ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 8, August 2023

Aerial Video Retrieval with Text Cues

[1] Shima M. Al Mehmadi, [2] Laila Bashmal, [3] Yakoub Bazi, [4] Mohamad M. Al Rahhal

[1][2][3][4] King Saud University, Riyadh, Saudi Arabia
E-mail: [1] 441203085@student.ksu.edu.sa, [2] 439204359@student.ksu.edu.sa, [3] ybazi@ksu.edu.sa,
[4] mmalrahhal@ksu.edu.sa

Abstract*— In today's digital era, multimedia content such as images, videos, text, and audio are commonplace. With the increase in the number of Unmanned Aerial Vehicles (UAVs) in the sky, UAV videos have emerged as a new form of communication. To efficiently and effectively search for a specific video from a large dataset, text-to-video retrieval is recommended. In this paper, we present a text-to-event retrieval model for UAV videos. The model comprises two parts: the first part extracts frame-level features from the video using Vision Transformer (ViT), and the second part extracts textual representations from the query using Bidirectional Encoder Representations from Transformers (BERT). Both parts are jointly trained on text-video pairs using bidirectional contrastive loss. The effectiveness of the proposed method was evaluated on the CapERA dataset, an extended version of the event recognition in aerial video (ERA) dataset, and the results demonstrate its efficacy.

***Index Terms*— Unmanned Aerial Vehicles, Vision Transformer, Contrastive Loss**

I. INTRODUCTION

Because of the great development in remote sensing (RS) technology, the number of high spatial resolution images increases day after day. This has resulted in an abundance of information that introduces new challenges in the study of RS images. Furthermore, as RS devices and technologies have advanced, many applications of RS images have made substantial progress, including scene classification [1], object detection [2], image retrieval [3], image captioning [4], and semantic segmentation [5].

Data in the RS technique will be collected using sensor technology based on satellites or UAVs. Satellite-based technologies cover a broader geographical area than UAV-based technologies. Consequently, the data volume of RS is continuously increasing depending on the enormous number of launched earth observation satellites. One of the most well-known satellites is Jilin-1 [6], a Chinese RS satellite sensor. Unlike satellites, UAVs can provide real-time allowing for quick decision-making and high-resolution video at a very low cost. In addition, UAVs can dramatically reduce reliance on weather conditions, such as clouds, dust and fog. And they are providing greater flexibility to deal with a variety of problems. For instance, the ERA dataset [7] is a new aerial video dataset that covers a wide range of events. Besides that, each event was represented by a record of 5 seconds, for a total of 2,864 video clips.

In this work, we are concerned to propose a novel architecture for the text-to-video retrieval task in the RS field. The idea of retrieving RS data from a given text has been investigated in the context of images only. Zhang *et al.* [8] proposed a feature decoupling and reconstruction method using a vision transformer to solve the issue of redundant information that resulted from straightforward mapping in state-of-the-art models. First, image and text features are extracted using a ViT and BERT, respectively. Then, the resulting features are decoupled into modal invariant and modal heterogeneous features. And at the end, these features will be reconstructed back. After the features of text and RS images are extracted in [9], graph modules are applied to induce the interactive fusion of text and RS image features. In addition, the model highlights the components related to the query by using the Text-Image Association Module. Lv *et al.* [10] presented a fusion-based correlation learning model with two stages: modality-specific feature learning and common feature space learning. In the first stage, a deep CNN was used to extract distinct image and text features. Where in the second stage, a common space is constructed to compare the heterogeneous data with consecutive loss functions for realizing the semantic correlation between image and text pairs.

Due to the scarcity, or rather absence, of literatures on RS text-to-video retrieval models, inspiration has come from works in the field of computer vision. Recently, in [11] the input video is represented with Reading-strategy-Inspired Visual Representation Learning (RIVRL). It is divided into two parts: a previewing part and an intensive-reading part. The first part is responsible for capturing general information about videos. While the second part is intended to get more detailed information about videos and it is aware of the resultant data in the previewing part. Feng *et al.* [12] proposed a Temporal Multi-modal Graph Transformer with Global-Local Alignment (TMMGT-GLA). The visual input is represented as a series of semantic correlation graphs to utilize the structural information between multi-modal features. In addition, graph and temporal self-attention layers are used to learn cross-modal relations and temporal associations in an effective way.

Retrieving a video from a given textual query for RS aerial videos has not been investigated yet. This is a crucial issue since most users prefer to convey their information needs using natural language. As a result, this paper presents a model for retrieving RS videos. Which is one of the most important multi-modal learning challenges that tries to find the most relevant video for a given textual query. The model includes two parts. The first part is the video encoder, which uses the vision transformer (ViT) to extract frame-level information from video. The second part is the text encoder, that employs Bidirectional Encoder Representations from Transformers (BERT) to extract textual representations from the query. The two parts are trained jointly on video-text pairs by minimizing a bidirectional contrastive loss.

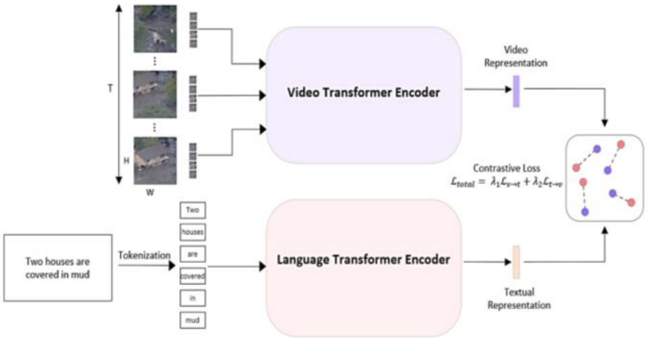
T
H
W
Video Transformer Encoder
Video Representation
Contrastive Loss
$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{v\to t} + \lambda_2 \mathcal{L}_{t\to v}$
Two houses are covered in mud
Tokenization
Two
houses
are
covered
in
mud
Language Transformer Encoder
Textual Representation

Fig. 1. The general flow chart of the proposed text-video retrieval method.

II. PROPOSED METHOD

This section will explain the proposed method of the text-to-video retrieval task in detail. The backbone of our methodology is the vision transformer, and the model consists of two parts: a language transformer encoder and a video transformer encoder, as shown in Fig. 1. In the beginning, we supposed to have a set of R video-text pairs as $D = \{X_i, c_i\}_{i=1}^{R}$, where X_i is the video of size $X_i \in \mathbb{R}^{T\times H\times W\times C}$ and c_i is the related textual caption. The following is a detailed explanation of the proposed model.

A. Language Transformer Encoder

The mechanism of the language transformer encoder begins with taking the sentence input as word tokens $c_i = (w_1, w_2, w_3, \cdots, w_n)$, where n is the length of the sentence. Then, a learnable embedding layer E_c was applied to convert the "word vector" into a series of textual features with dimension d_c. And by using a learnable positional embedding layer, the output textual features series has included additional information about the position of each word. Moreover, two special tokens, CLS and SEP, are added at the beginning and end of the word series. Therefore, the sentence is represented as:

$$g_{c0} = [w_{class};\ w_1E_c;\ w_2E_c;\ \cdots;\ w_nE_c] +\ E_{pos} \qquad (1)$$

w_{class} is the general token and $E_{pos} \in \mathbb{R}^{(n+1)\times d_c}$ is the positional embedding information. Then, the initial g_{c0} is passing through multiple identical layers of the language transformer encoder produce the last representation g_{cL} at the latest layer L. Each encoder's layer begins with a multi-head self-attention (MSA) block and ends with a multi-layer perceptron (MLP) block with GELU activation function in between. These two blocks are joined by residual skip connection and follow up by a normalization layer (LN):

$$g'_l\ = MSA\left(LN(g_{l-1})\right) + g_{l-1},\ l = 1,2,\cdots,L \qquad (2)$$

$$g_l\ = MLP\left(LN(g'_l)\right) + g'_l, \qquad l = 1,2,\cdots,L \qquad (3)$$

Then the output textual features are normalized by L2-normlization technique as $\{f_{ci}\}_{i=1}^{b}$.

B. Video Transformer Encoder

In video encoder architecture, the input video is splitting into a sequence of f nonoverlapping video frames $(x_p^1;\ x_p^2;\ \cdots; x_p^f)$. Each frame has processed as single image with its own image encoder [13] and has dimension of $(3p^2)$ where p is the height or width of the frame and f represents the total number of frames $f = (224 \times 224)/p^2$. The resulting frame sequence is then projected and flattened by a linear projection layer E_v to dimension d_v. A positional embedding layer is applied to include the positional information, as done in the language transformer encoder. After that, the video frame sequence supplies the transformer encoder as:

$$g_{v0} = [x_{class};\ x_p^1\ E_v;\ x_p^2\ E_v;\ \cdots;\ x_p^f\ E_v] +\ E_{pos} \qquad (4)$$

here the linear embedding layer is $E_v \in \mathbb{R}^{(p^2.\ c)\times d_v}$, and the positional embedding layer is $E_{pos} \in \mathbb{R}^{(f+1)\times d_v}$. Also, x_{class} represents the patch representations. The last video representation g_{vL} is obtained by utilizing equations (2) and (3). Also, the output video frame features are normalized by L2-normlization as $\{f_{vi}\}_{i=1}^{b}$. Furthermore, contrary to the language transformer encoder, normalization occurs before MSA and MLP blocks in the video transformer encoder.

C. Contrastive Loss

To improve the model's performance, we utilized contrastive loss, which has proven to be effective in measuring pairwise similarity. We computed contrastive loss in both text and video spaces. In the text space, we aimed to ensure that the visual features are very close to their corresponding text features. Similarly, in the video space, we aimed to ensure that the text features are very close to their corresponding visual features.

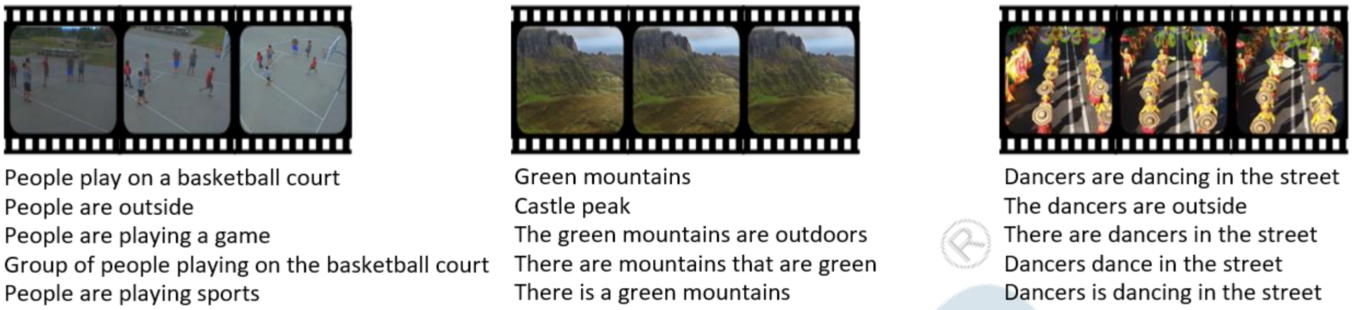
People play on a basketball court
People are outside
People are playing a game
Group of people playing on the basketball court
People are playing sports
Green mountains
Castle peak
The green mountains are outdoors
There are mountains that are green
There is a green mountains
Dancers are dancing in the street
The dancers are outside
There are dancers in the street
Dancers dance in the street
Dancers is dancing in the street

Fig.2. Sample video clips from CapERA dataset with their captions.

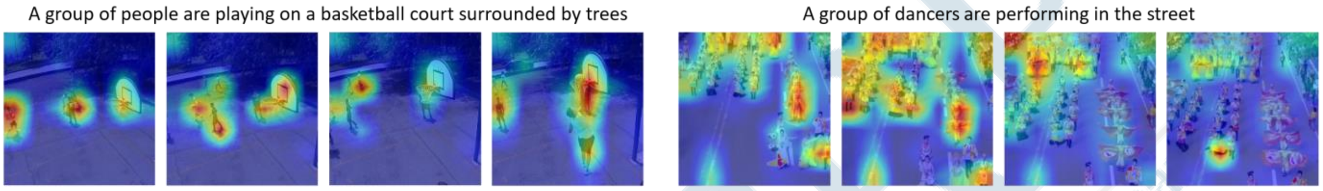
A group of people are playing on a basketball court surrounded by trees
A group of dancers are performing in the street

Fig. 3. Attentions over text and video for different samples

III. EXPERIMENTS RESULTS

A. Dataset Description

The aerial video dataset ERA [7] includes an extensive range of events. Moreover, it captures dynamic events in a variety of situations with significantly different scales. It is a collection of 2,864 videos divided between 1,473 training set and 1,391 test set, and each video is clipped from a long YouTube UAV video to 5 seconds with 24 frames per second and a spatial size of 640×640. Researchers of ERA dataset have started by making a general taxonomy of 24 most commonly seen events in aerial scenes. Also, they set up additional category called "non-event" to ensure that models can distinguish between events and normal videos. In this paper, we utilized an upgraded version of the ERA dataset called CapERA [14]. It describes each video by five descriptions. The first description of each video is manually annotated, while the rest are automatically produced by paraphrasing and translation tools. In Fig. 2, we show an example of three videos with their five captions.

B. Experimental Settings

Our text-to-video retrieval model was trained on the contrastive language image pre-training (CLIP) model [15] with over 400 million text-image pairs. Furthermore, we adopt BERT with 12 layers as a language transformer. The size of the vocabulary is equal to 49,408, and to embed the sequence into the features of dimension $d_c = 512$ a word embedding layer is utilized. The Bert-base-uncased tokenizer is used with a fixed length equal to $n = 77$. While ViT32 with 12 layers is used as a video transformer. It splits the image into $f = 49$ patches of 32×32 pixels each and flattened to the dimension $d_v = 768$. Moreover, we choose one of the five captions for each video at random during training. In addition, we set the learning rate to $3e - 4$ to optimize the model and use the Adam optimizer. We trained the model for 50 epochs and a mini-batch of size 10.

C. Evaluation Metrics

The results in this paper are evaluated by the Recall@k (R@k) metric. It is the fraction of the relevant items that are successfully retrieved with a given textual query, and it is expressed as:

$$R@K = (TP@k)/(TP@k + FN@k) \qquad (5)$$

The TP@k represents the true positive condition, while the FN@k is the false negative. To evaluate the performance, we use three different values of k (1,5,10) with the R@k indicator.

D. Results

The outcomes presented in Table 1 are categorized into two retrieval methods: text-to-video and video-to-text. Furthermore, the retrieval performance of the video encoder is examined by using varying numbers of sampled frames. It can be observed that the R@k scores for video-to-text retrievals are slightly lower than text-to-video retrievals, as each video in the CapERA dataset has five relevant captions, while each text sentence has only one relevant video. Moreover, it is evident that the R@k score outcomes are relatively high when using eight sampled frames for both text-to-video and video-to-text retrieval. However, when the number of sampled frames is increased to 16, the results decrease across all metrics.

Fig. 3 displays the attention maps generated by our model for two video frames retrieved from two distinct textual queries. The attention maps assist us in identifying the specific spatiotemporal regions of focus. The attention maps on the left indicate that the model emphasizes the people and basket mentioned in the query, indicating that the model should concentrate on fine-grained temporal details. On the other hand, the attention maps on the right highlight the dancers more prominently. Despite the differences, the attention maps reveal that the proposed model is adept at accurately learning deep spatiotemporal features for UAV video retrieval.

Table 1. Retrieval Results on Era Dataset

	CapERA Dataset					
	Text-to-video Retrieval			Video-to-text Retrieval		
Recall@k No. of frames	R@1	R@5	R@10	R@1	R@5	R@10
4 frames	8.92	27.79	41.59	7.83	23.36	34.29
8 frames	**9.20**	**28.09**	**42.49**	**8.91**	**25.23**	**35.37**
16 frames	7.40	26.47	39.57	7.33	20.85	31.78

IV. CONCLUSION

In conclusion, we proposed a self-attention model of text-to-video retrieval for aerial videos. In addition, the enhancement of text and video representations has been done by improving the bidirectional contrastive loss. We can clearly see the effectiveness of our model from the quantitative results on the CapERA dataset.

REFERENCES

[1] X. Lu, X. Zheng, and Y. Yuan, “Remote Sensing Scene Classification by Unsupervised Representation Learning,” *IEEE Trans. Geosci. Remote Sensing*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017, doi: 10.1109/TGRS.2017.2702596.

[2] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, “Object Detection in Optical Remote Sensing Images Based on Weakly Supervised Learning and High-Level Feature Learning,” *IEEE Trans. Geosci. Remote Sensing*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015, doi: 10.1109/TGRS.2014.2374218.

[3] T. Abdullah, Y. Bazi, M. M. Al Rahhal, M. L. Mekhalfi, L. Rangarajan, and M. Zuair, “TextRS: Deep Bidirectional Triplet Network for Matching Text to Remote Sensing Images,” *Remote Sensing*, vol. 12, no. 3, p. 405, Jan. 2020, doi: 10.3390/rs12030405.

[4] G. Hoxha and F. Melgani, “A Novel SVM-Based Decoder for Remote Sensing Image Captioning,” *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–14, 2022, doi: 10.1109/TGRS.2021.3105004.

[5] B. Qu, X. Li, D. Tao, and X. Lu, “Deep semantic understanding of high resolution remote sensing image,” in *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*, Kunming, China: IEEE, Jul. 2016, pp. 1–5. doi: 10.1109/CITS.2016.7546397.

[6] “Jilin-1 Satellite Sensor | Satellite Imaging Corp.” https://www.satimagingcorp.com/satellite-sensors/jilin-1-satellite-sensor-1m/ (accessed Mar. 08, 2022).

[7] L. Mou, Y. Hua, P. Jin, and X. X. Zhu, “ERA: A Data Set and Deep Learning Benchmark for Event Recognition in Aerial Videos [Software and Data Sets],” *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 125–133, Dec. 2020, doi: 10.1109/MGRS.2020.3005751.

[8] H. Zhang *et al.*, “A Transformer-Based Cross-Modal Image-Text Retrieval Method using Feature Decoupling and Reconstruction,” in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, Kuala Lumpur, Malaysia: IEEE, Jul. 2022, pp. 1796–1799. doi: 10.1109/IGARSS46834.2022.9883242.

[9] F. Yao, N. Liu, P. Li, D. Yin, C. Liu, and X. Sun, “Cross-Modal Remote Sensing Image Retrieval Via Intra- and Inter-Modal Feature Matching,” in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, Kuala Lumpur, Malaysia: IEEE, Jul. 2022, pp. 1792–1795. doi: 10.1109/IGARSS46834.2022.9884328.

[10] Y. Lv, W. Xiong, X. Zhang, and Y. Cui, “Fusion-Based Correlation Learning Model for Cross-Modal Remote Sensing Image Retrieval,” *IEEE Geosci. Remote Sensing Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3131592.

[11] J. Dong *et al.*, “Reading-Strategy Inspired Visual Representation Learning for Text-to-Video Retrieval,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5680–5694, Aug. 2022, doi: 10.1109/TCSVT.2022.3150959.

[12] Z. Feng, Z. Zeng, C. Guo, and Z. Li, “Temporal Multimodal Graph Transformer With Global-Local Alignment for Video-Text Retrieval,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1438–1453, Mar. 2023, doi: 10.1109/TCSVT.2022.3207910.

[13] M. M. A. Rahhal, Y. Bazi, N. A. Alsharif, L. Bashmal, N. Alajlan, and F. Melgani, “Multilanguage Transformer for Improved Text to Remote Sensing Image Retrieval,” *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol. 15, pp. 9115–9126, 2022, doi: 10.1109/JSTARS.2022.3215803.

[14] L. Bashmal, Y. Bazi, M. M. Al Rahhal, M. Zuair, and F. Melgani, “CapERA: Captioning Events in Aerial Videos,” *Remote Sensing*, vol. 15, no. 8, p. 2139, Apr. 2023, doi: 10.3390/rs15082139.

[15] A. Radford *et al.*, “Learning Transferable Visual Models From Natural Language Supervision.” arXiv, Feb. 26, 2021. Accessed: Feb. 19, 2023. [Online]. Available: http://arxiv.org/abs/2103.00020