A Review on the Video Summarization and Glaucoma Detection

Tales Correia[1], António Cunha[2,3], and Paulo Coelho[1,4(✉)]

[1] School of Technology and Management, Polytechnic of Leiria, 2411-901 Leiria, Portugal
tales.correia@ipleiria.pt, paulo.coelho@ipleiria.pt

[2] Escola de Ciências e Tecnologias, University of Trás-os-Montes e Alto Douro, Quinta de Prados, 5001-801 Vila Real, Portugal
atcunha@utad.pt

[3] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), 4200-465 Porto, Portugal

[4] Institute for Systems Engineering and Computers at Coimbra (INESC Coimbra), DEEC, Pólo II, 3030-290 Coimbra, Portugal

Abstract. Glaucoma is a severe disease that arises from low intraocular pressure, it is asymptomatic in the initial stages and can lead to blindness, due to its degenerative characteristic. There isn't any available cure for it, and it is the second most common cause of blindness in the world. Regular visits to the ophthalmologist are the best way to prevent or contain it, with a precise diagnosis performed with professional equipment. From another perspective, for some individuals or populations, this task can be difficult to accomplish, due to several restrictions, such as low incoming resources, geographical adversities, and traveling restrictions (distance, lack of means of transportation, etc.). Also, logistically, due to its dimensions, relocating the professional equipment can be expensive, thus becoming inviable to bring them to remote areas. As an alternative, some low-cost products are available in the market that copes with this need, namely the D-Eye lens, which can be attached to a smartphone and enables the capture of fundus images, presenting as major drawback lower quality imaging when compared to professional equipment. Some techniques rely on video capture to perform summarization and build a full image with the desired features. In this context, the goal of this paper is to present a review of the methods that can perform video summarization and methods for glaucoma detection, combining both to indicate if individuals present glaucoma symptoms, as a pre-screening approach.

Keywords: Review · Glaucoma · Machine learning · Video summarization

A. Cunha et al. (Eds.): MobiHealth 2022, LNICST 484, pp. 144–156, 2023.
https://doi.org/10.1007/978-3-031-32029-3_14

1 Introduction

The human eyes take a key role in daily life, allowing humans to be able to see and perceive things all around their reach. Maintaining the eye in good health is extremely important, and the best way to keep track of its health is to have regular visits to the ophthalmologist. One of the most common eye checkups performed is the retinal fundus capture, which consists of the exposition of the back of the eye, where is located the macula and fovea, two of the principal areas where happens the actual formation of an image from the light incidence [1].

With the recent advances in studies in the fields of video summarization and machine learning, with a simple video recording taken from a smartphone using some gadgets with amplifying lens, is possible to convert it to a single image with the desired features to point out what could be an indication of glaucoma. This process doesn't intend to substitute the diagnosis process in a professional fundus machine, with specialist assistance, although it allows a quick summarization based on the captured video, to perform a pre-screening that may alert the individuals to seek professional assistance. Aiming to search for a method capable of performing retinal analysis based on low-quality imaging, thus, being capable of making pre-diagnosis with low-budget equipment, instead of professional medical equipment only found in clinics and hospitals. The idea is that with only a cell phone with a simple device, such as D-Eye attached, in hand, trained personnel can perform a retinal recording from patients that live in remote areas or have no facilitated access to other equipment. It is expected that it could help to orientate the most affected people to a further medical analysis, depending on the precision of the reading. Figure 1 shows a frame from a video captured of a patient eye using D-eye lens.

The advent of Video Summarization came from the need to gather important content from a significant amount of data, extracting the desired information from previously set keyframes that match the objective. With that in mind, selecting the keyframes and the features allied with the method to retrieve this information is crucial. Over the past years, many methods and techniques have been introduced and some of them will be shared later in this document, many of them being derived from machine learning, more specifically deep learning.
Some examples of deep learning methods are Convolutional Neural Networks (CNNs), very useful for gesture recognition [2], speech recognition [3], and as will be presented in this document, video summarization [4–18]. Some of the reviewed works use advanced deep learning methods [19–21] and apply 3D-CNN methods, which are ideal when treating high volumetric data. It is also one of the objectives to include these methods of video summarization, applied to the lower-quality images.

Section 2 of this paper will be presented the criterion and methods that were used to select the reviewed works, and Sect. 3 will be presented the state of the art for both video summarization and glaucoma detection. Section 4 will present the conclusion and what can be done in terms of improvements in future works.

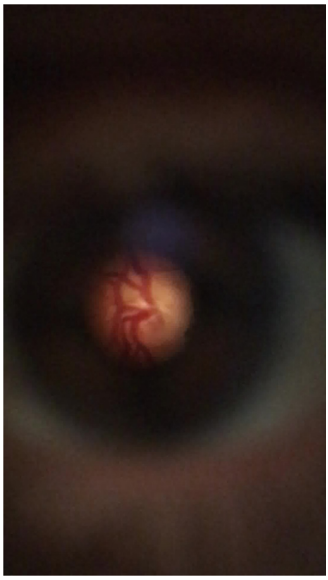

Fig. 1. Frame of retinal capture, using D-Eye lens.

2 Materials and Methods

2.1 Research Questions

This review was based on the following questions: (RQ1) Which methods can perform video summarization? (RQ2) Which methods can help identify glaucoma in a summarized image? (RQ3) From those methods which are realistic to be implemented specifically on low-quality acquired images? (RQ4) It is possible to adapt those methods in a multiple combination with other methods?

2.2 Inclusion Criteria

The study of the methods for both video summarization and glaucoma detection was performed with the following inclusion criteria: (1) studies that performed video summarization; (2) studies that performed glaucoma detection; (3) studies that apply deep-learning techniques; (4) studies with a relevant medical background; (5) studies that present their respective results/scores and datasets; (6) studies that were published between 2018 and 2022; (7) studies published in English.

2.3 Search Strategy

The reviewed studies with the selected inclusion criteria were searched in ScienceDirect and IEEE Xplore databases. The following research terms were used

to research this review: "video summarization" AND "glaucoma detection". Every study was independently evaluated by the authors, determining their suitability with the agreement of all reviewers. There was a total of 44 reviewed studies and after more criterion selection, 18 studies were selected in the final analysis. The research was performed on 16 May 2022.

3 Results

As presented in Fig. 2, 42 studies were identified from the selected sources, without duplicated papers. Two additional records were added to the results, gathered from different queries to the databases. After analyzing each research article's metadata, namely the title, abstract, and keywords, 6 studies were excluded from the analysis because they were medically specific and did not directly relate to evaluating the video summarization or glaucoma detection. The full text of the remaining 38 articles was assessed considering the inclusion criteria, and consequently, 20 articles were excluded. Finally, the remaining 18 papers were examined and included in qualitative and quantitative syntheses. The criteria for organizing and presenting the articles description was based on the relevance of the study returned by the research platforms, followed by the year of publication, from the most recent to oldest.

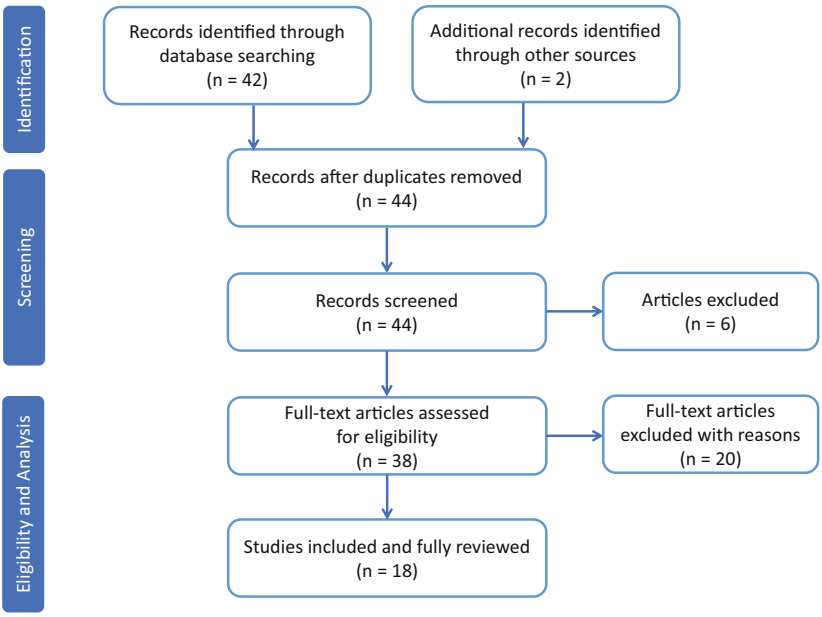

Fig. 2. Flow diagram of the selection of the papers.

3.1 Video Summarization

Some of the known techniques for video summarization rely on a set of tasks that the machine needs to perform to summarize relevant parts from video frames, taking into consideration the key features that have been selected to be present in the final set or image [22]. Figure 3 presents an example of the basic structure of how a video summarization algorithm performs. The input video is usually segmented frame by frame and based on what features the algorithm was trained to summarize, it will select the best shots or frames as output, depending on a chosen method or criteria.

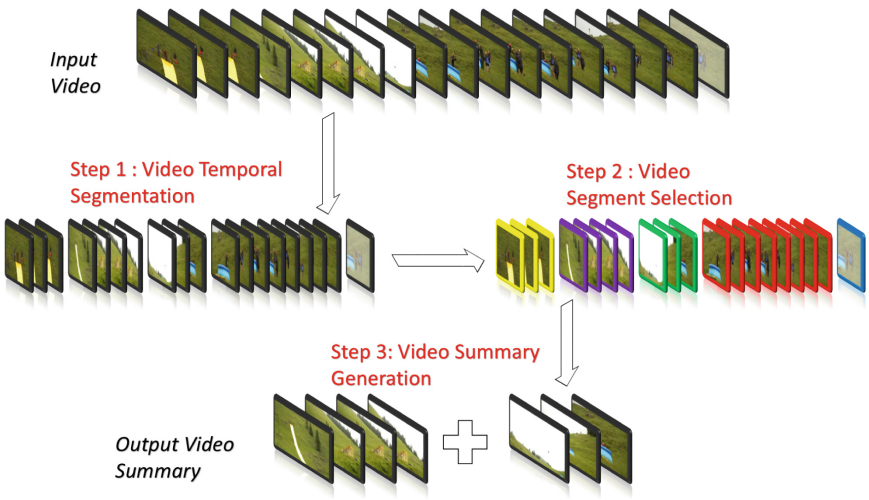

Fig. 3. Example of a video summarization algorithm structure - source [23].

The Long Short-Term Memory (LSTM) is a modern Recurrent Neural Network (RNN) useful to capture temporal dependencies between frames, but it has the issue of only being capable of handling short video segments, within a range from 30 to 80 frames. To overcome this, the method proposed by Lin et al. [4] employs a three-dimension Convolutional Neural Network (CNN) to extract features allied with a Deep Hierarchical LSTM Network with attention mechanism for Video Summarization (DHAVS) framework. This method was applied in SumMe [24], and TVSum [25] datasets and compared with recent results from other works with similar approaches. The F-Score obtained from DHAVS in SunMe was 45.2% and in the TVSum was around 60%.

In contrast to what was presented in [4], Zhao et al. [5] claims that RNN-based methods neglect global dependencies and multi-hop relationships between frames. To overcome that situation, a Hierarchical Multimodal Transformer (HMT) method is proposed for being capable of summarizing lengthy videos

by hierarchically separating the layers of dependency between shots, thus reducing memory and computational consumption. The metrics were also like the previously mentioned work as well as the datasets, where this method achieved an F-Score of 44.1% on SunMe and 60.1% on TVSum.

An attention mechanism is proposed to work with a dual-path attentive network by Liang et al. [6] to overcome the systematical stiffness of the recurrent neural networks. It was stated that their method improves the processing time and reduces the computational power while is possible to train the model in parallel, thus being scalable in bigger datasets. The results for F-Score from training and testing in SumMe and TVSum, with 51.7% and 61.5% respectively, were higher than what was presented in [4] and [5].

Feng et al. [11] proposed a video summarization technique that uses two different feature extraction that converts frame-level features into shot-level features based on CNN which was named Video Summarization with netVLAD and CapsNet (VCVS). Their method improved computational and hardware work while using a feature fusion algorithm with capsule neural (CapsNet) networks to enhance the video features. The F-score presented is 49.5% on SumMe and 61.22% on TVSum.

Some video summarization methods [12–14], can only extract the content of static images of those videos. Huang et al. [15] comes with a method to do both video and motion summarization, relying on transitions effects detection (TED) for automatic shots segmentation, using CapsNet, and a self-attention method to summarize the results. The scores for this method were 46.6% on SumMe and 58% on TVSum.

A multiscale hierarchical attention approach is proposed by [16] for pattern recognition using intra-block and inter-block attention methods, exploring short and long-range temporal representations of a video. The implemented method was developed this way because the attention mechanism is easier to implement than RNN. The achieved results are scores of 51.1% on SumMe and 61.0% on TVSum.

Chai et al. [17] propose a graph-based structural analysis in a three-step method that can detect the differences in continuous frames and establish the correct summarization of the video. For the tests they used VSUMM and Youtube datasets [18], in which compared to similar analyzed works, they achieved an F-score of 67.5% and 56.7%, respectively.

Another interesting approach, presented by Hussain et al. [7] shows a survey on multi-view video summarization (MVS), claiming that this technique is not addressed regularly as other mainstream summarization methods. Consisting in gathering video records from simultaneous cameras within different angles, the paper reviews the recent and most significant works that englobe MVS.

A self-attention binary neural tree (SABTNet) method is proposed by Fu et al. [8] to perform video summarization, subdividing the video and then extracting it to shot-level features, altogether with a self-attention imbued. This work is the first to introduce such an approach, and similarly to the previously presented,

the method was tested on SumMe and TVSum datasets, with F-scores of 50.7% and 61.0% respectively.

The work from Harakannanavar et al. [9] an approach based on ResNet-18, a CNN with eighteen layers, was used with kernel temporal segmentation (KTS) for the videos to create a temporally consistent summary. This method was benchmarked with the usual datasets SumMe and TVSum, obtaining 45.06% and 56.13% on F-scores respectively.

An interesting method is proposed in [10], a CNN with a Global Diverse Attention (SUM-GDA) mechanism, implying that the GDA provides relations within pair-frames and those pairs with all others in the video, stating that it overcomes the long-range issue from RNN models. They performed tests with supervised, unsupervised, and semi-supervised scenarios, with the usual datasets with the addition of VTW dataset [26]. The F-scores obtained from the tests were, as expected, higher in the supervised training, in which was obtained 52.8% on SumMe, 61% on TVSum, and 47.9% on VTW. Table 1 presents a resumed overview of the previously mentioned methods.

Table 1. Summary of the study analysis for video summarization.

Method	F-Score (%)		Remarkable Features
	SumMe	TVSum	
3D-CNN with DHAVS [4]	42.2	60	Employs LSTM to long videos.
HMT [5]	44.1	60.1	Reduces computational consumption by separating dependency between shots.
Dual-Path Attentive Network [6]	51.7	61.5	Improves computational consumption, improves process time and the model can be trained in parallel.
VCVS with CapsNet [11]	49.5	61.22	Improves computational and hardware work.
TED with CapsNet [15]	46.6	58	Summarization can be done in video and motion, not only in static images.
Multiscale Hierarchical Attention [16]	51.1	61	Captures of short and long-range dependencies also can perform motion detection.
SABTNet [8]	50.7	61	Shot-level segmentation and feature extraction.
ResNet-18 with KTS [9]	45.06	56.13	Temporal consistent summarization.
SUM-GDA [10]	52.8	61	Provides pair-frame relations within all video

It is relevant to imply that all those works are the most recent in terms of video summarization, making them a starting point as testing approaches to glaucoma detection. The following papers to be presented are more oriented to methods that have a direct impact on this matter.

3.2 Glaucoma Detection

A multimodal model to automatically detect glaucoma was proposed by [27] to combine deep neural networks focused on macular optical coherence tomography (OCT) and color fundus photography (CFP). Their dataset consisted of the UK Biobank dataset [28] with 1193 healthy and 1283 healthy and glaucomatous frames respectively. The OCT developed model was based on Densenet with MRSA initialization. For the CFP model, transfer learning with Inception Resnet V4 model, pre-trained on ImageNet data was used. Then it was introduced a gradient-boosted decision tree with XGBoost to create four separate baseline models (BM1 to BM4), enhancing specific features that they wanted to highlight. After testing the model, it was stated that mixing demographic and clinical features boosted the accuracy of diagnosis, obtaining around 97% of precision in correct results.

Trying to solve the issues of overfitting and big sets of data for training, Nayak et al. [29] proposed a method with a feature extraction called evolutionary convolutional network (ECNet) to perform automated glaucoma detection. They also applied an evolutionary algorithm named real-coded genetic algorithm (RCGA), which maximizes the inter-class distance and minimizes the intra-class variability to optimize the weight of the layers. Then a set of classifiers is applied, such as K-nearest neighbor (KNN), extreme learning machine (ELM), backpropagation neural network (BNN), and support vector machine (SVM), and kernel ELM (K-ELM), to enhance the model. They used a dataset, obtained from Kasturba Medical College, Manipal, India using a Zeiss FF 450 fundus camera, containing 1426 retinal fundus images, 589 healthy, and 837 with glaucoma. As for the results, the classifier that made the best score was SVM with 97.2% of obtaining the correct diagnosis. Figure 4 illustrates the difference between a normal eye and glaucoma affected with a picture taken from the fundus of the retina, exposing the optic disc and cup. The difference in size between the optic cup and optic disc, also known as CDR (cup to disc ratio) is one of the most common clinical glaucoma diagnoses.

Li et al. [30] proposed an attention-based CNN for glaucoma detection (AG-CNN), using large-scale attention-based glaucoma (LAG) database [31], a large-scale fundus image that has 5824 images within positive and negative glaucoma, obtained from Beijing Tongren Hospital. When this work was proposed, there was no other work that incorporated human attention in medical recognition. These attention maps were obtained through a simulated eye-tracking experiment and incorporated into the LAG dataset. The method had an accuracy of 95.3%.

An artificial intelligence technique method presented by Venugopal et al. [32] relies on Phase Quantized Polar Transformed Cellular Automaton (PQPT-CA) for training on fundus images for glaucoma detection in early stages, using ACRIMA database, with 705 fundus images within glaucoma and normal ones. This approach was chosen because of the recent results in image processing, slightly changing the existing architecture of the automaton to fit the proposed method, they could use it to extract the features boosting the accuracy by around 24%, being 21% faster, and reducing the false-positive results in 54%.

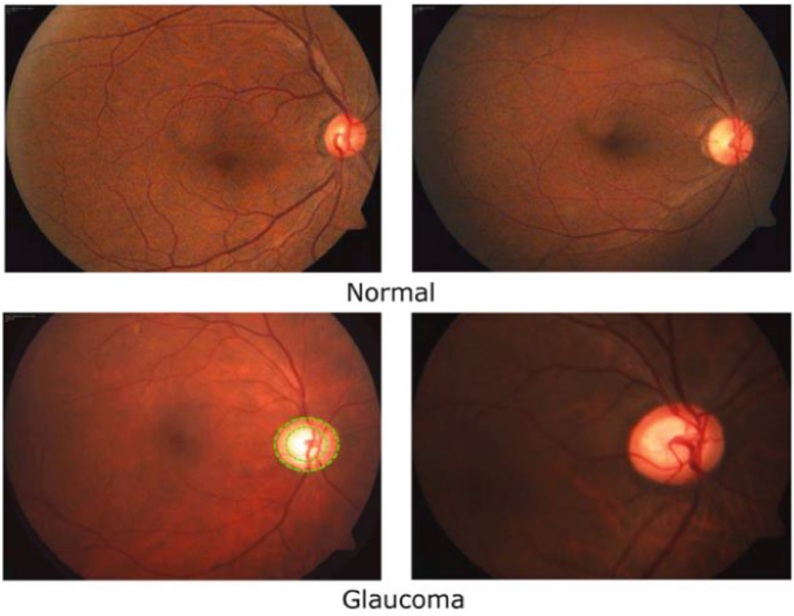

Fig. 4. Optic disc and optic cup comparison - source [29].

As for Zulfira et al. [33], they proposed a method that uses the classical parameter cup-to-disc ratio (CDR) allied with peripapillary atrophy (PPA) to enhance the precision of classification. They use an active contour snake (ACS) to segment the desired areas to calculate the CDR and Otsu's segmentation and threshold technique to acquire the PPA, and then the features are extracted with a grey-level co-occurrence matrix (GLCM). To classify glaucoma, dynamic ensemble selection (DES) is applied to make the final discrimination. The model was evaluated with three different databases where the ground truth was provided by ophthalmologists. Applying this method to RIM-ONE dataset [34] it was obtained an accuracy score of 96%.

Proposing the usage of 3D spectral-domain OCT, claiming that are potential information in these scans to help in glaucoma detection, Garcia et al. [35] brings a new perspective by presenting a method that uses the spatial dependencies of the features extracted from a B-scan of an OCT. Their database was composed of 176 healthy and 144 ill eyes. The method employed consisted of a slide-level feature extractor and a volume-based predictive model. They also used an LSTM network to combine the recurrent dependencies that will be further mixed into the latent space to provide a holistic feature vector that was generated by the proposed method of sequential-weighting module (SWM). The best results were achieved by using RAGNet-VGG16 architecture with an accuracy of 92%.

Gupta et al. [36] comes with a robust network to detect glaucoma in retinal fundus images based on CDR. They used two main modules, CLAHE [37] to

improve the retinal images and the second module to find the CDR after the image segmentation based on EfficientNet and U-net. They performed the tests in DRISHTI-GS and RIM-ONE datasets. The result for this method using the Dice coefficient for similarity was 96%, and the pixel accuracy for optic disc and cup was 96.54% and 96.89% respectively.

Table 2 presents a resumed overview of the previously mentioned methods.

Table 2. Summary of the study analysis for glaucoma detection.

Method	Dataset/Number of images	Accuracy (%)
OCT & CFP & Systemic & Ocular Model [27]	UK Biobank/2476	97.0
RCGA with SVM [29]	Kasturba Medical College/1426	97.2
Full AG-CNN [30]	LAG/5824	95.1
DES-MI [33]	RIM-ONE/250	96.0
RAGNet-VGG16 with SWM [35]	Private Dataset/905	92.0
CLAHE with EfficientNet + U-Net [36]	RIM-ONE/766	99.6

4 Conclusions

It is noticeable that in the past years, new methods and algorithms have been implemented to bring solutions to recurring problems, and more than ever machine learning is taking a huge part in those approaches. As instigated before in (RQ2), when talking about specific objectives, like glaucoma detection, it is known that the algorithm must adapt to extract the correct and desired key features.

Answering (RQ1), some of the works showed that RNNs and LSTM are not the best methods to treat long video summarization due to limitations on data length, being more useful on short length videos, within 30–80 frames range, like speech recognition videos. CNNs have presented some of the best results within long-length video summarization, thus being useful in some areas like medical, face recognition, security, and summarization of large data in general. Most of the reviewed articles were dependent on CNN methods, due to their excellent performance, but it is important to note that being a powerful tool also means that it will need an equivalent computational power.

Overall, the key to achieving satisfactory results in video summarization depends on a good comprehension of the features needed to be summarized and choosing the method or combinations of methods that best suit the desired outcome, also selecting an ideal classifier can help achieve better results, that answers the (RQ4). In glaucoma detection, there is a great field of study that still can be developed.

From what (RQ3) brought, it is also important to state that in the reviewed papers, all of them used public or private databases from high-quality images or videos, and one of the main ideas of this work is the pursuit of the best method

that can provide a reliable summary with low computational consumption, due to the usage of smartphones with lower-quality image acquisition.

Aiming for early detection with a fast and trustworthy algorithm is what this paper intends to propose for its next iteration. Instead of the conventional methods proposed in past works, a new algorithm capable of using a low-quality smartphone video of fundus recording and converting the resulting video to a single image with relevant features to finally bring a significant diagnostic, and of course, with a professional medical validation for those results.

Acknowledgements. This work is funded by FCT/MEC through national funds and, when applicable, co-funded by the FEDER-PT2020 partnership agreement under the project UIDB/00308/2020.

References

1. Cowan, C.S., et al.: Cell types of the human retina and its organoids at single-cell resolution. Cell **182**(6), 1623–1640 (2020)
2. Xu, L., Zhang, K., Yang, G., Chu, J.: Gesture recognition using dual-stream CNN based on fusion of sEMG energy kernel phase portrait and IMU amplitude image. Biomed. Sig. Process. Control **73**, 103364 (2022). https://doi.org/10.1016/j.bspc.2021.103364
3. Atila, O., Şengür, A.: Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition. Appl. Acoust. **182**, 108260 (2021)
4. Lin, J., Zhong, S.H., Fares, A.: Deep hierarchical LSTM networks with attention for video summarization. Comput. Electr. Eng. **97**, 107618 (2022). https://doi.org/10.1016/j.compeleceng.2021.107618
5. Zhao, B., Gong, M., Li, X.: Hierarchical multimodal transformer to summarize videos. Neurocomputing **468**, 360–369 (2022). https://doi.org/10.1016/j.neucom.2021.10.039
6. Liang, G., Lv, Y., Li, S., Wang, X., Zhang, Y.: Video summarization with a dual-path attentive network. Neurocomputing **467**, 1–9 (2022). https://doi.org/10.1016/j.neucom.2021.09.015
7. Hussain, T., Muhammad, K., Ding, W., Lloret, J., Baik, S.W., de Albuquerque, V.H.C.: A comprehensive survey of multi-view video summarization. Pattern Recogn. **109**, 107567 (2021). https://doi.org/10.1016/j.patcog.2020.107567
8. Fu, H., Wang, H.: Self-attention binary neural tree for video summarization. Pattern Recogn. Lett. **143**, 19–26 (2021). https://doi.org/10.1016/j.patrec.2020.12.016
9. Harakannanavar, S.S., Sameer, S.R., Kumar, V., Behera, S.K., Amberkar, A.V., Puranikmath, V.I.: Robust video summarization algorithm using supervised machine learning. Global Transitions Proc. **3**(1), 131–135 (2022). https://doi.org/10.1016/j.gltp.2022.04.009
10. Li, P., Ye, Q., Zhang, L., Yuan, L., Xu, X., Shao, L.: Exploring global diverse attention via pairwise temporal relation for video summarization. Pattern Recogn. **111**, 107677 (2021). https://doi.org/10.1016/j.patcog.2020.107677
11. Feng, X., Zhu, Y., Yang, C.: Video summarization based on fusing features and shot segmentation. In: Proceedings of 2021 7th IEEE International Conference on Network Intelligence and Digital Content, IC-NIDC 2021, pp. 383–387 (2021)

12. Badre, S.R., Thepade, S.D.: Summarization with key frame extraction using thepade's sorted n-ary block truncation coding applied on haar wavelet of video frame. In: 2016 Conference on Advances in Signal Processing, CASP, pp. 332–336 (2016)
13. Fei, M., Jiang, W., Mao, W.: Memorable and rich video summarization. J. Vis. Commun. Image Represent. **42**, 207–217 (2017). https://doi.org/10.1016/j.jvcir.2016.12.001
14. Mehmood, I., Sajjad, M., Rho, S., Baik, S.W.: Divide-and-conquer based summarization framework for extracting affective video content. Neurocomputing **174**, 393–403 (2016). https://doi.org/10.1016/j.neucom.2015.05.126
15. Huang, C., Wang, H.: A novel key-frames selection framework for comprehensive video summarization. IEEE Trans. Circ. Syst. Video Technol. **30**(2), 577–589 (2020)
16. Zhu, W., Lu, J., Han, Y., Zhou, J.: Learning multiscale hierarchical attention for video summarization. Pattern Recogn. **122**, 108312 (2022). https://doi.org/10.1016/j.patcog.2021.108312
17. Chai, C., et al.: Graph-based structural difference analysis for video summarization. Inf. Sci. **577**, 483–509 (2021). https://doi.org/10.1016/j.ins.2021.07.012
18. De Avila, S.E.F., Lopes, A.P.B., Da Luz, A., De Albuquerque Araújo, A.: VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recogn. Lett. **32**(1), 56–68 (2011). https://doi.org/10.1016/j.patrec.2010.08.004
19. Huang, S., Li, X., Zhang, Z., Wu, F., Han, J.: User-ranking video summarization with multi-stage spatio-temporal representation. IEEE Trans. Image Process. **28**(6), 2654–2664 (2019)
20. Agyeman, R., Muhammad, R., Choi, G.S.: Soccer video summarization using deep learning. In: Proceedings - 2nd International Conference on Multimedia Information Processing and Retrieval, MIPR 2019, pp. 270–273 (2019)
21. Riahi, A., Elharrouss, O., Al-Maadeed, S.: EMD-3DCNN-based method for COVID-19 detection. Comput. Biol. Med. **142**, 105188 (2022). https://doi.org/10.1016/j.compbiomed.2021.105188
22. Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I.: Video summarization using deep neural networks: a survey. Proc. IEEE **109**(11), 1838–1863 (2021)
23. Lei, Z., Zhang, C., Zhang, Q., Qiu, G.: FrameRank: a text processing approach to video summarization. In: Proceedings - IEEE International Conference on Multimedia and Expo, vol. 2019, pp. 368–373 (2019)
24. Gygli, M., Grabner, H., Riemenschneider, H., Van Gool, L.: Creating summaries from user videos. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 505–520. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10584-0_33
25. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: TVSum: summarizing web videos using titles. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07–12 June, pp. 5179–5187 (2015)
26. VTW Dataset. http://aliensunmin.github.io/project/%0Avideo-language/
27. Mehta, P., et al.: Automated detection of glaucoma with interpretable machine learning using clinical data and multimodal retinal images. Am. J. Ophthalmol. **231**, 154–169 (2021). https://doi.org/10.1016/j.ajo.2021.04.021
28. Sudlow, C., et al.: UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. **12**(3), 1–10 (2015)

29. Nayak, D.R., Das, D., Majhi, B., Bhandary, S.V., Acharya, U.R.: ECNet: an evolutionary convolutional network for automated glaucoma detection using fundus images. Biomed. Sig. Process. Control **67**, 102559 (2021). https://doi.org/10.1016/j.bspc.2021.102559
30. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: a large-scale database and CNN model. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 10563–10572 (2019)
31. Li, L., et al.: A large-scale database and a CNN model for attention-based glaucoma detection. IEEE Trans. Med. Imaging **39**(2), 413–424 (2020). https://ieeexplore.ieee.org/document/8756196/
32. Venugopal, N., Mari, K., Manikandan, G., Sekar, K.R.: Phase quantized polar transformative with cellular automaton for early glaucoma detection. Ain Shams Eng. J. **12**(4), 4145–4155 (2021). https://doi.org/10.1016/j.asej.2021.04.018
33. Zulfira, F.Z., Suyanto, S., Septiarini, A.: Segmentation technique and dynamic ensemble selection to enhance glaucoma severity detection. Comput. Biol. Med. **139**, 104951 (2021). https://doi.org/10.1016/j.compbiomed.2021.104951
34. RIM-ONE (2020). https://www.ias-iss.org/ojs/IAS/article/view/2346
35. García, G., Colomer, A., Naranjo, V.: Glaucoma detection from raw SD-OCT volumes: a novel approach focused on spatial dependencies. Comput. Methods Programs Biomed. **200**, 105855 (2021)
36. Gupta, N., Garg, H., Agarwal, R.: A robust framework for glaucoma detection using CLAHE and EfficientNet. Vis. Comput. 1–14 (2021). https://doi.org/10.1007/s00371-021-02114-5
37. Pizer, S.M., et al.: Adaptive histogram equalization and its variations. Comput. Vis. Graph. Image Process. **39**(3), 355–368 (1987). https://linkinghub.elsevier.com/retrieve/pii/S0734189X8780186X