academic publishers
INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE (ISSN: 2692-5206)
Volume 04, Issue 07, 2024
Published Date: 10-09-2024

WORKING WITH VIDEOS

PhD, Azizbek Ruzmetov
Kimyo International University in Tashkent
Yetmishboyev Shakhzodbek
Master's student of Kimyo International University in Tashkent

Abstract: This article presents a number of ways to work with video images, such as choosing a suitable title, subtitle and object. Also, within the framework of this topic, researches of many researchers are studied, suggestions and recommendations are given to users.
Keywords: Video image, Artificial intelligence, subtitle, object, graphic.

INTRODUCTION

More and more users are joining video editing every day. For some, this process becomes a fun activity, and for some users, it becomes a way to earn money. The video editor provides users with all the necessary features for video editing. At the same time, the interface and functionality of the video editor will attract not only professionals, but also users who understand the basics of video editing[1]. Image restoration techniques have been widely used in various fields, including computer science, medicine [2], remote sensing, and security monitoring [3]. To date, artificial intelligence and graphics automatically describe the visual data of a video, this person cannot but be amazed. An image caption automatically generates a natural language description of the image[4]. This is a complex task that involves both image comprehension and language creation. In addition, the purpose of subtitling an image is to create a sentence that correctly expresses the content of the image and is grammatically correct[5].

Video captioning, also known as video image storytelling, creates natural language descriptions of events and actions in a video. This task is similar to captioning an image, but more complex, as it requires understanding the temporal dynamics of visual content and events. Recently, in order to make video image captioning more convenient and easier, methods such as "Convolutional Neural Networks" (CNN)[6], "Recurrent Neural Networks" (RNNs)[7] and "Knowledge Graphs" (KGs)[8] have been introduced. was used. These methods use neural networks to learn features from images and videos, and then use those features to generate captions. This includes object acquisition, attribute detection, how they interact with other objects, relationships between objects, and how anyone can quickly understand what's happening in a video image.

It is known that most of the communication between machines and people depends on the understanding and interpretation of natural language [9]. Therefore, there are various applications of image description in real scenario[10]: Image captioning is becoming popular as a challenging and important area of study in artificial intelligence and computer vision[11] and it is of increasing importance is earning[12].

LITERATURE ANALYSIS AND METHODS

A common method of providing automatic video content description in human-understandable language is called video captioning[13]. The video captioning task allows the use of artificial intelligence, computer vision, and knowledge graphs[14]. Previously, video captioning was a task of visual content detection and sentence-wise captioning with hand-crafted features[15]. The purpose of creating video

captions is to provide a sequence of words to explain the visual content of that video. In order to understand video material, temporal dynamics must be captured, and video contains significantly more information than still images[16].

Captions describe a video image as a sentence or paragraph in natural language. You need to be very careful when choosing a title. That's why a single title serves to show the quality of an entire video image. Several scientists and researchers have conducted research in this regard. Researcher Liunian Li and his group[17] developed a Grounded Language-Image Pre-training (GLIP) model for learning object-level, linguistically aware and semantically rich visual images. The work proposes a deep integration of text and image, linguistically informing the detection model and building a solid foundational model. It also proposed to pre-train GLIP on extensible and semantically rich reasoning data through its reformulation and deep synthesis.

In his research, **Xuelong Li**[18] implemented text-based attention and semantic attention to reduce the semantic gap between visual and natural language. Finally, the authors gathered all the information to generate the desired answers in the form of a title for the optical question-answering system. Songtao Ding and his research group[19] proposed an attention theory for image captioning in psychology and combined low-level features, i.e., image quality, with high-level features, i.e., image regions, to focus attention on particular areas of the image. The authors applied attention theory to the psychology of visual headlines and used filtered image features[20].

Notably, researcher **Xinlei Chen**[21] used sentence-based bidirectional mapping between visual images. Their work led to the creation of new image captions and was able to reconstruct the visual features given in the image description. The authors tested their work on sentence generation and retrieval, and image retrieval. They used different datasets to evaluate the performance of their model such as PASCAL sentence[22], Flickr8K, Flickr30K and MSCOCO[23]. Researcher Huaishao[24] proposed a model known as "CLIP4Clip". This clip model knowledge is used for video language retrieval. They used video encoders and text decoders on datasets such as MSR-VTT, MSVC, LSMDC, ActivityNet, and DiDeMo[25].

RESULTS AND DISCUSSIONS

The term "functions" refers to signal processing operations performed by the system consisting of fundamental mathematical operations, such as edge enhancement, detection, and motion blur. These functions work to extract or enhance these fundamental features in the input signal and include integral and fractional order differentiation, Hilbert transform, and integration. For "differentiation" and "Hilbert" transformations, both the integral order and the continuous range of fractional order transformations must be performed first. So while there are 3 main types of functions to perform, a total of 34 functions can be achieved by including a number of integral and fractional routines. Importantly, all of these 34 features can be accessed without any hardware changes, just by tweaking system settings. It should also be noted that in addition to these 34 functions, the system can handle a continuous range of arbitrary fractional and higher-order differentiation and "Hilbert" fractional transformations.

The experiments presented here demonstrate real-time video image processing by simultaneously performing 34 functions including edge enhancement, edge detection, and motion blur[26]. Edge detection is the basis for object detection, feature extraction, and data compression [27]. This can be achieved by differentiating the temporal signal using high-integral or fractional-order derivatives, which obtain information about object boundaries in images or videos. It also performs a motion blur function based on signal integration, which represents the clear lines of moving objects in images or videos. It usually occurs when the recorded image changes during the recording process of a single exposure and can be widely used in computer animation and graphics[28]. Edge enhancement or sharpening based on the "Signal Hilbert" transform is also widely used[29] and is a key processing function. This improves the edge contrast of images or videos and at the same time improves their sharpness. The "standard Hilbert" transform performs a 90-degree phase shift and is commonly used in signal processing to generate a complex analytical signal from a real-valued signal. Also, arbitrary or fractional-order Hilbert has been proven to be useful for object image edge enhancement[30]. These processing functions facilitate not only traditional image or video processing [31] but also emerging technologies such as robotic vision and machine learning [32].

CONCLUSION

In short, with the advent of deep learning technology, the field of computer vision has developed rapidly. Deep learning technology has been used for image recognition, and it has made great progress in object recognition in recent years. Deep learning technology can process and analyze features by learning and simulating the cognitive capabilities of the human brain.

Unlike traditional feature extraction methods, deep convolutional neural networks can achieve high accuracy by extracting features using multi-layer convolution operations. In addition, they are robust to geometric changes, deformations, and lighting, and can overcome challenges caused by environmental changes. Deep learning methods can adapt the feature description using the training data, and they are highly flexible and have high generalization ability.

In recent years, with the rapid development of computer vision, object detection (OD) has been widely used in many fields as an important part of computer vision. Based on image processing, OD extracts features from images, and then extracts and analyzes object information such as category, location, and orientation. OD is widely used in many real-time situations, such as video monitoring, abnormal behavior analysis, and mobile robots. This approach can yield valuable information by extracting and analyzing features.

REFERENCES:

1. Liang, J.; Deng, Y.; Zeng, D. A deep neural network combined CNN and GCN for remote sensing scene classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.; 2020; 13, pp. 4325-4338. [DOI: https://dx.doi.org/10.1109/JSTARS.2020.3011333]
2. Chaudhuri, U.; Banerjee, B.; Bhattacharya, A.; Datcu, M. Attention-driven graph convolution network for remote sensing image retrieval. IEEE Geosci. Remote Sens. Lett.; 2021; 19, 8019705. [DOI: https://dx.doi.org/10.1109/LGRS.2021.3105448]
3. Ma, C.; Zeng, S.; Li, D. Image restoration and enhancement in monitoring systems. Proceedings of the 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS); Vientiane, Laos, 11–12 January 2020; pp. 753-760.
4. Zhang W, Tang S, Su J, Xiao J, Zhuang Y Tell and guess: cooperative learning for natural image caption generation with hierarchical refined attention. Multimed Tools Appl. 2021; 80: 16267-16282.
5. Cheng C, Li C, Han Y, Zhu Y A semi-supervised deep learning image caption model based on pseudo label and n-gram. Int J Approx Reason. 2021; 131: 93-107.
6. Lecun Y, Bottou L, Bengio Y, Haffner P Gradient-based learning applied to document recognition. Proc IEEE. 1998; 86: 2278-2324. doi:10.1109/5.726791
7. Dupond S. A thorough review on the current advance of neural net- work structures. Ann Rev Control. 2019; 14: 200-230.
8. Ehrlinger L, Wöß W Towards a definition of knowledge graphs. Paper presented at: SEMANTiCS (Posters, Demos, SuCCESS). 2016.
9. Bai S, An S. A survey on automatic image caption generation. Neurocomputing. 2018; 311: 291-304.
10. Wajid MA, Zafar A Multimodal information access and retrieval notable work and milestones. Paper presented at: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE. 2019:1-6.
11. Smeaton AF, Quigley I Experiments on using semantic distances between words in image caption retrieval. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1996:174-180.
12. Li K, Zhang Y, Li K, Li Y, Fu Y Visual semantic reasoning for image-text matching. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:4654-4662.
13. Chen S, Yao T, Jiang Y-G Deep learning for video captioning: a review. Paper presented at: IJCAI. 2019.
14. Kojima A, Tamura T, Fukunaga K Natural language description of human activities from video images based on concept hierarchy of actions. Int J Comput Vis. 2002; 50: 171-184.
15. Guadarrama S, Krishnamoorthy N, Malkarnenkar G, et al. Youtube2text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. Paper presented at: 2013 IEEE International Conference on Computer Vision. 2013:2712-2719.

16. Pei W, Zhang J, Wang X, Ke L, Shen X, Tai Y-W Memory-attended recurrent network for video captioning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:8347-8356.
17. Li LH, Zhang P, Zhang H, et al. Grounded language-image pre-training. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:10965-10975.
18. Li X, Yuan A, Lu X Vision-to-language tasks based on attributes and attention mechanism. IEEE Trans Cybern. 2019; 51: 913-926.
19. Ding S, Qu S, Xi Y, Sangaiah AK, Wan S Image caption generation with high-level image features. Pattern Recogn Lett. 2019; 123: 89-95.
20. Ishtiaque S, Wajid MS A review on medical image compression techniques. Int J Digit Appl Contemp Res 2017:17.
21. Chen X, Lawrence Zitnick C Mind's eye: a recurrent visual representation for image caption generation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:2422-2431.
22. Rashtchian C, Young P, Hodosh M, Hockenmaier J Collecting image annotations using amazon's mechanical turk. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. 2010:139-147.
23. Lin T-Y, Maire M, Belongie S, et al. Microsoft coco: common objects in context. Paper presented at: European Conference on Computer Vision, Springer. 2014:740-755.
24. Luo H, Ji L, Zhong M, et al. Clip4clip: An empirical study of clip for end-to-end video clip retrieval, arXiv preprint arXiv:2104.08860. 2021.
25. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 × 16 words: transformers for image recognition at scale, arXiv preprint arXiv:2010.11929. 2020.
26. Zhou, Y; Zheng, H; Kravchenko, II; Valentine, J. Flat optics for image differentiation. Nat. Photonics; 2020; 14, pp. 316-323. [DOI: https://dx.doi.org/10.1038/s41566-020-0591-3]
27. Zhu, T et al. Plasmonic computing of spatial differentiation. Nat. Commun.; 2017; 8, [DOI: https://dx.doi.org/10.1038/ncomms15391]
28. H. Ji, C. Q. Liu, Motion blur identification from image gradients, CVPR (2008).
29. Davis, JA; McNamara, DE; Cottrell, DM. Analysis of the fractional Hilbert transform. Appl. Opt.; 1998; 37, pp. 6911-6913. [DOI: https://dx.doi.org/10.1364/AO.37.006911]
30. Tan, M et al. Highly versatile broadband RF photonic fractional hilbert transformer based on a Kerr soliton crystal microcomb. J. Light. Technol.;2021;39,pp.75817587.[DOI:https://dx.doi.org/10.1109/JLT.2021.3101816]
31. Capmany, J et al. Microwave photonic signal processing. J. Light. Technol.;2013;31,pp.571586.[DOI:https://dx.doi.org/10.1109/JLT.2012.2222348]
32. Yang, T et al. Experimental observation of optical differentiation and optical Hilbert transformation using a single SOI microdisk chip. Scie. Rep.; 2014; 4, [DOI: https://dx.doi.org/10.1038/srep03960]