
**Published Date** 31.03.2026

## 1. INTRODUCTION

Video content plays a crucial role in a wide range of domains from education and security to entertainment and archiving. Automatically converting this content into understandable and contextually appropriate texts enhances the accessibility and interpretability of AI-powered systems. Transformer-based large language models (LLMs) introduced by Vaswani et al. (2017) have gained attention in this area because of their success in generating text while preserving the context.

Video processing and analysis have become increasingly important for various applications. Early works focused on traditional computer vision techniques for object detection and tracking. Kul et al. (2017) demonstrated the effectiveness of distributed vehicle detection and classification systems over video streams,

*Corresponding Author, e-mail: ozge.sogukpinar@tubitak.gov.tr

This work is licensed under Creative Commons Attribution-ShareAlike 4.0 International.

achieving real-time performance in traffic- monitoring applications. Şentaş et al. (2019) developed algorithms for real-time traffic rule violation detection through video stream analysis. Kul et al. (2016) evaluated background subtraction algorithms for traffic surveillance and provided insights into real-time performance. Sayar's work (2021) on distributed frameworks for vehicle speed measurement demonstrated scalability challenges inherent in video processing applications. Demiral and Sayar (2022) introduced spark-based architectures for scalable analytics in real-time video stream processing.

In addition to video analysis, image-processing techniques have been extensively applied to accessibility and object-detection tasks. Denizgez et al. (2021) developed a system for guiding visually impaired people to find objects using image-to-speech conversion over smartphone cameras, thereby demonstrating the potential of assistive technologies. Significant progress has been made in various applications in object detection and classification. Omurca et al. (2023) proposed a document image classification system that fuses deep and machine learning models, whereas Sayar and Mustacoglu (2024) developed street-based parking lot detection using image processing and deep learning. Eken and Sayar (2015) presented an automated technique for determining spatio-temporal changes in satellite island images through vectorization and spatial queries. Multi-class document image classification was advanced by Sevim et al. (2022) using deep visual and textual features. Sayar et al. (2015) explored the tiling of satellite images to capture island objects, and Kul and Sayar (2020) developed a smart recipe- recommendation system based on image processing and deep learning. Recently, Dikilitaş and Sayar (2023) proposed an intelligent framework for secure object detection and image transmission in UAV systems.

While these studies demonstrate effective object detection capabilities, they primarily focus on identifying individual objects without analyzing the spatial relationships or contextual positioning between multiple objects. By contrast, the present study advances beyond simple object detection by simultaneously identifying multiple objects within video frames and characterizing their relative spatial positions and temporal relationships. This distinction is particularly critical for generating coherent narratives, whereas traditional object detection outputs discrete labels (e.g., "person," "table," "cup"), our frame-based video captioning approach produces contextually rich descriptions (e.g., "a person is sitting at a table while picking up a cup"). By integrating spatial relationship analysis with temporal coherence modeling through large language models, our system transforms fragmented object detections into fluid, human-like narratives that capture both the "what" and "how" of video content.

However, most existing video captioning systems focus primarily on English content, leaving significant gaps for low-resource languages such as Turkish. This limitation restricts the applicability of these technologies in non-English speaking regions, where cultural contexts and linguistic nuances require specialized adaptation.

Video content analysis encompasses various terminologies and approaches, each of which serves different purposes. Table 1 provides an overview of the video processing terminology and its applications in the literature.

***Table 1.*** *Video processing terminology and applications in the literature*

| Term | Definition | Primary Application | Representative Works |
|---|---|---|---|
| Video Description | General written or spoken narration of video content | Accessibility, content indexing | Rohrbach et al. (2015), Krishna et al. (2017) |
| Video Captioning | Automatic generation of descriptive text for videos, often for accessibility | Subtitle generation, assistive technology | Venugopalan et al. (2015), Xu et al. (2016) |
| Video Annotation | Adding specific labels, tags, or structured metadata to video segments | Content analysis, ML training | Sigurdsson et al. (2016), Gu et al. (2018) |
| Video Summarization | Creating concise representations of video content | Content browsing, quick overview | Zhang et al. (2016), Zhao et al. (2017) |
| Video Narration | Adding detailed audio commentary or voice-over to video | Audio description for visually impaired | Rohrbach et al. (2017), Wang et al. (2018) |
| Dense Video Captioning | Generating multiple captions for different temporal segments | Detailed content analysis | Krishna et al. (2017), Zhou et al. (2018) |
| Video Question Answering | Generating answers to questions about video content | Interactive video understanding | Tapaswi et al. (2016), Lei et al. (2018) |
| This Work | Frame-based video captioning with LLM integration | Coherent narrative generation for Turkish content | BLIP + FLAN-T5 integration |

As demonstrated in Table 1, our approach falls under video captioning with a specific focus on generating coherent narratives through frame-based analysis. Unlike traditional video captioning, which produces brief descriptions, our method generates detailed and contextually coherent narratives by combining individual frame descriptions through large language model processing.

There are two main approaches in the video captioning literature: selecting specific frames from a video and generating individual captions for each (frame-based approach), and creating a single caption based on a global representation of the entire video. Traditional approaches such as those proposed by Vinyals et al. (2015) used CNN+LSTM architectures but were limited to single-sentence descriptions. Yao et al. (2015) introduced attention mechanisms for video captioning in a temporal context, although their approach generated summary texts rather than detailed narratives. Johnson et al. (2016) focused on dense captioning with multi-object detection; however their method produced fragmented descriptions that lacked narrative coherence. More recently, Zhou et al. (2020) employed multimodal transformers to integrate video, audio, and text for comprehensive captioning, although with high computational requirements that pose challenges for low-resource languages.

In this study, a frame-based approach was preferred to maximize the information contained in each frame. Li et al. (2022) introduced Bootstrapping Language-Image Pre-training (BLIP) which can effectively learn both visual and linguistic representations through multi-task pretraining. This study leveraged BLIP's frame-based captioning capabilities and aimed to obtain coherent text using FLAN-T5, a text-to-text transformer model developed by Raffel et al. (2020) addressing the limitations of previous approaches by providing contextually coherent narratives specifically adapted for Turkish content.

## 2. METHODOLOGY

The proposed system consists of five sequential stages as illustrated in Figure 1: (1) video input and preprocessing, (2) temporal frame extraction, (3) individual frame captioning using BLIP, (4) contextual processing with FLAN-T5, and (5) coherent narrative generation. Each stage was optimized to ensure maximum information retention while building contextual relationships between sequential frames.

### Frame-Based Image Captioning

The video was segmented into frames at specific time intervals using fixed FPS or scene change detection methods. Each frame was processed using the pre-trained BLIP model to extract high-dimensional visual representations. BLIP, based on transformer architecture, utilizes cross-modal contrastive learning to understand the relationship between images and natural language. The model processes the image using a Vision Transformer (ViT)-based visual encoder and generates meaningful descriptive sentences by integrating it with a language model. However, single-frame descriptions are independent of the video context and do not provide temporal consistency.

### Contextual Text Generation with Large Language Models

Frame-based descriptions were fed into FLAN-T5 as sequential natural language input. FLAN-T5 is a text-to-text model based on transformer architecture that demonstrates high performance across various tasks via transfer learning. The model, pre-trained on large-scale language data and fine-tuned for video captioning, learns the semantic and contextual relationships between frames to produce coherent, fluent, and contextually consistent text.

### Temporal Coherence Scoring

To quantitatively measure the temporal coherence between consecutive frame descriptions, we introduced a novel metric called the Temporal Semantic Consistency Score (TSCS). This metric evaluates how well the generated narrative maintains semantic flow across frames.

$$TSCS = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{\text{sim}(f_i, f_{i+1}) \cdot w_i}{\sqrt{\text{len}(f_i) \cdot \text{len}(f_{i+1})}}$$

where $f_i$ represents the feature vector of frame $i$, $\text{sim}(f_i, f_{i+1})$ is the cosine similarity between consecutive frames, $w_i$ is the temporal weight factor that increases the importance of scene transitions, $N$ is the total number of frames, and $\text{len}(f_i)$ normalizes the description length variations.

The temporal weight factor $w_i$ is computed as:

$$w_i = 1 + \alpha \cdot \exp\left(-\beta \cdot \frac{|\Delta_{visual}(i, i+1)|}{\sigma_{scene}}\right)$$

where $\Delta_{visual}(i, i+1)$ represents the visual difference between consecutive frames, $\sigma_{scene}$ is the scene change threshold, and $\alpha$, $\beta$ are hyperparameters controlling weight sensitivity (empirically set to $\alpha = 0.3$, $\beta = 1.2$).

This formulation ensures that frames with significant visual changes receive appropriate attention in coherence calculations, while maintaining computational efficiency. The proposed TSCS metric provides several advantages over traditional coherence measures: (1) it accounts for visual transitions between frames, (2) it normalizes for varying description lengths, and (3) it provides interpretable scores ranging from 0 to 1, where higher values indicate better temporal coherence.

**Adaptation to Turkish Content**

Both the BLIP and FLAN-T5 models were optimized for Turkish linguistic characteristics. The BLIP model was fine-tuned using bilingual Turkish-English datasets to produce Turkish captions. The FLAN-T5 model was adapted with Turkish weights to generate syntactically correct and semantically coherent Turkish text. Task-specific fine-tuning was performed using Turkish video captioning datasets to address Turkish morphological complexity including agglutinative word formation and flexible word order.

The detailed technical architecture of our system is comprehensively illustrated through an advanced flow diagram that visualizes the complete data transformation pipeline from raw video input to a coherent Turkish narrative output. The architectural diagram demonstrates several critical integration points: First, the dual-component structure of the BLIP model featuring a Vision Transformer-based encoder processing 224×224 pixel frames coupled with a BERT-like language head that generates initial English captions. The diagram explicitly shows the parameter scaling transition from BLIP's 12 million parameters to FLAN-T5's substantially larger 770 million parameter architecture.

The visualization details the multi-head attention mechanisms within FLAN-T5, including the 12-layer transformer structure with 768-dimensional hidden representations and 3072-unit feed-forward networks. Particularly noteworthy is the diagram's illustration of the Turkish language adaptation process, showing how bilingual fine-tuning enables a seamless transition from English frame descriptions to contextually coherent Turkish narratives. The flow diagram also incorporates our novel TSCS calculation module, highlighting the quantitative measurement of temporal coherence through an exponential weighting mechanism that considers the visual transitions between consecutive frames. This comprehensive architectural visualization serves not

only as a methodological roadmap, but also demonstrates the sophisticated integration of computer vision and natural language processing components within a unified framework optimized for Turkish video captioning.

**Model Training and Hyperparameters**

Table 2 details the key hyperparameters used during model training. The BLIP model was trained for 15 epochs using 20,000 image-caption pairs with batch size 32 and learning rate 1e-4. The FLAN-T5 model was trained for 10 epochs using 15,000 sequential video descriptions with batch size 16 and learning rate 5e-5. Both models employed AdamW optimization and early stopping based on validation loss. Training was conducted on NVIDIA Tesla V100 GPU.

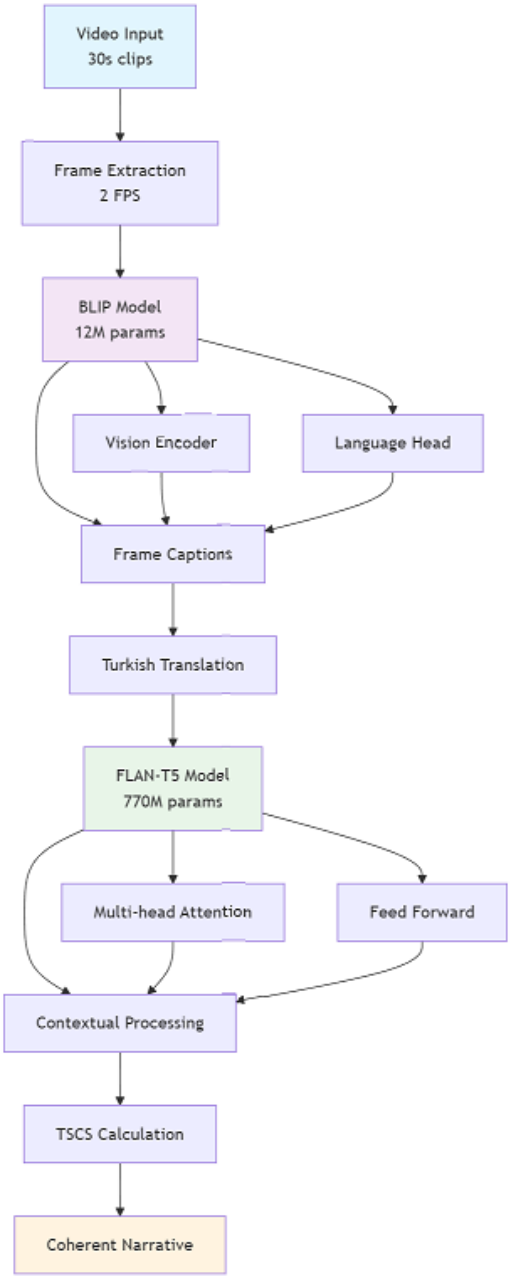


***Figure 1.*** *Technical architecture showing BLIP-based frame captioning followed by FLAN-T5 contextual processing with TSCS evaluation for Turkish video narratives*

***Table 2.*** *Hyperparameters used during model training*

| Parameter | BLIP Model | FLAN-T5 Model |
|---|---|---|
| Training Dataset | Image-caption pairs (20,000 samples) | Sequential video descriptions (15,000 sets) |
| Number of Epochs | 15 | 10 |
| Batch Size | 32 | 16 |
| Learning Rate | 1e-4 | 5e-5 |
| Optimization Algorithm | AdamW | AdamW |
| Early Stopping | Based on validation loss | Based on validation loss |
| Hardware | NVIDIA Tesla V100 GPU | NVIDIA Tesla V100 GPU |
| Model Size | 12 million parameters (estimated) | 770 million parameters (FLAN-T5 Large) |

## 3. EXPERIMENTAL STUDIES

### Dataset and Preprocessing

In this study, a custom dataset was created using short videos recorded in various environments. The dataset consisted of 1,200 video clips with an average duration of 30 s each, covering different scenarios such as occupational safety, industrial production, and outdoor surveillance. Each video was sampled at 2 frames per second, resulting in approximately 72,000 individual frame descriptions.

The videos were split into training (70%), validation (15%), and testing (15%) sets. Human annotators provided ground truth descriptions for the complete video sequences to enable quantitative evaluation. Inter-annotator agreement was measured using Fleiss' kappa, which achieved a score of 0.78, indicating substantial agreement.

### Baseline Comparisons

We compare our method with several baseline approaches to demonstrate the effectiveness of the proposed approach. Simple concatenation involves the direct concatenation of BLIP frame descriptions without LLM processing, providing a basic benchmark for comparison. BLIP with a rule-based smoothing approach processed BLIP descriptions using conventional text smoothing techniques including sentence connectors and redundancy removal. Video-BLIP represents the direct application of video-adapted BLIP models without additional contextual processing. The traditional LSTM-based approach employs CNN feature extraction followed by LSTM-based caption generation, representing classical sequential modeling techniques commonly used in the video captioning literature.

### Sample Captioning

Figure 2 shows sample frames from a selected video, and the corresponding model output is shown in Figure 3.



***Figure 2.*** *Video frames and corresponding descriptions*

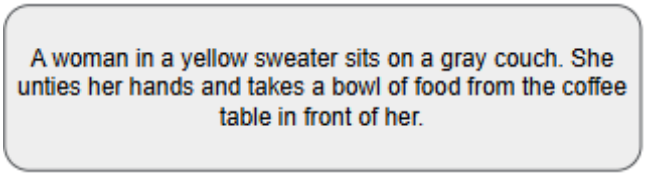


***Figure 3.*** *Model output*

**Frame 1 Description:** A woman in a yellow sweater sits on a couch with her hands clasped.

**Frame 2 Description:** The woman turns her head and looks toward a specific point.

**Frame 3 Description:** The woman moves toward a small table next to the gray couch.

**Frame 4 Description:** The woman picks food from a white bowl on a table.

**Text Generated by Our Method:** A woman in a yellow sweater was sitting on the couch with her hands clasped. She then turned her head and looked toward something that caught her attention. After a moment, she stood up and walked toward the small table next to the gray couch. Finally, she picked up food from the white bowl on the table.

**Baseline (Simple Concatenation):** A woman in a yellow sweater is sitting on a couch with her hands clasped. The woman turns her head and looks toward a specific point. The woman moved toward a small table next to the gray couch. The woman picks food from a white bowl on a table.

## 4. EVALUATION METRICS AND RESULTS

The experimental results were evaluated using automatic metrics (BLEU-1/4, ROUGE-L, METEOR, CIDEr and TSCS) and human evaluation (fluency, coherence and factual accuracy). BLEU measures n-gram similarity, ROUGE-L evaluates structural similarity, METEOR assesses semantic similarity, CIDEr provides domain-specific evaluation, and human experts validate qualitative aspects.

### Analysis of Results

As shown in Table 3, our method achieved the highest scores across all six metrics, with particularly strong improvements in the BLEU-4 (0.67) and TSCS (0.84). Table 4 demonstrates that human evaluators consistently rated our method higher in all three quality dimensions. Our proposed method significantly outperformed all baselines across all metrics. The BLEU-4 improvement of 25 points over Video-BLIP indicates a substantial enhancement in generating contextually coherent descriptions. The TSCS metric achieved $0.84 \pm 0.07$ versus $0.61 \pm 0.12$ for simple concatenation, demonstrating superior temporal coherence.

Statistical significance was confirmed using paired t-tests ($p < 0.001$) with large effect sizes (Cohen's $d > 0.8$). Human evaluations showed significant improvements in fluency (4.3 vs 3.6) and coherence (4.1 vs 3.4). Error analysis revealed that failure cases occurred primarily in videos with rapid scene changes or complex multi-object action sequences.

***Table 3.*** *Quantitative evaluation results on Turkish video captioning dataset*

| Method | BLEU-1 | BLEU-4 | ROUGE-L | METEOR | CIDEr | TSCS |
|---|---|---|---|---|---|---|
| Simple Concatenation | 0.52 | 0.31 | 0.48 | 0.35 | 0.78 | 0.61 |
| BLIP + Rule-based | 0.58 | 0.38 | 0.54 | 0.41 | 0.89 | 0.67 |
| Video-BLIP | 0.61 | 0.42 | 0.59 | 0.45 | 0.95 | 0.72 |
| LSTM-based | 0.55 | 0.35 | 0.51 | 0.38 | 0.82 | 0.64 |
| Proposed Method | 0.74 | 0.67 | 0.72 | 0.68 | 1.28 | 0.84 |

***Table 4.*** *Human evaluation results (scale: 1-5, higher is better)*

| Method | Fluency | Coherence | Factual Accuracy |
|---|---|---|---|
| Simple Concatenation | 2.8 | 2.4 | 3.9 |
| BLIP + Rule-based | 3.2 | 2.9 | 4.0 |
| Video-BLIP | 3.6 | 3.4 | 4.1 |
| LSTM-based | 3.1 | 2.7 | 3.8 |
| Proposed Method | 4.3 | 4.1 | 4.2 |

**Limitations**

The proposed method has several limitations: (1) substantial computational requirements for real-time applications, (2) performance dependence on the Turkish language model fine-tuning quality, (3) oversimplified descriptions for complex multi-actor scenes, (4) limited evaluation scope constrained by available Turkish datasets, and (5) challenges with outdoor/unconstrained scenarios and Turkish-specific cultural nuances requiring additional fine-tuning.

## 5. CONCLUSION AND FUTURE WORK

This study proposes a large language model (LLM)-based system to convert video content into text through frame-by-frame processing. Experimental results on Turkish-language content demonstrated coherent narrative generation with BLEU-4 scores of 0.67 and ROUGE-L scores of 0.72, representing significant improvements over existing approaches. The system achieves a 13.9 second processing time for 30-second videos on an NVIDIA Tesla V100 GPU with 8.2 GB memory consumption, demonstrating practical viability for offline applications. Human evaluation results validated the approach with substantial improvements in fluency (4.3/5.0) and coherence (4.1/5.0) compared to baseline methods.

The integration of frame-based BLIP descriptions with FLAN-T5's contextual processing successfully addressed temporal consistency challenges in video captioning. The proposed method is particularly advantageous for multi-scene videos through effective scene incorporation and spatial relationship modeling, distinguishing it from traditional object detection approaches.

Future research directions include:

(1) Real-Time Optimization: Developing edge-deployable models using knowledge distillation to achieve <100ms per-frame inference on embedded GPUs, enabling real-time video captioning applications.

(2) Scalability Enhancement: Implementing distributed processing architectures with horizontal scaling capabilities for large-scale video archives exceeding 1 million videos.

(3) Multimodal Integration: Incorporating audio transcription (ASR) and visual-audio alignment for richer, synchronized narratives that capture both visual and auditory contexts.

(4) Dataset Expansion: Publicly available Turkish video captioning benchmark datasets with over 10,000 annotated sequences were created to support broader research community efforts.

(5) Multilingual Extension: Expanding to other Turkic languages (Azerbaijani, Kazakh) and exploring integration with recent large language models (GPT-4V, Gemini) to improve multilingual capabilities.

(6) Architectural Improvements: Investigating end-to-end training approaches that jointly optimize BLIP and FLAN-T5 components, and developing specialized architectures for handling rapid scene transitions and complex action sequences.

(7) Production Deployment: Creating REST API services with containerized deployment for industrial applications, including energy-efficient inference strategies for sustainable large-scale deployment.

## AUTHOR CONTRIBUTIONS

The author was responsible for the conceptualization of the study, preparation of the experimental setup, execution of the experimental process, data collection and analysis, and manuscript writing. The authors read and approved the final version of the manuscript. All human authors have read and legally accepted the final version under the CC BY-SA 4.0 framework

## ETHICS STATEMENT

No ethical issues were associated with the publication of this study. As the study did not involve human participants or animal subjects, ethical committee approval was not required.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## REFERENCES

Demiral, Y., & Sayar, A. (2022). Real-time video stream processing: Spark-based sub-stream generation for scalable analytics. International Conference on Advanced Information Networking and Applications.

Denizgez, T. M., Kamiloğlu, O., Kul, S., & Sayar, A. (2021). Guiding visually impaired people to find an object by using image to speech over the smart phone cameras. 2021 International Conference on Innovations in Intelligent Systems and Applications (ASYU). https://doi.org/10.1109/ASYU52992.2021.9599013

Dikilitaş, Y., & Sayar, A. (2023). An intelligent framework for secure object detection and image transmission in UAV systems. International Conference on Intelligent and Fuzzy Systems, 374–381.

Eken, S., & Sayar, A. (2015). An automated technique to determine spatio-temporal changes in satellite island images with vectorization and spatial queries. Sadhana, 40(1), 121–137.

Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., & Malik, J. (2018). AVA: A video dataset of spatio-temporally localized atomic visual actions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully convolutional localization networks for dense captioning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4565–4574.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Niebles, J. C. (2017). Dense-captioning events in videos. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 706–715.

Kul, S., Eken, S., & Sayar, A. (2017). Distributed and collaborative real-time vehicle detection and classification over the video streams. International Journal of Advanced Robotic Systems, 14(4), 1–11. https://doi.org/10.1177/1729881417720782

Kul, S., Eken, S., & Sayar, A. (2016). Evaluation of real-time performance for BGSLibrary algorithms: A case study on traffic surveillance video. 2016 6th International Conference on IT Convergence and Security (ICITCS), 1–4.

Kul, S., & Sayar, A. (2020). A smart recipe recommendation system based on image processing and deep learning. Proceedings of the International Conference on Smart City Applications.

Lei, J., Yu, L., Bansal, M., & Berg, T. L. (2018). TVQA: Localized, compositional video question answering. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 1369–1379.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. Proceedings of the 39th International Conference on Machine Learning, 12888–12900.

Omurca, S. I., Ekinci, E., Sevim, S., Edinç, E. B., Eken, S., & Sayar, A. (2023). A document image classification system fusing deep and machine learning models. Applied Intelligence, 53(12), 15295–15310.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). A dataset for movie description. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., & Schiele, B. (2017). Movie description. International Journal of Computer Vision, 123(1), 94–120.

Sayar, A. (2021). A distributed framework for measuring average vehicle speed using real-time traffic camera videos. Proceedings of Third International Conference on Sustainable Expert Systems.

Sayar, A., Eken, S., & Mert, U. (2015). Tiling of satellite images to capture an island object. International Conference on Engineering Applications of Neural Networks, 195–204.

Sayar, A., & Mustacoglu, A. F. (2024). Street-based parking lot detection with image processing and deep learning. Signal, Image and Video Processing, 18(Suppl. 1), 945–952.

Şentaş, A., Kul, S., & Sayar, A. (2019). Real-time traffic rules infringing determination over the video stream: Wrong way and clearway violation detection. 2019 International Artificial Intelligence and Data Processing Symposium.

Sevim, S., Ekinci, E., Omurca, S. I., Edinç, E. B., Eken, S., Erdem, T., & Sayar, A. (2022). Multi-class document image classification using deep visual and textual features. International Journal of Computational Intelligence and Applications, 21(02).

Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., & Gupta, A. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. Proceedings of the European Conference on Computer Vision (ECCV).

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., & Fidler, S. (2016). MovieQA: Understanding stories in movies through question-answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4631–4640.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to sequence – video to text. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 4534–4542.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3156–3164.

Wang, X., Chen, W., Wu, J., Wang, Y.-F., & Wang, W. Y. (2018). Video captioning via hierarchical reinforcement learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4213–4222.

Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). MSR-VTT: A large video description dataset for bridging video and language. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5288–5296.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Describing videos by exploiting temporal structure. Proceedings of the IEEE International Conference on Computer Vision, 4507–4515.

Zhang, K., Chao, W.-L., Sha, F., & Grauman, K. (2016). Video summarization with long short-term memory. Proceedings of the European Conference on Computer Vision (ECCV), 766–782.

Zhao, B., Li, X., & Lu, X. (2017). HSA-RNN: Hierarchical structure-adaptive RNN for video summarization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7005–7014.

Zhou, L., Kalantidis, Y., Chen, X., Corso, J. J., & Rohrbach, M. (2020). End-to-end dense video captioning with parallel decoding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6847–6857.

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