Journal of Big Data

RESEARCH Open Access

Bilingual video captioning model for enhanced video retrieval

Norah Alrebdi[1*] and Amal A. Al-Shargabi[1]

*Correspondence:
n.a.alrebdi@gmail.com

[1] Department of Information Technology, College of Computer, Qassim University, 51452 Buraydah, Saudi Arabia

Abstract

Many video platforms rely on the descriptions that uploaders provide for video retrieval. However, this reliance may cause inaccuracies. Although deep learning-based video captioning can resolve this problem, it has some limitations: (1) traditional keyframe extraction techniques do not consider video length/content, resulting in low accuracy, high storage requirements, and long processing times; (2) Arabic language support in video captioning is not extensive. This study proposes a new video captioning approach that uses an efficient keyframe extraction method and supports both Arabic and English. The proposed keyframe extraction technique uses time- and content-based approaches for better quality captions, fewer storage space requirements, and faster processing. The English and Arabic models use a sequence-to-sequence framework with long short-term memory in both the encoder and decoder. Both models were evaluated on caption quality using four metrics: bilingual evaluation understudy (BLEU), metric for evaluation of translation with explicit ORdering (METEOR), recall-oriented understudy of gisting evaluation (ROUGE-L), and consensus-based image description evaluation (CIDE-r). They were also evaluated using cosine similarity to determine their suitability for video retrieval. The results demonstrated that the English model performed better with regards to caption quality and video retrieval. In terms of BLEU, METEOR, ROUGE-L, and CIDE-r, the English model scored 47.18, 30.46, 62.07, and 59.98, respectively, whereas the Arabic model scored 21.65, 36.30, 44.897, and 45.52, respectively. According to the video retrieval, the English and Arabic models successfully retrieved 67% and 40% of the videos, respectively, with 20% similarity. These models have potential applications in storytelling, sports commentaries, and video surveillance.

Keywords: Artificial intelligence, Computer vision, Natural language processing, Video retrieval, English video captioning, Arabic video captioning

Introduction

In recent times, video-processing communities have paid significant attention to video captioning. This task combines two main areas of artificial intelligence: computer vision (CV) and natural language processing (NLP). This combination results in many exciting applications, such as the DALL.E [1] and OrCam MyEye [2], which use deep learning methods to enable computers to understand and analyze visual content as humans do [3, 4]. NLP uses various computational techniques to learn, analyze, and produce human language [5], which enables computers to understand the visual content of videos and describe it in natural language. Video captioning comprises multiple steps, such as keyframe extraction, feature extraction, and caption generation. Figure 1 shows the general workflow for video captioning.

Using video captioning to improve video retrieval has garnered significant attention [6–8]. With the rising popularity of video platforms, such as TikTok, YouTube, and Snapchat, there has been an exponential increase in the number of published and shared videos. According to a survey conducted in 2019 [9], over 500 h of videos are uploaded to YouTube every minute. Given the enormous volume of available videos, developing efficient video retrieval techniques is important to help users find specific videos quickly.

Keyframe extraction is an essential step in video processing because it selects crucial frames and removes unnecessary frames to summarize a video [10, 11]. However, most video captioning studies have not focused on keyframe extraction. Each second of a video typically produces approximately 24–30 frames, commonly called frames per second (FPS) [12]. Therefore, the video captioning task must extract keyframes for processing instead of processing all the frames, which would result in a massive number of frames requiring several resources and a long processing time. However, most existing studies use simple approaches, such as time- and frame-based approaches, which neglect video content. The content-based keyframe extraction approach is considered a supervised process that studies the frames before extraction to enhance the keyframe extraction process, thus enabling the extraction of all the important details in the videos with minimum duplication. Another problem in video captioning studies is the lack of models in Arabic and other languages, excluding English. Therefore, this study proposes a deep learning video captioning model that enhances video retrieval and implements a new keyframe extraction method that contains two phases: time- and content-based. This new approach improves the accuracy of captions and reduces processing time and storage requirements. This study is the first to conduct video captioning in Arabic.

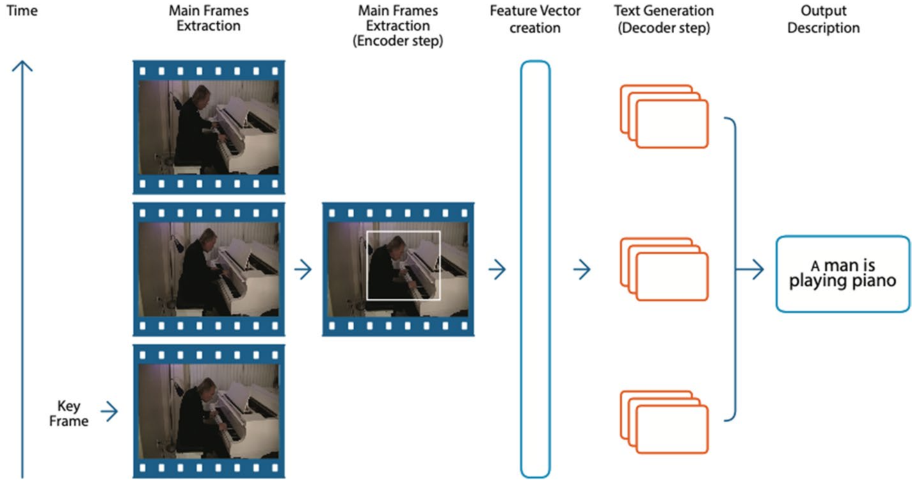


Fig. 1 General workflow for video captioning

The contributions of this study are summarized as follows:

- Implementing a new keyframe extraction method that contains two phases: time- and content-based.
- Conducting the first Arabic video captioning model.

The remainder of this paper is organized as follows. "Related work" section reviews the current studies on video captioning, such as keyframe and caption generation models. "Methodology" and "Experimental setup" sections describe the proposed methodology and experimental setup. "Results" section presents the results of the study. Finally, "Conclusion and future work" section concludes the paper and suggests future work.

Related work

This section reviews the literature on keyframe extraction and video captioning. Each of these aspects is discussed in the following subsections.

Keyframe extraction

Most video captioning studies focus on enhancing video captioning models but use simple approaches for keyframe extraction, despite their importance in improving performance. Several keyframe extraction approaches exist, such as shot boundary [13], segmentation [14], and clustering-based [14]. However, this study classifies keyframe extraction approaches into three categories: time-, frame-, and content-based, depending on how they are used in video captioning. The time-based approach is implemented using a specific duration, such as extracting one frame every 1 or 2 s. However, the frame-based approach uses a pre-specified number of frames, that is, extracting 10 or 100 frames per video. Conversely, the content-based approach extracts keyframes based on the frames' content using different techniques, such as content similarity measurement [15] or shot boundary detection [16].

Xu et al. extracted one keyframe per second [17]. Conversely, the authors of another study extracted a specified number (240) of keyframes for each video, regardless of its length [18]. However, other studies have used different content-based approaches to extract keyframes. For example, Qian et al. proposed a reinforcement learning filtration network that uses an actor-double-critic algorithm to filter duplicate frames [19]. Furthermore, to extract video keyframes, one study used dynamic programming and clustering after reducing the dimensionality of video frames [20]. Additionally, the model of one of the studies used ResNet [21] in a locally consistent deformable convolution to detect important regions in each frame, from which it identified whether a given frame is a keyframe [22]. Yet another study developed a model to detect the video's abrupt transitions and extract keyframes [23]. This method first extracts binary edge information before measuring the Euclidean distance [24] on the histogram features of adjacent frames and the Z-score on the magnitude of the Sobel gradient images [25]. Subsequently, the model calculates the variance coefficient, and the frames with the highest values are selected as keyframes. Furthermore, the model presented in the video classification study extracts keyframes using the structural similarity index measurement (SSIM) value [15]. This model measures the similarity of a specific region, depicting its main actions. Additionally, a study used coarse extraction to extract keyframes using the motion states of objects and local and global changes [13].

Although extracting keyframes using a specific number of frames or a certain duration is a simple and fast method, it is random and disregards the frame content. Therefore, this method may hinder the video captioning models' performance. However, despite advancements in the filtration keyframe extraction method [16], this method must pass through each frame, which is time-consuming. Additionally, this method completely disregards duplicate frames and filters similar frames. Therefore, this method produces multiple frames as keyframes owing to the nature of video frames, where adjacent frames are typically similar. However, the dimensionality reduction method reduces the processing time and storage requirements. This is a difficult process that requires the compression or expansion of frames to a specific size.

Three main approaches, time-, frame-, and content-based, are used to summarize the keyframe extraction approaches found in the literature review. Table 1 summarizes the reviewed keyframe extraction methods and their limitations.

As shown in the table, the time- and frame-based keyframe extraction approaches are easy to implement but inaccurate. By contrast, the methods related to the content-based approach require a long processing time. Thus, in this study, keyframe extraction was implemented in two steps using two approaches: the time-based approach to speed up the first extraction step and the content-based approach to reduce frame duplications.

Video captioning

Early video captioning studies described videos using a subject-verb-object (SVO) template. For example, a study captioned videos from the DARPA dataset using Stanford parser [27] to extract SVO features [28]. Another study used the same dataset but added information about the activities, such as who acted, what the action was, where it occurred, and why [29]. However, only 118 words, including verbs, nouns, pronouns, adjectives, adverbs, prepositions, prepositional phrases, determiners, and particles, were used in the sentence generation process. Thomason et al. added

Table 1 A summary of the reviewed keyframe extraction methods

Ref.	Year	Approach/method	Weaknesses
[13]	2022	Content-based: coarse extraction and partial-fine re-extraction of spatiotemporal slices	Not suitable for all videos (especially videos that contain fast scenes)
[15]	2021	Content-based: SSIM	Applied to a specific regions, not a complete frame
[17]	2015	Time-based	Inaccurate because it is based on time (one frame each second)
[18]	2015	Frame-based	Inaccurate because it is based on the number of frames (240 frames per video)
[19]	2021	Content-based: filtration network RL-based	Requires efficient training
[20]	2022	Content-based: multiview fusion method-based	Complicated method Requires specific frame sizes
[22]	2022	Content-based: local consistent deformable convolution	Long processing time
[23]	2020	Content-based: Sobel gradient images and variance coefficient measure	Long processing time compared to other similar approaches [26]

location information to sentence templates to improve the SVO [30]. Therefore, this model uses detection confidence to detect features and probabilistic knowledge from text-mining to select an SVO and place (P). Another study used deep neural networks (DNNs) to extract features and create a tree structure, whereas Stanford's parser [27] was used to construct a descriptive caption [17]. Krishnamoorthy et al. proposed a model that requires only a few training captions because it learns from web-based text-mined knowledge [31]. Table 2 summarizes the template-based video captioning studies reviewed.

However, most video captioning studies have used a free-template-based approach using the decoder–encoder model. For example, Qian et al. developed a model that used a CNN and gated recurrent unit (GRU) in the encoder and decoder, respectively, to extract 2D and 3D features and generate captions [19]. Similarly, Peng et al. developed a model for extracting 2D and 3D features and used local and global text attention to enhance the quality of the generated captions [32]. However, Liu et al. proposed a sibling convolutional encoder comprising content and semantic branches, which encode notable visual information and semantic information, respectively [33].

Lee and Kim extracted visual features, temporal features, and semantic data using bidirectional long short-term memory (BLSTM) rather than long short-term memory (LSTM) to efficiently detect events over time [34]. This study focused on prepositions and conjunctions, using context gating and soft attention to improve sentence generation. Similarly, a decoder presented in another study uses soft attention to link verbal and visual materials and improves the creation of semantic captions [35]. A variational stacked local attention network was proposed to increase the diversity of captions [36]. The captions produced in this study included adverbs, adpositions, adjectives, determiners, and numbers.

However, the models proposed in some studies use reinforcement learning (RL) and the encoder–decoder approach; for example, the decoding part in a model proposed in [37] served as an agent, whereas the video features and sequence of words served as the environment. The agent takes a probability distribution using the policy and performs an action, which is next-word prediction. After generating all the caption words, the agent takes the score as a reward to update the internal parameters. Zheng et al. proposed a stacked multimodal attention network (SMAN) using an RL and a coarse-to-fine training approach [38]. The model uses SMAN to capture and describe visual and textual data. The RL and training strategies enhance the generated captions. Adversarial learning is another technique used in video

Table 2 Summary of the reviewed structured-template-based studies

Ref.	Year	Extracted information	Method
[17]	2015	Subject, verb, object	Stanford parser—DNN
[28]	2012	Subject, verb, object	Stanford parser
[29]	2012	Subject, verb, object, additional details (adverbs, adjectives, …)	Detectors—Kanade–Lucas–Tomas—HMMs—dynamic-programming algorithm
[30]	2014	Subject, verb, object, place	Factor graph model
[31]	2013	Subject, verb, object	Text-mining knowledge

Table 3 Summary of the free-template-based studies

Ref.	Year	Method	Dataset	Evaluation metrics			
				B	M	R	C
[19]	2021	CNN-GRU	MSVD	57.9	37.4	74.7	96.3
			MSR-VTT	45.1	28.6	61.8	51.5
[32]	2021	CNN-GRU	MSVD	55.1	36.4	72.2	85.7
			MSR-VTT	42.3	28.9	61.7	49.2
[33]	2021	CNN-RNN	MSVD	54.2	34.8	71.7	88.2
			MSR-VTT	40.9	27.5	60.2	47.5
[34]	2021	CNN-BiLSTMs	MSVD	41.8	–	–	60.1
			ActivityNet	32.1	–	–	25.7
[35]	2022	CNN-LSTM	MSVD	43.7	32.3	68.8	70.7
[36]	2022	CNN-LSTM	MSVD	57.4	36.9	75.6	98.1
			MSR-VTT	46.5	32.8	55.8	62.4
[37]	2021	CNN-LSTM and RL	MSVD	52.3	35.0	71.9	84.3
			MSR-VTT	41.1	27.5	60.4	47.0
[38]	2022	CNN-GRU and RL	MSVD	52.5	35.0	72.4	94.5
			MSR-VTT	41.3	28.7	62.1	53.8
[40]	2018	CNN-LSTM and GAN	MSVD	42.9	30.4	–	–
			MSR-VTT	36.0	26.1	–	–
			M-VAD	–	63.0	–	–
			MPII-MD	–	72.0	–	–

MSVD microsoft research video description, *MSR-VTT* microsoft research video to text, *MPII-MD* Max Planck Institute for Informatics-Movie Description

captioning models. A study developed a model that generates captions using a generative adversarial network (GAN) [39] with an LSTM [40]. Table 3 shows the free-template-based video captioning studies reviewed.

As illustrated in Table 2, the SVO template is the most widely used template in structured-template-based video captioning. However, presently, very few studies have used this approach owing to the low accuracy of the generated captions. According to the authors' perspective, the structured-template-based approach tends to be more of a classification task than a captioning task. Thus, most of the current studies follow a free-template-based video captioning approach. Table 3 illustrates that the encoder–decoder technique is the most widely used in free-template-based video captioning studies, and it has achieved superior results compared with the others. Thus, according to the literature review, this study follows the free-template-based video captioning approach and uses an encoder–decoder technique.

Methodology

This section presents this study's methodology, including the dataset used, keyframe and feature extraction phases, English and Arabic video captioning models, and the evaluation phase. Figure 2 shows the general architecture of the proposed method. The methodology of each phase is described in the following subsections.

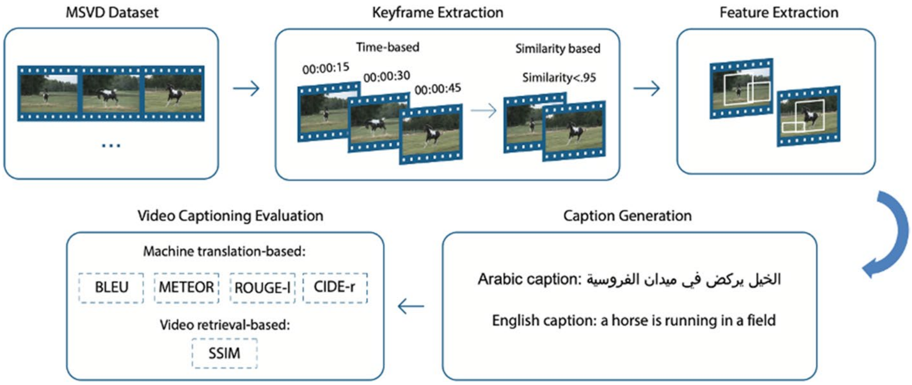


Fig. 2 Architecture of the proposed methodology

Dataset

This study used the MSVD dataset, which is the most widely used dataset for video captioning. MSVD is a general-domain video-captioning dataset that Microsoft provided in 2010. The videos in the dataset were gathered from YouTube and described by AMT workers. The original dataset contained 2089 videos and 85,550 English captions (with an average of 41 captions per video) [41]. Figure 3 shows a sample video and its associated captions.

The dataset used in this study contained 1970 videos and 80,827 captions. The lengths of the videos were between 1.7 and 60 s. Table 4 shows the statistics of the

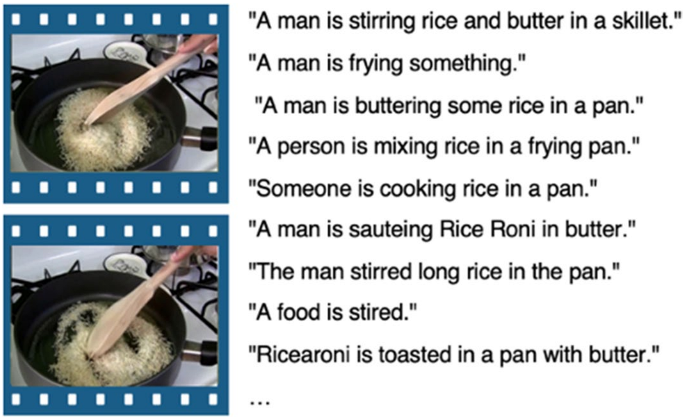

"A man is stirring rice and butter in a skillet."
"A man is frying something."
"A man is buttering some rice in a pan."
"A person is mixing rice in a frying pan."
"Someone is cooking rice in a pan."
"A man is sauteing Rice Roni in butter."
"The man stirred long rice in the pan."
"A food is stired."
"Ricearoni is toasted in a pan with butter."
...

Fig. 3 Sample of reference captions of an MSVD video

Table 4 Video specifications of the MSVD dataset

Metric	Value
Longest video	60.03 s
Shortest video	1.74 s
Maximum number of FPS	59.92
Minimum number of FPS	6.0
Maximum number of frames	1799
Minimum number of frames	41

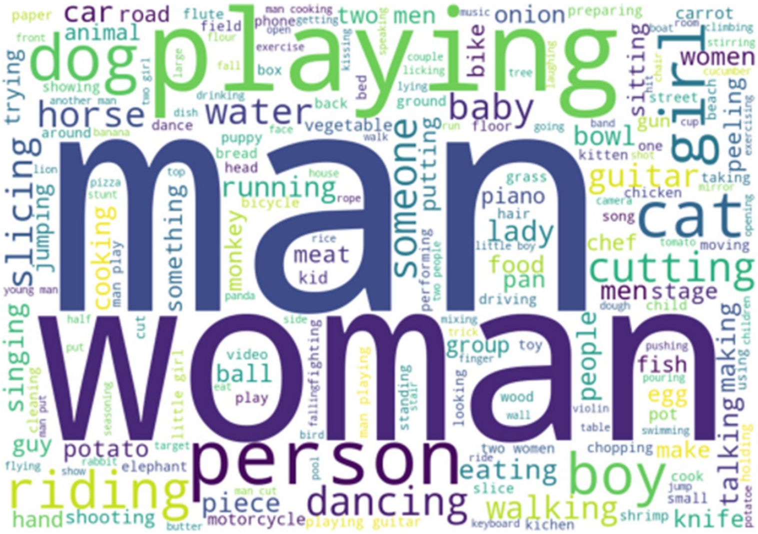


Fig. 4 Word cloud of the MSVD dataset

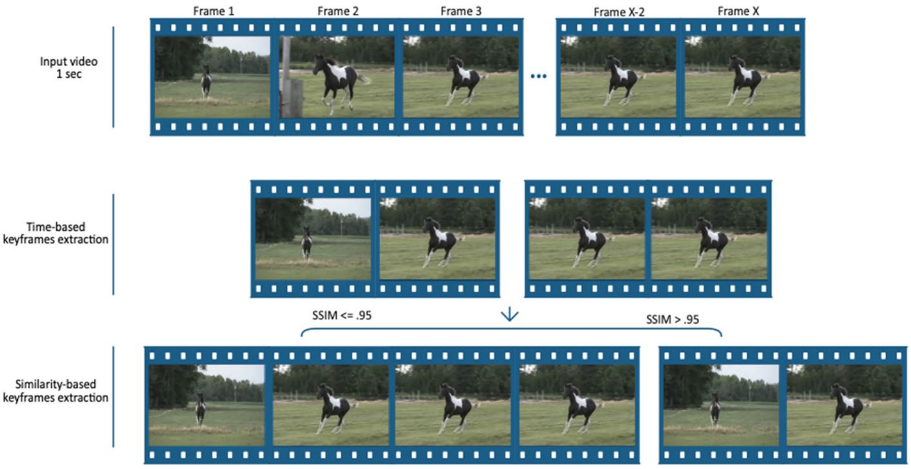


Fig. 5 Overall architecture of keyframe extraction

videos in the dataset. The video content varied, showing general activities, such as sports, cooking, and playing. Figure 4 shows a word cloud of the MSVD captions.

Keyframe extraction

Video keyframe time- and content-based extractions are the two steps followed to achieve the study's objective, which was to efficiently extract video keyframes at a low cost. Figure 5 shows the general workflow of the two steps. The following subsections describe each step's methodology.

Time-based phase

A frame is a single still image obtained from a series of still images in a video. A video generates 24–30 frames per second (FPS) [12]. Therefore, a short video (approximately 10–20 s) produces 240–480 frames at the minimum. Particularly, video frames

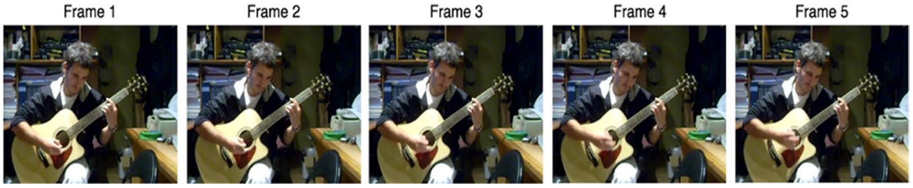


Fig. 6 Sequenced frame similarities

in the MSVD dataset ranged from 41 to 1799, as shown in Table 4. Processing all the frame is expensive, particularly if they are duplicates. Furthermore, because each video in the MSVD dataset showed a single event, changes in the video over time were minimal. A sequence of five frames from a video is shown in Fig. 6. To remove duplicate frames in a simple and inexpensive way, several experiments were conducted during this phase (extracting 1 FPS, one frame every half second, and one frame every quarter second).

Content-based phase

According to the time-based extracted keyframes, duplicate frames still exist, as shown in Fig. 6. Therefore, a content-based keyframe extraction phase must reduce duplication, processing time, and storage requirements. Various methods exist for extracting keyframes using the content. However, using the specifications of the dataset used (a few shots and one event per video), the authors believe that a similarity-based approach is appropriate for this phase. SSIM was used to extract keyframes for a classification task in [42], which is the underpinning method for this study. The method was used to extract video captioning keyframes during this phase. Therefore, several experiments were conducted to determine suitable similarity thresholds (70%, 80%, 90%, and 95%).

During this phase, the similarity between every two adjacent frames is calculated from one side (forward direction) using SSIM, as follows [43]:

$$SSIM(X,Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{({\mu^2}_X + {\mu^2}_Y + C_1)({\sigma^2}_X + {\sigma^2}_Y + C_2)} \quad (1)$$

However, if the similarity between two adjacent frames is less than the threshold, both frames are considered keyframes; otherwise, the first frame is selected as a keyframe, and the second is ignored. This process continues until no adjacent frames are at least 90% similar. Figure 7 shows an example of the similarity-based keyframe extraction phase.

Feature extraction

To standardize the array size of the extracted features, the model feeds 40 frames from the extracted keyframes from the previous step (time- and content-based keyframe extraction) of each video into the visual geometry group 16 (VGG16), which is pre-trained on ImageNet [44]. However, the remaining space in the array is empty if the videos contain fewer than 40 keyframes.

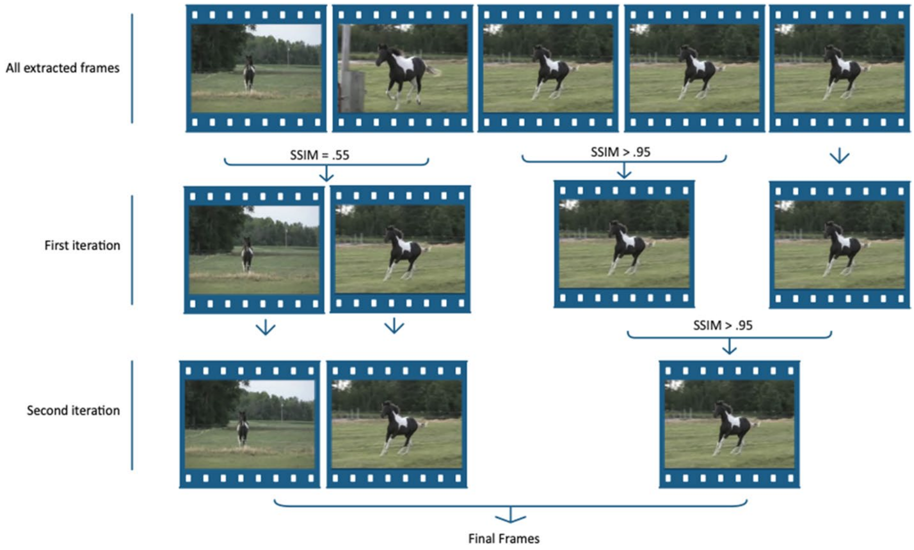


Fig. 7 Architecture of similarity-based keyframe extraction phase

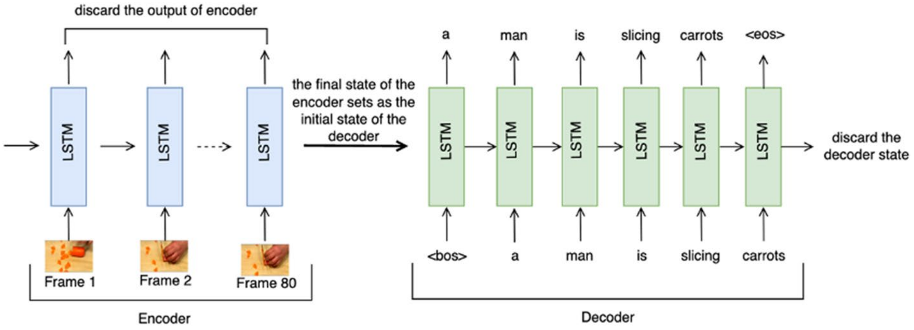

discard the output of encoder
LSTM
LSTM
LSTM
Frame 1
Frame 2
Frame 80
Encoder
the final state of the encoder sets as the initial state of the decoder
a
man
is
slicing
carrots
<eos>
<bos>
a
man
is
slicing
carrots
discard the decoder state
Decoder

Fig. 8 Architecture of the training model [45]

Video captioning model

This study used the same model to perform video captioning in English and Arabic and adopted an encoder–decoder model, particularly the LSTM–LSTM [45] model, owing to its effectiveness in previous video captioning studies. LSTM is a type of RNN proposed in [46] containing a memory cell that maintains information for a long time to resolve the long-term dependency issue. The overall architecture of the training model is shown in Fig. 8.

Figure 8 shows the implemented architecture comprising two LSTM layers with different numbers of units. The first LSTM layer is used in the encoder to learn the structure of the video content, whereas the second layer is used in the decoder to learn the structure of the descriptive sentence. Forty extracted frames are fed sequentially into the LSTM cells in the encoder. Additionally, the output of each cell is input into the next cell. Finally, the last cell's output, i.e., the "final state," is inputted into the decoder. However, each decoder cell inputs one word of the caption from the training caption list, corresponding to the encoder frames.

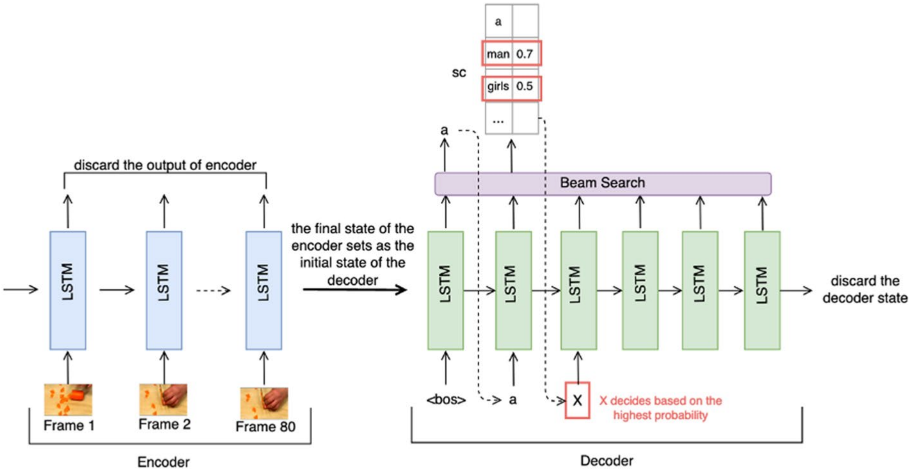

discard the output of encoder
LSTM
Frame 1
Frame 2
Frame 80
Encoder
the final state of the encoder sets as the initial state of the decoder
a
man 0.7
girls 0.5
...
sc
Beam Search
<bos>
X
X decides based on the highest probability
discard the decoder state
Decoder

Fig. 9 Architecture of the caption generation model

Similarly for the training model, the generation model uses the LSTM in the encoder and decoder. Beam search [47] is a search method that uses conditional probability to select any number of proper sentences using a user-specified beam size k. Despite being slow, beam search was chosen because it generates optimal sentences according to some studies [33, 48]. As the first step in beam search, the likelihood of each token being the first term in a sentence is calculated. The k tokens with the highest likelihood are then selected. Similarly, the likelihood is calculated for each candidate token with all tokens to select the best combination of two sequence words. This process continues until the sentence ends with an <eos>, where the final candidate sentence is the one with the highest likelihood. Figure 9 shows the caption generation model's architecture.

Evaluation approach

The proposed model was assessed from the perspectives of machine translation and video retrieval. This section describes the approaches used to address each aspect.

Machine translation-based approach

Similar to most video captioning studies, this study evaluated the proposed method using four machine translation metrics. First, this model was evaluated using bilingual evaluation understudy (BLEU) [49], which calculates common terms between reference and generated captions. The second is the metric for the evaluation of translation with explicit ORdering (METEOR) [50], which considers synonymy and stemming. The third metric is the recall-oriented oriented understudy of gisting evaluation (ROUGE) [51], which calculates the similarity of sequences between the reference and generated captions. The last is the image captioning metric, the consensus-based image description evaluation (CIDE-r) [52], which calculates the cosine similarity between reference and generated captions. The scalar value of BLEU is usually [0–100], METEOR and ROUGE are [0–1], and CIDE-r is from 0 to infinity. However, for clarity, this study normalized the results of the first three measures to be from 0 to 100.

Table 5 Comparison of Arabic-translated captions

English sentence	Translated caption in Arabic		
	Google	**MyMemory**	**Bing**
The lady cracked the egg in to the bowl	كسرت السيدة البيضة في الوعاء	السيدة كسرت البيضة في الوعاء	قامت السيدة بتكسير البيضة في الوعاء
A man is playing a large flutelike instrument	رجل يعزف على آلة موسيقية كبيرة **تشبه الفلوت**	رجل يعزف على آلة موسيقية كبيرة	رجل يعزف على آلة كبيرة تشبه الفلوت
A woman is frying onion pieces and adding shredded carrot	امرأة تقلى البصل وتضيف جزرة مبشورة	امرأة تقلى قطع البصل وتضيف الجزر المبشور	امرأة تقلى قطع البصل وتضيف الجزر المبشور
A person is paring an apple with a knife	شخص **يقشر** تفاحة بسكين	شخص **يقلب** تفاحة بسكين	شخص **يقلص** تفاحة بسكين

The bold typeface indicates the correct translation, while bolditalic typeface indicates the incorrect ones

Table 6 Dataset specifications

Dataset split	# of videos	# of captions	# of captions used (English experiment)	# of captions used (Arabic experiment)
Training	1200	49,142	47,472	34,477
Validation	100	3990	3870	2836
Testing	670	27,695	27,695	27,695
Total	1970	80,827	79,037	65,008

Video retrieval-based approach

In addition to the machine translation-based evaluation, this study evaluated the proposed model based on video retrieval using the cosine similarity [53] between the retrieval query (ground truth) and video description (generated caption).

Experimental setup

This section describes the data preprocessing for the English and Arabic models. Furthermore, the data preparation and parameter settings used are presented.

Data preprocessing

The English dataset underwent two different types of preprocessing: all caption letters were converted into lowercase and all punctuation, including commas and full stops, were removed. For the Arabic experiment, the preprocessed English MSVD dataset mentioned in "Feature extraction" section was translated into Arabic because there is currently no Arabic video captioning dataset. As shown in Table 5, three Arabic-supported translators were compared: Google [54], MyMemory [55], and Bing [56]. Based on this comparison, the authors selected Google Translator because it outperformed the others in the translation of some vocabulary. For example, it perfectly translated "large flutelike instrument" and "paring" action as "تشبه الفلوت" and "يقشر" instead of "يقلص" or "يقلب". Additionally, the Google Translator has been used in many studies, such as [57].

Dataset preparation

As mentioned previously, this study used the MSVD dataset containing 1970 videos. As in many earlier studies [33, 58–60], the dataset was randomly divided into training, validation, and testing sets with 1200, 100, and 670 videos, respectively. The English reference captions were between one and forty-five words long, whereas the Arabic captions ranged from one to forty-six words long. These differences in the lengths of the reference captions may have negatively affected the caption generation model. Therefore, this study conducted several experiments to identify the best range of training sentences for both the English and Arabic datasets. Finally, using most captions' lengths, the model was trained on captions between six and thirty words, including the <bos> and <eos> tags. However, captions with fewer than 30 words were added.

The total number of trained English captions was 51,342, with 16–78 captions for each video. Particularly, 47,472 captions were included in the training set and 3870 in the validation set. Conversely, the Arabic dataset had a total of 37,313 with 6–63 reference

Table 7 Number of extracted keyframes in the implemented experiments

Criteria	Training	Validation	Testing	Total
First experiment				
Time-based 50	21,248	1843	13,826	36,917
Similarity-based 90	19,499	1744	12,866	34,109
Second experiment				
Time-based 25	43,097	3730	28,008	74,835
Similarity-based 90	37,374	3367	24,669	65,410
Third experiment				
Time-based 25	43,097	3730	28,008	74,835
Similarity-based 95	41,128	3597	27,013	71,738
Original experiment	304,564	26,550	210,752	541,866

Table 8 Processing time of the implemented experiments

Criteria	Training	Validation	Testing
First experiment			
Time-based 50	15 m	2 m	14 m
Similarity-based 90	6 h 3 m	6 m	2 h 35 m
Second experiment			
Time-based 25	26 m	3 m	25 m
Similarity-based 90	12 h 34 m	1 h 12 m	7 h 33 m
Third experiment			
Time-based 25	26 m	3 m	25 m
Similarity-based 95	13 h 16 m	41 m	15 h 4 m
Original experiment	46 m	4 m	40 m

captions for each video. The training set contained 34,477 captions, while the validation set contained 2836 captions. All captions were used as the testing set, regardless of the caption length. Table 6 shows the specifications of the English and Arabic datasets used in the study.

Parameters settings

In the training step, the model was trained using the most common 6000 and 8000 tokenizers in the training set out of 9771 and 14,964 tokenizers in English and Arabic, respectively. Additionally, the LSTM encoder used 40 cells using the maximum number of trained frames, with 4096 features extracted from each frame. Conversely, the LSTM decoder used 30 cells with 6000 and 8000 tokens in English and Arabic, respectively. The batch size was set to 320. The adaptive moment estimation (Adam) algorithm [61] is an adaptive algorithm used with a learning rate of 0.0003. The experiments were conducted using Keras and TensorFlow. However, during the generation step, the model used the same encoder–decoder model with a beam search technique.

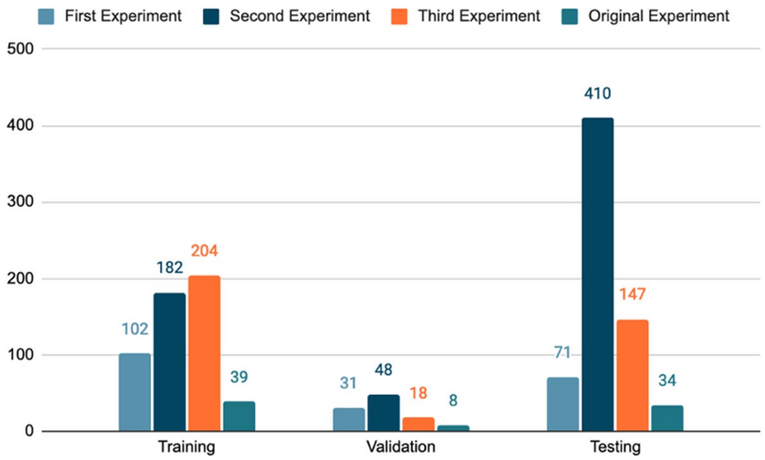


Fig. 10 Comparison of the processing time of the proposed and the original feature extraction

Table 9 Evaluation of the performance in the English experiment

	Split	B-4	M	R	C
1st experiment	Training	87.29	53.47	84.06	192.97
	Validation	50.47	31.26	63.91	60.72
	Testing	46.14	**30.97**	62.16	**59.32**
2nd experiment	Training	81.78	49.03	80.94	168.65
	Validation	49.47	30.30	62.32	64.16
	Testing	46.22	**30.99**	**62.37**	**61.33**
3rd experiment	Training	75.63	44.82	77.43	141.20
	Validation	46.39	30.91	63.49	63.40
	Testing	**47.18**	**30.46**	62.07	**59.98**
Original experiment	Training	77.45	46.03	78.50	150.54
	Validation	50.86	31.45	63.23	66.98
	Testing	46.96	30.33	62.17	58.98

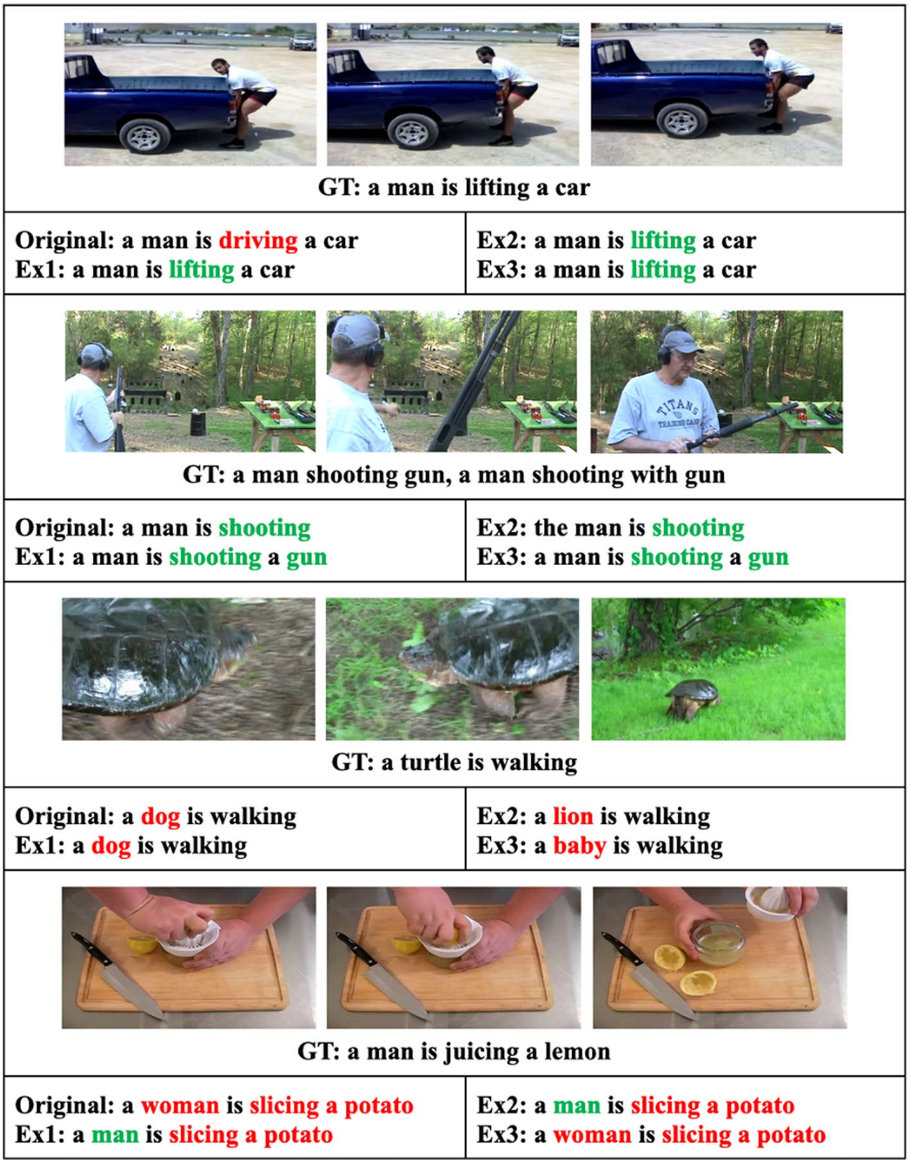
The bold typeface indicates the highest results obtained

Results

This section presents the results of keyframe and feature extractions and video captioning models (English and Arabic). Moreover, a qualitative analysis and major findings are provided.

Keyframe extraction

As previously mentioned, the proposed keyframe extraction approach consists of time- and content-based phases. Time-based keyframe extraction effectively reduces the number of frames extracted. For example, the number of frames extracted at a rate of one frame every quarter of a second in the time-based phase was 43,097; 3,730; and 28,008 in 26, 3, and 25 min for training, validation, and testing, respectively. Similarly, content-based keyframe extraction with a threshold of 95 produced satisfactory frame-reduction results. For example, the number of extracted frames for the



GT: a man is lifting a car

Original: a man is driving a car
Ex1: a man is lifting a car

Ex2: a man is lifting a car
Ex3: a man is lifting a car

GT: a man shooting gun, a man shooting with gun

Original: a man is shooting
Ex1: a man is shooting a gun

Ex2: the man is shooting
Ex3: a man is shooting a gun

GT: a turtle is walking

Original: a dog is walking
Ex1: a dog is walking

Ex2: a lion is walking
Ex3: a baby is walking

GT: a man is juicing a lemon

Original: a woman is slicing a potato
Ex1: a man is slicing a potato

Ex2: a man is slicing a potato
Ex3: a woman is slicing a potato

Fig. 11 Examples of English captions. The green and red fonts illustrate correct and incorrect descriptions, respectively

Table 10 Evaluation of the performance in the Arabic experiment

	Split	B-4	M	R	C
Proposed experiment	Training	54.54	49.35	69.40	1125.85
	Validation	22.36	36.87	46.79	57.08
	Testing	21.65	**36.30**	44.90	**45.52**
Original experiment	Training	51.83	48.06	67.74	117.69
	Validation	24.37	36.91	46.57	59.15
	Testing	22.46	36.16	45.60	43.53

The bold typeface indicates the highest results obtained

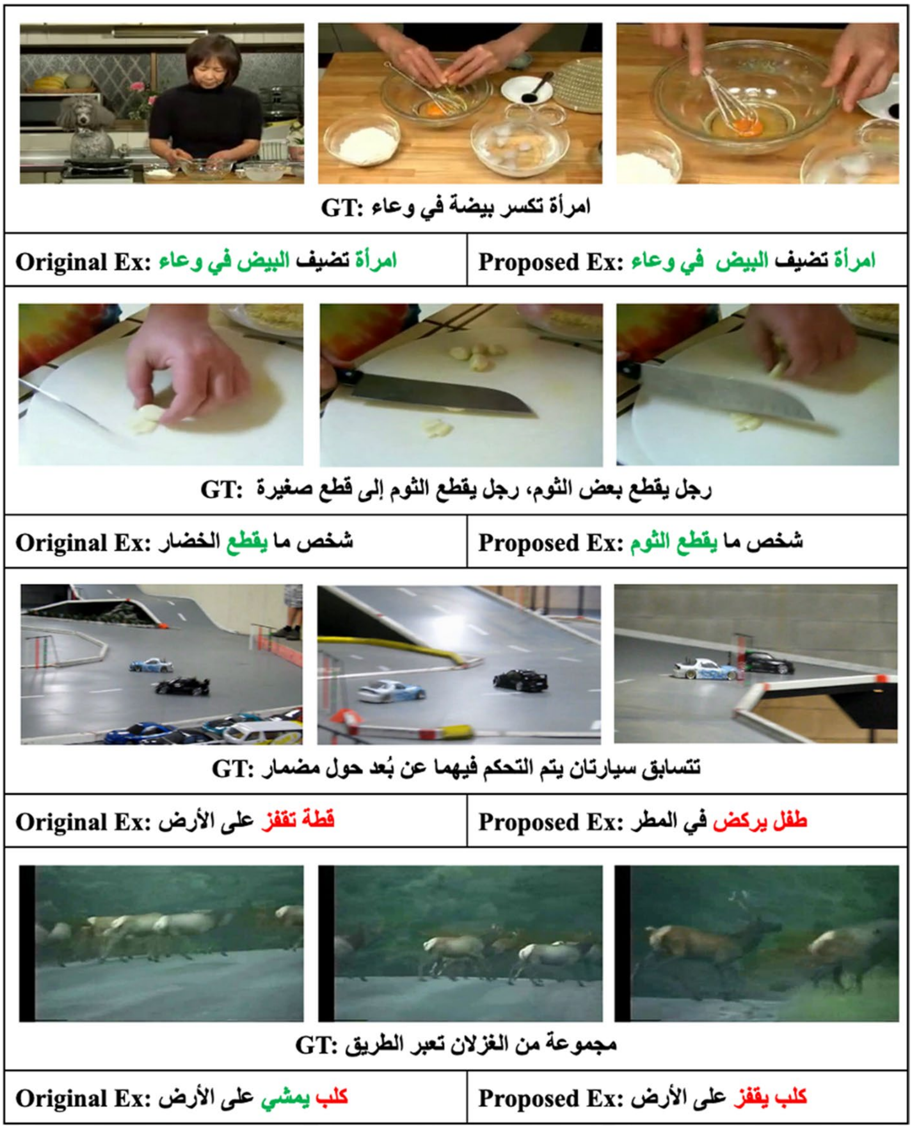

GT: امرأة تكسر بيضة في وعاء
Original Ex: امرأة تضيف البيض في وعاء
Proposed Ex: امرأة تضيف البيض في وعاء
GT: رجل يقطع بعض الثوم، رجل يقطع الثوم إلى قطع صغيرة
Original Ex: شخص ما يقطع الخضار
Proposed Ex: شخص ما يقطع الثوم
GT: تتسابق سيارتان يتم التحكم فيهما عن بُعد حول مضمار
Original Ex: قطة تقفز على الأرض
Proposed Ex: طفل يركض في المطر
GT: مجموعة من الغزلان تعبر الطريق
Original Ex: كلب يمشي على الأرض
Proposed Ex: كلب يقفز على الأرض

Fig. 12 Examples of Arabic captions. The green and red fonts illustrate correct and incorrect descriptions, respectively

training, validation, and testing sets was 41,128; 3597; and 27,013, respectively. However, the processing time was lengthy, approximately 13 h, 41 min, and 15 h for training, validation, and testing, respectively. Table 7 compares the number of keyframes extracted using various thresholds for the original, time-, and content-based keyframe extractions. Table 8 shows the processing times for the original experiment and multiple experiments using the proposed time- and similarity-based keyframe extraction approaches (these experiments are referred as "proposed experiments" later in the text). Original experiment refers to the experiment in which all the frames of the videos were extracted as keyframes.

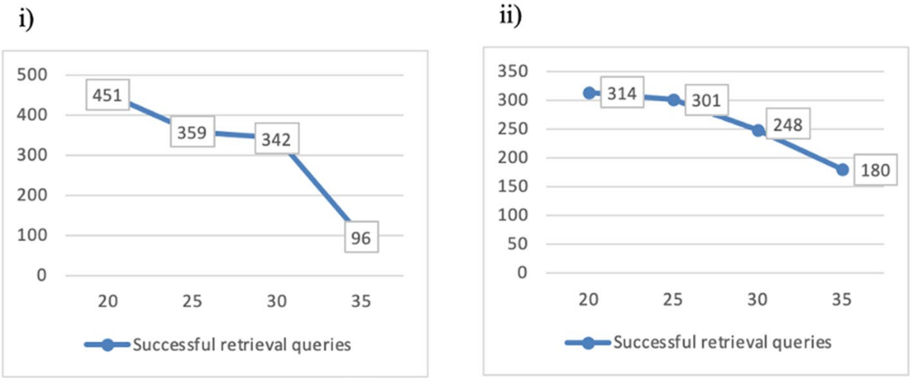
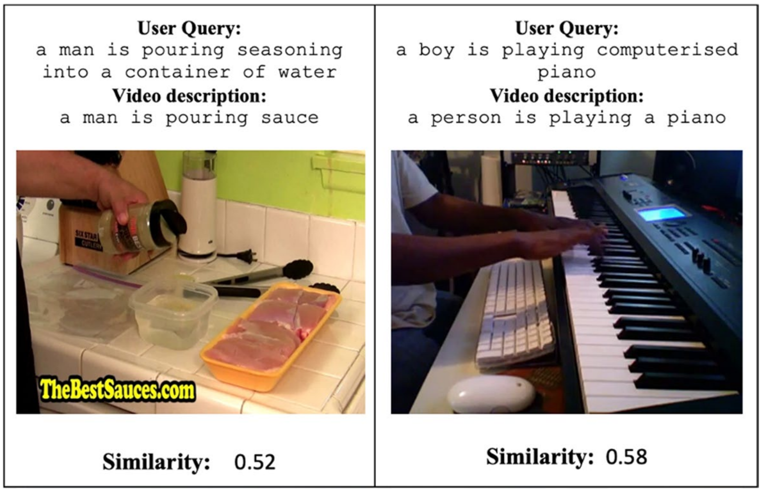


Fig. 13 Performance of English retrieval queries



User Query:
a man is pouring seasoning into a container of water
Video description:
a man is pouring sauce

Similarity: 0.52

User Query:
a boy is playing computerised piano
Video description:
a person is playing a piano

Similarity: 0.58

Fig. 14 Successful retrieval queries in English. The "user query" is the reference caption, while the "video description" is the model-generated caption. The "similarity" is the similarity rate between the two sentences

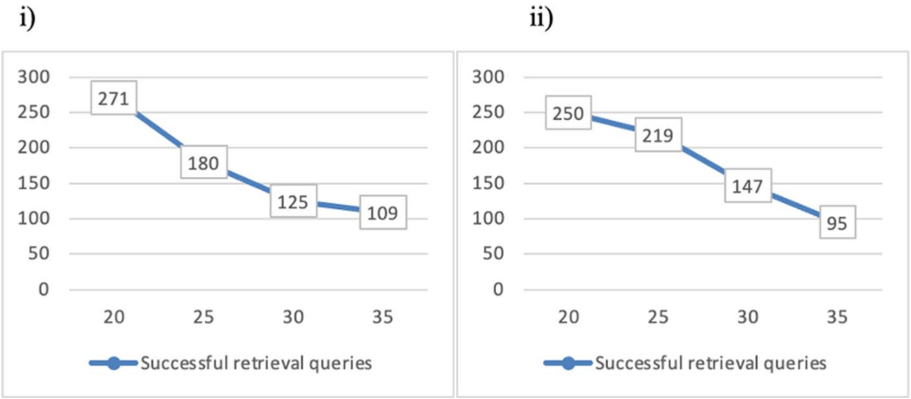


Fig. 15 Performance of English retrieval queries

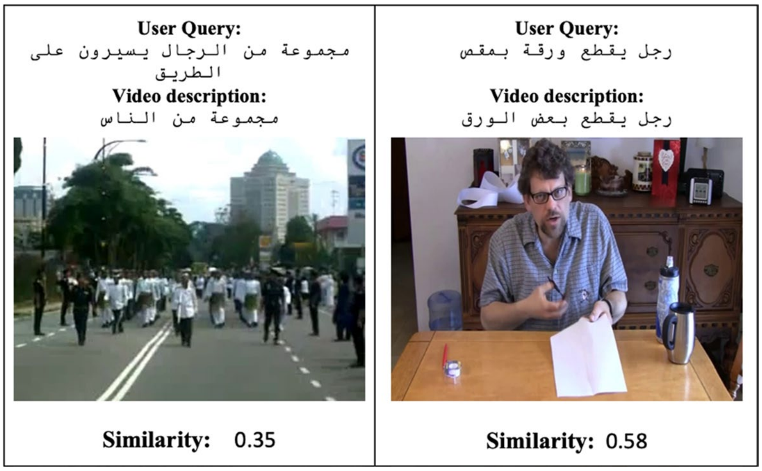


Fig. 16 Arabic successful retrieval queries. The "user query" is the reference caption, while the "video description" is the model-generated caption. The "similarity" means the similarity rate between the two sentences

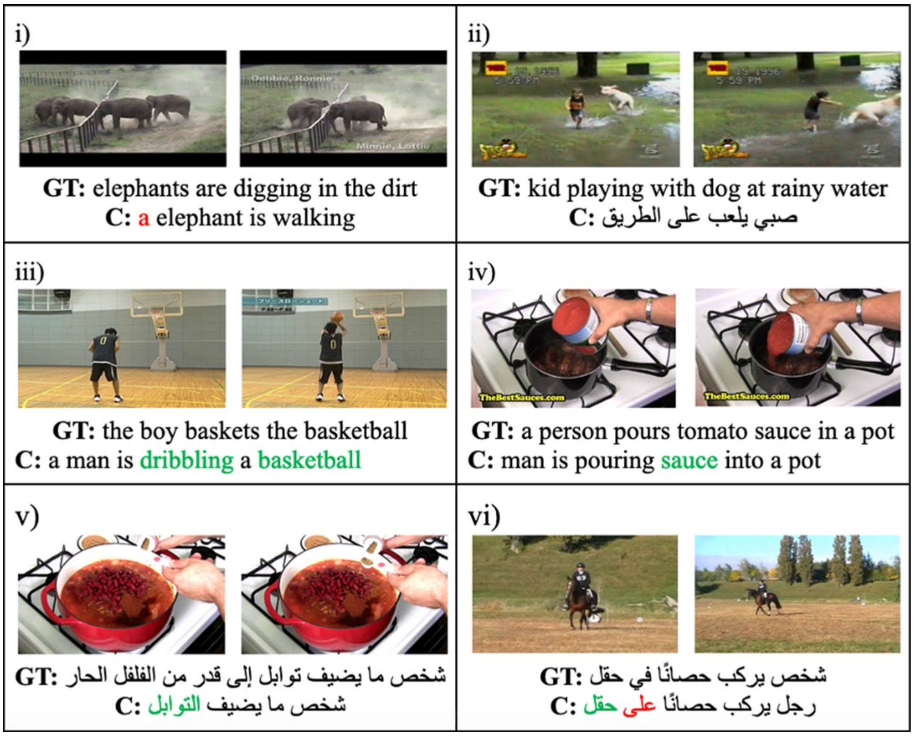
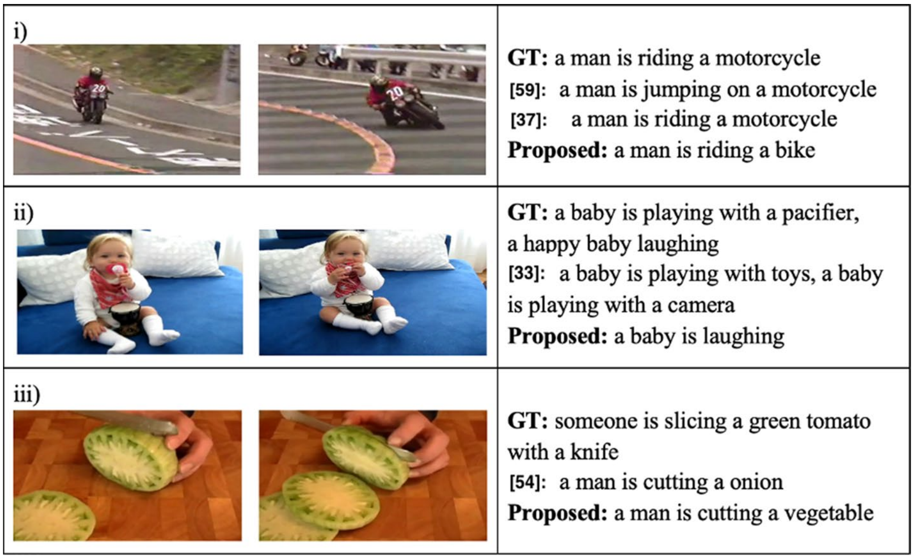

i)
GT: elephants are digging in the dirt
C: a elephant is walking
ii)
GT: kid playing with dog at rainy water
C: صبي يلعب على الطريق
iii)
GT: the boy baskets the basketball
C: a man is dribbling a basketball
iv)
GT: a person pours tomato sauce in a pot
C: man is pouring sauce into a pot
v)
GT: شخص ما يضيف توابل إلى قدر من الفلفل الحار
C: شخص ما يضيف التوابل
vi)
GT: شخص يركب حصانًا في حقل
C: رجل يركب حصانًا على حقل

Fig. 17 Different scenarios for the English and Arabic-generated captions

Feature extraction

Although the number of extracted keyframes decreased because of the two keyframe extraction phases (time- and content-based), the processing time for feature extraction did not decrease. The processing times for feature extraction in the suggested and original experiments are shown in Fig. 10. As shown in the figure, the proposed experiments



GT: a man is riding a motorcycle
[59]: a man is jumping on a motorcycle
[37]: a man is riding a motorcycle
Proposed: a man is riding a bike

GT: a baby is playing with a pacifier, a happy baby laughing
[33]: a baby is playing with toys, a baby is playing with a camera
Proposed: a baby is laughing

GT: someone is slicing a green tomato with a knife
[54]: a man is cutting a onion
Proposed: a man is cutting a vegetable

Fig. 18 Examples of the effectiveness of the proposed model

took longer to complete the feature extraction process compared with the time taken by the original experiment.

Caption evaluation

This section introduces the results of the English and Arabic experiments using machine translation and video retrieval.

Machine translation-based approach

As discussed above, many experiments were conducted using different parameters during the keyframe extraction step. Table 9 compares the results of the proposed experiments with those of the original experiment, which used all the extracted frames. As shown in Table 9, the proposed experiments outperformed the original experiment in terms of at least two metrics. Notably, no experiment achieved better results than the others on all the metrics. Figure 11 shows a few English captions.

According to the results of the English experiments, the best results were achieved in the third experiment in terms of the number of extracted frames, processing time, and caption quality. Therefore, the Arabic experiment was conducted using the same parameters as in the third experiment. Table 10 shows the evaluation results of the proposed and original experiments in terms of all the four metrics. As shown, the proposed method outperformed the original methods for METEOR and CIDE-r. Figure 12 shows examples of captions generated by the Arabic model.

Video retrieval-based approach

The third experiment in the English version included several video retrieval experiments with various thresholds. Additionally, two video retrieval experiments were conducted, one of which removed stop words from the similarity calculation, while the other considered them. The proportion of successful retrieval queries in both experiments is shown in Fig. 13. Figure 14 shows examples of successful retrieval queries.

Similar to the English version, Fig. 15 shows the correct results for the Arabic version's video retrieval queries. Figure 16 shows examples of successful video retrieval.

Qualitative analysis

Video captioning models face the challenge of specifying the main event in the video. Therefore, video captioning models sometimes generate accurate descriptions for the sub-events instead of the main event. For example, in Fig. 17, example (i), the proposed model described the video content as "elephant is walking." This description is correct, but unlike most ground truths, the model did not describe the main event, "digging in the dirt" or "kicking up dust." Similarly, example (ii) in Fig. 17 shows a dog; however, the Arabic model described the video as "a kid playing on the street," "صبي يلعب على الطريق" and completely ignored the dog.

However, the English captions of examples (iii) and (iv), which are "a man is dribbling a basketball" and "man is pouring sauce into a pot," accurately provide specific descriptions. In example (iii), the model described the event using a specialized sports term, "dribbling," instead of a typical one, such as "playing." Similarly, in example (iv), the model described the object as "sauce" rather than a more typical term, such as "a liquid thing." Furthermore, the Arabic model was successful in accurately describing some details, such as "التوابل" in example (v) and "الحقل" in example (vi).

Captions by the proposed model and ones from several existing studies are compared in Fig. 18. As shown in example (i), the caption of the proposed model accurately reflects the ground truth and the caption generated in [40], uses the GAN technique, and outperforms the caption generated using the LSTM model [62]. Furthermore, the caption in [35] mentioned a "camera" and "toy," which do not appear in the video. Example (ii) shows how the proposed model successfully produced a suitable caption. The third example (iii) shows a "green tomato," i.e., an unripe tomato. However, the proposed model did not identify the unripe tomato and instead described it as a "vegetable," contrary to the description generated by the model in [57], which described it as an "onion."

Major findings

The study findings can be summarized as follows:

- The proposed keyframe extraction method showed positive results in terms of METEOR, as shown in Table 9, which shows that all the proposed experiments outperformed the baseline. The authors attribute this performance to a reduction in the number of frames per event, reducing the bias for specific terms.
- While content-based keyframe extraction requires more time, it reduces storage costs while maintaining caption quality. Users should weigh the trade-off between time and cost when deciding to adopt this model.
- The model performance varied among videos; some videos were well described (as shown in Fig. 11). This result can be attribute to some subjects, actions, and objects

frequently appearing in the training set, thus allowing the model to accurately describe them, while data that appeared less frequently resulted in lower performance. For example, the dataset had strong coverage of cooking-related actions such as cutting and slicing as well as cooking-related objects such as pans, pots, vegetables, and meat, leading to improved performance in these types of videos. Similarly, actions like "horse riding" and musical performances on instruments like the piano, flute, or guitar were well covered and described.

- Several studies have concluded that video captioning is suitable for video retrieval, and this study supports this conclusion by using both Arabic and English (refer to "Video retrieval-based approach" section).

Conclusion and future work

To improve video retrieval, a video captioning model capable of producing captions in Arabic and English was proposed in this study. The model achieved high accuracy using only a few keyframes. Both the English and Arabic models were suitable for video retrieval, with success rates of 67% and 40%, respectively. This model has potential applications in various fields, such as automating sports commentaries, converting movies into written works, generating security reports, and assisting visually impaired individuals to understand visual content. Despite its potential benefits, this study faced several challenges, such as the use of small datasets, the absence of preprocessed Arabic datasets, and the need for high computational resources. Future research needs to explore different parameters, such as batch size, learning rate, and number of encoder and decoder tokens, and use a variant size of the feature array using the video length to address these challenges. Additionally, the authors suggest that researchers can incorporate additional features, such as written text or audio, in videos to improve video captioning models, which might improve performance. Moreover, enlarging existing datasets or creating new ones can enhance video captioning models. Finally, the authors emphasize the need for concerted global research efforts to develop video captioning models for languages other than English.

Abbreviations

BLEU	Bilingual evaluation understudy
CIDE-r	Consensus-based image description evaluation
CV	Computer vision
DNN	Deep neural network
FPS	Frames per second
GAN	Generative adversarial network
GRU	Gated recurrent unit
LSTM	Long short-term memory
METEOR	Metric for evaluation of translation with explicit ordering
NLP	Natural language processing
P	Place
RL	Reinforcement learning
ROUGE-L	Recall-oriented understudy of gisting evaluation
SSIM	Structural similarity index measurement
SMAN	Stacked multimodal attention network
SVO	Subject-verb-object

Acknowledgements

The authors gratefully acknowledge Qassim University, represented by the Deanship of "Scientific Research, on the financial support for this research under the number (COC-2022-12-J-30388) during the academic year 1444 AH/2022AD".

Author contributions

NA: conceptualization, methodology, experimentations, validation, writing—original draft; AAA-S: conceptualization, supervision, review and editing. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The code and experimental Arabic data are available at https://github.com/NorahAbdulaziz/VideoCaptionModel/tree/main and https://kaggle.com/datasets/4b6d062755848ce3ad3fd74ca875dfbd12d7744b7e131cf45e7d14470be0a2bb, respectively.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 8 April 2023 Accepted: 4 January 2024
Published online: 16 January 2024

References

1. Ramesh A et al. Zero-shot text-to-image generation. In: International conference on machine learning; 2021.
2. OrCam MyEye 2. 0—for people who are blind or visually impaired. https://www.orcam.com/en/myeye2/. Accessed 20 Nov 2022.
3. Bebis G, Egbert D, Member S, Shah M. Review of computer vision education. IEEE Trans Educ. 2003;46:1–20.
4. Wiley V, Lucas T. Computer vision and image processing: a paper review. Int J Artif Intell Res. 2018;2:29–36.
5. Hirschberg J, Manning CD. Advances in natural language processing. Science (80−). 2015;349:261–6.
6. Nabati M, Behrad A. Multimodal video-text matching using a deep bifurcation network and joint embedding of visual and textual features. Expert Syst Appl. 2021;184: 115541.
7. Du XY, et al. Captioning videos using large-scale image corpus. J Comput Sci Technol. 2017;32:480–93.
8. Aggarwal A, et al. Video caption based searching using end-to-end dense captioning and sentence embeddings. Symmetry. 2020;2020(12): 992.
9. Hale J. More than 500 hours of content are now being uploaded to you tube every minute—tubefilter. 2019. https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/. Accessed 05 Nov 2022.
10. Paul MKA, Kavitha J, Rani PAJ. Key-frame extr techniques. Rev Recent Pat Comput Sci. 2018;1:3–16.
11. Meena P, Kumar H, Yadav SK. A review on video summarization techniques. Eng Appl Artif Intell. 2023;118: 105667.
12. Video D. What is frame. Frame in the world of animated video. https://darvideo.tv/dictionary/frame/. Accessed 5 Apr 2022.
13. Dong Y, Zhang Y, Zhang J, Zhang X, Zhang CY. Video key frame extraction based on scale and direction analysis. J Eng. 2022. https://doi.org/10.1016/j.cmpb.2019.105236.
14. Tang H et al. Deep unsupervised key frame extraction for efficient video classification. arXiv. 2022;1–16.
15. Savran Kızıltepe R, Gan JQ, Escobar JJ. A novel keyframe extraction method for video classification using deep neural networks. Neural Comput Appl. 2021;35:1–12.
16. Rafiq M, Rafiq G, Choi GS. Video description: datasets & evaluation metrics. IEEE Access. 2021;9:121665–85.
17. Xu R, Xiong C, Chen W, Corso JJ. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence. 2015. p. 2346–52.
18. Yao L et al. Describing videos by exploiting temporal structure. In: Proceedings of the IEEE international conference on computer vision. 2015. p. 4507–15.
19. Qian T, Mei X, Xu P, Ge K, Qiu Z. Filtration network: a frame sampling strategy via deep reinforcement learning for video captioning. J Intell Fuzzy Syst. 2021;40:11085–97.
20. Chen K, et al. A video key frame extraction method based on multiview fusion. Mob Inf Syst. 2022. https://doi.org/10.1155/2022/8931035.
21. Elahi GMME, Yang YH. Online learnable keyframe extraction in videos and its application with semantic word vector in action recognition. Pattern Recognit. 2022;122:108273.
22. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770–8.
23. Nandini HM, Chethan HK, Rashmi BS. Shot based keyframe extraction using edge-LBP approach. J King Saud Univ Comput Inf Sci. 2022;34:4537–45.
24. Danielsson PE. Euclidean distance mapping. Comput Graph Image Process. 1980;14:227–48.
25. Sobel I, Feldman GM. A 3x3 isotropic gradient operator for image processing. In: The Stanford artificial intelligence project. 1968. p. 1–6.
26. Chakraborty S, Thounaojam DM. SBD-Duo: a dual stage shot boundary detection technique robust to motion and illumination effect. Multimed Tools Appl. 2021;80:3071–87.

27. Klein D, Manning CD. Accurate unlexicalized parsing. In: Proceedings of the 41st annual meeting of the association for computational linguistics. 2003. p. 423–30.
28. Hanckmann P, Schutte K, Burghouts GJ. Automated textual descriptions for a wide range of video events with 48 human actions. In: Lect Notes Comput Sci. 2012. p. 372–80.
29. Barbu A et al. Video in sentences out. arXiv. 2012. p. 1–13.
30. Thomason J, Venugopalan S, Guadarrama S, Saenko K, Mooney R. Integrating language and vision to generate natural language descriptions of videos in the wild. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics; 2014. p. 1218–27.
31. Krishnamoorthy N, Malkarnenkar G, Mooney R, Saenko K, Guadarrama S. Generating natural-language video descriptions using text-mined knowledge. In: Proceedings of the Twenty-seventh AAAI conference on artificial intelligence; 2013. p. 541–7.
32. Peng Y, Wang C, Pei Y, Li Y. Video captioning with global and local text attention. Int J Comput Graph. 2021;38:1–12.
33. Liu S, Ren Z, Yuan J. SibNet: sibling convolutional encoder for video captioning. IEEE Trans Pattern Anal Mach Intell. 2021;43:3259–72.
34. Lee S, Kim I. DVC-Net. A deep neural network model for dense video captioning. IET Comput Vis. 2021;15:12–23.
35. Naik D, Jaidhar CD. Semantic context driven language descriptions of videos using deep neural network. J Big Data. 2022;9(17):1–22.
36. Deb T et al. Variational stacked local attention networks for diverse video captioning. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV). 2022. p. 4070–79.
37. Li Q, Yang L, Tang P, Wang H. Enhancing semantics with multi-objective reinforcement learning for video description. Electron Lett. 2021;57:977–9.
38. Zheng Y, Zhang Y, Feng R, Zhang T, Fan W. Stacked multimodal attention network for context-aware video captioning. IEEE Trans Circuits Syst Video Technol. 2022;32:31–42.
39. Creswell A, et al. Generative adversarial networks: an overview. IEEE Signal Process Mag. 2018;35:53–65.
40. Yang Y, et al. Video captioning by adversarial LSTM. IEEE Trans Image Process. 2018;27:5600–11.
41. Chen DL, Dolan WB. Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. 2011. p. 190–200.
42. Vaidya J, Subramaniam A, Mittal A. Co-segmentation aided two-stream architecture for video captioning. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022. p. 2774–84.
43. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004;13:600–12.
44. Lab SV, ImageNet., Lab. Stanford vision. https://www.image-net.org/. Accessed 21 Nov 2022.
45. Shreya. Video-captioning: video captioning is an encoder decoder mode based on sequence to sequence learning. github. 2020. https://github.com/Shreyz-max/Video-Captioning. Accessed 05 Jan 2022.
46. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9:1735–80.
47. Zhang A, Lipton ZC, Li M, Smola AJ. Dive into deep learning. Cambridge: Cambridge University Press; 2022.
48. Xu J, Wei H, Li L, Guo J. Video description model based on temporal–spatial and channel multi-attention mechanisms. Appl Sci. 2020;10:4312.
49. Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002. p. 311–8.
50. Lavie A, Sagae K, Jayaraman S. The significance of recall in automatic metrics for MT evaluation. Lect Notes Comput Sci. 2004;3265:134–43.
51. Lin C-Y. ROUGE: a package for automatic evaluation of summaries. In: Text summarization branches out. 2004. p. 74–81.
52. Vedantam R, Zitnick CL, Parikh D, CIDEr. Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2015. p. 4566–75.
53. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manag. 1988;24:513–23.
54. Google translate. https://translate.google.co.in. Accessed 1 Jan 2023.
55. MyMemory—Machine translation meets human translation. https://mymemory.translated.net. Accessed 01 Jan 2023.
56. Bing microsoft translator. https://www.bing.com/translator. Accessed 01 Jan 2023.
57. Singh A, Singh TD, Bandyopadhyay S. Attention based video captioning framework for Hindi. Multimed Syst. 2021;28(1):195–207.
58. Qi S, Yang L. Video captioning via a symmetric bidirectional decoder. IET Comput Vis. 2021;15:283–96.
59. Ye H et al. Hierarchical modular network for video captioning. in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2022. p. 17939–948.
60. Kim H, Lee SA. Video Captioning method based on multi-representation switching for sustainable computing. Sustainability. 2021;13:2250.
61. Kingma DP, Ba JL, Adam. A method for stochastic optimization. In: 3rd Int. Conf. Learn. Represent. ICLR 2015—Conf. Track Proc. 2014.
62. Venugopalan S et al. Translating videos to natural language using deep recurrent neural networks. arXiv. 2015. p. 1–18.

Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.