Video Summarization Overview

Other titles in Foundations and Trends® in Computer Graphics and Vision

Deep Learning for Multimedia Forensics
Irene Amerini, Aris Anagnostopoulos, Luca Maiano and Lorenzo Ricciardi Celsi
ISBN: 978-1-68083-854-1

Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art
Joel Janai, Fatma Güney, Aseem Behl and Andreas Geiger
ISBN: 978-1-68083-688-2

Discrete Graphical Models - An Optimization Perspective
Bogdan Savchynskyy
ISBN: 978-1-68083-638-7

Line Drawings from 3D Models: A Tutorial
Pierre Bénard and Aaron Hertzmann
ISBN: 978-1-68083-590-8

Publishing and Consuming 3D Content on the Web: A Survey
Marco Potenziani, Marco Callieri, Matteo Dellepiane and Roberto Scopigno
ISBN: 978-1-68083-536-6

Crowdsourcing in Computer Vision
Adriana Kovashka, Olga Russakovsky, Li Fei-Fei and Kristen Grauman
ISBN: 978-1-68083-212-9

Video Summarization Overview

Mayu Otani
CyberAgent, Inc.
otani_mayu@cyberagent.co.jp

Yale Song
Microsoft Research
yalesong@microsoft.com

Yang Wang
University of Manitoba
ywang@cs.umanitoba.ca

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Computer Graphics and Vision

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

M. Otani *et al.*. *Video Summarization Overview*. Foundations and Trends® in Computer Graphics and Vision, vol. 13, no. 4, pp. 284–335, 2022.

ISBN: 978-1-63828-071-2

Foundations and Trends® in Computer Graphics and Vision
Volume 13, Issue 4, 2022

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Computer Graphics and Vision publishes survey and tutorial articles in the following topics:

- Rendering
- Shape
- Mesh simplification
- Animation
- Sensors and sensing
- Image restoration and enhancement
- Segmentation and grouping
- Feature detection and selection
- Color processing
- Texture analysis and synthesis
- Illumination and reflectance modeling
- Shape representation
- Tracking
- Calibration
- Structure from motion
- Motion estimation and registration
- Stereo matching and reconstruction
- 3D reconstruction and image-based modeling
- Learning and statistical methods
- Appearance-based matching
- Object and scene recognition
- Face detection and recognition
- Activity and gesture recognition
- Image and video retrieval
- Video analysis and event recognition
- Medical image analysis
- Robot localization and navigation

Information for Librarians

Foundations and Trends® in Computer Graphics and Vision, 2022, Volume 13, 4 issues. ISSN paper version 1572-2740. ISSN online version 1572-2759. Also available as a combined paper and online subscription.

Contents

Video Summarization Overview

Mayu Otani[1], Yale Song[2] and Yang Wang[3]

[1] *CyberAgent, Inc., Japan; otani_mayu@cyberagent.co.jp*
[2] *Microsoft Research, USA; yalesong@microsoft.com*
[3] *University of Manitoba, Canada; ywang@cs.umanitoba.ca*

ABSTRACT

With the broad growth of video capturing devices and applications on the web, it is more demanding to provide desired video content for users efficiently. Video summarization facilitates quickly grasping video content by creating a compact summary of videos. Much effort has been devoted to automatic video summarization, and various problem settings and approaches have been proposed. Our goal is to provide an overview of this field. This survey covers early studies as well as recent approaches which take advantage of deep learning techniques. We describe video summarization approaches and their underlying concepts. We also discuss benchmarks and evaluations. We overview how prior work addressed evaluation and detail the pros and cons of the evaluation protocols. Last but not least, we discuss open challenges in this field.

Mayu Otani, Yale Song and Yang Wang (2022), "Video Summarization Overview", Foundations and Trends® in Computer Graphics and Vision: Vol. 13, No. 4, pp 284–335. DOI: 10.1561/060000099.

1

Introduction

The wide spread use of internet and affordable video capturing devices have dramatically changed the landscape of video creation and consumption. In particular, user-created videos are more prevalent than ever with the evolution of video streaming services and social networks. The rapid growth of video creation necessitates advanced technologies that enable efficient consumption of desired video content. The scenarios include enhancing user experience for viewers on video streaming services, enabling quick video browsing for video creators who need to go through a massive amount of video rushes, and for security teams who need to monitor surveillance videos.

Video summarization facilitates quickly grasping video content by creating a compact summary of videos. One naive way to achieve video summarization would be to increase the playback speed or to sample short segments with uniform intervals. However, the former degrades the audio quality and distorts the motion (Benaim *et al.*, 2020), while the latter might miss important content due to the random sampling nature of the method. Rather than these naive solutions, video summarization aims to extract the information desired by viewers for more effective video browsing.

The purpose of video summaries varies considerably depending on application scenarios. For sports, viewers want to see moments that are critical to the outcome of a game, whereas for surveillance, video summaries need to contain scenes that are unusual and noteworthy. The application scenarios grow as more videos are created, *e.g.* we are beginning to see new types of videos such as video game live streaming and video blogs (vlogs). This has led to a new problem of video summarization as different types of videos have different characteristics and viewers have particular demands for summaries. Such a variety of applications has stimulated heterogeneous research in this field.

Video summarization addresses two principal problems: "what is the nature of a desirable video summary" and "how can we model video content." The answers depend on application scenarios. While these are still open problems for most application scenarios, many promising ideas have been proposed in the literature. Early work made various assumptions about requirements for video summaries, *e.g.* uniqueness (less-redundancy), diversity, and interestingness. Some works focused on creating video summaries that are relevant to user's intention and involve user interactions. Recent research focuses more on data-driven approaches from annotated datasets to learn desired video summaries.

Computational modeling of desirable video content is also an important challenge in video summarization. Starting with low-level features, various feature representations have been applied, such as face recognition and visual saliency. Recently, feature extraction using deep neural networks has been mainly adopted. Some applications further utilize auxiliary information such as subtitles for documentary videos, game logs for sports videos, and brain waves for egocentric videos captured with wearable cameras.

The goal of this survey is to provide a comprehensive overview of the video summarization literature. We review various video summarization approaches and compare their underlying concepts and assumptions. We start with early works that proposed seminal concepts for video summarization, and also cover recent data-driven approaches that take advantage of end-to-end deep learning. By categorizing the diverse research in terms of application scenarios and techniques employed, we aim to help researchers and practitioners to build video summarization systems for different purposes and application scenarios.

We also review existing benchmarks and evaluation protocols and discuss the key challenges in evaluating video summarization, which is not straightforward due to the difficulty of obtaining ground truth summaries. We provide an overview of how previous works have addressed challenges around evaluation and discuss strengths and weaknesses of existing evaluation protocols. Finally, we discuss open challenges in this area.

References

Agnihotri, L., J. Kender, N. Dimitrova, and J. Zimmerman. (2005). "Framework for personalized multimedia summarization". In: *ACM SIGMM International Workshop on Multimedia Information Retrieval*.

Aizawa, K., K. Ishijima, and M. Shiina. (2001). "Summarizing wearable video". In: *Proc. International Conference on Image Processing (ICIP)*. 398–401.

Apostolidis, E., E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras. (2020). "Performance over random: a robust evaluation protocol for video summarization methods". In: *ACM International Conference on Multimedia*. 1056–1064.

Apostolidis, E., E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras. (2021). "Video summarization using deep neural networks: a survey". *Proceedings of the IEEE*. 109(11): 1838–1863.

Arev, I., H. S. Park, Y. Sheikh, J. Hodgins, and A. Shamir. (2014). "Automatic editing of footage from multiple social cameras". *ACM Transactions on Graphics*. 33(4).

Avila, S. E. F. de, A. P. B. Lopes, A. da Luz, and A. de Albuquerque Araújo. (2011). "VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method". *Pattern Recognition Letters*. 32(1): 56–68.

Awad, G., A. A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, J. Liu, A. F. Smeaton, Y. Graham, G. J. F. Jones, W. Kraaij, and G. Quénot. (2020). "TRECVID 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains". In: *TRECVID 2020*.

Babaguchi, N., Y. Kawai, T. Ogura, and T. Kitahashi. (2004). "Personalized abstraction of broadcasted American football video by highlight selection". *IEEE Transactions on Multimedia*. 6(4): 575–586.

Babaguchi, N., K. Ohara, and T. Ogura. (2007). "Learning personal preference from viewer's operations for browsing and its application to baseball video retrieval and summarization". *IEEE Transactions on Multimedia*.

Behrooz, M., L. Michael, and T. Sinisa. (2017). "Unsupervised video summarization with adversarial LSTM networks". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2982–2991.

Benaim, S., A. Ephrat, O. Lang, I. Mosseri, W. T. Freeman, M. Rubinstein, M. Irani, and T. Dekel. (2020). "SpeedNet: learning the speediness in videos". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 9922–9931.

Block, F., V. Hodge, S. Hobson, N. Sephton, S. Devlin, M. F. Ursu, A. Drachen, and P. I. Cowling. (2018). "Narrative bytes: data-driven content production in esports". In: *ACM International Conference on Interactive Experiences for TV and Online Video*. 29–41.

Cheng-Yang Fu Joon Lee, M. B. and A. C. Berg. (2017). "Video highlight prediction using audience chat reactions". In: *Conference on Empirical Methods in Natural Language Processing*.

Chi, P.-Y., S. Ahn, A. Ren, M. Dontcheva, W. Li, and B. Hartmann. (2012). "MixT: automatic generation of step-by-step mixed media tutorials". In: *Annual ACM Symposium on User Interface Software and Technology*. 93–102.

Chu, W.-S. and A. Jaimes. (2015). "Video co-summarization: video summarization by visual co-occurrence". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 3584–3592.

del Molino, A. G. and M. Gygli. (2018). "PHD-GIFs: personalized highlight detection for automatic GIF creation". In: *ACM International Conference on Multimedia*.

DeMenthon, D., V. Kobla, and D. Doermann. (1998). "Video summarization by curve simplification". In: *ACM International Conference on Multimedia*. 211–218.

DeVries, T., I. Misra, C. Wang, and L. van der Maaten. (2019). "Does object recognition work for everyone?" In: *CVPR Workshop on Computer Vision for Global Challenges*. 52–59.

Evangelopoulos, G., A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. (2013). "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention". *IEEE Transactions on Multimedia*. 15(7): 1553–1568.

Fajtl, J., H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino. (2018). "Summarizing videos with attention". In: *ACCV Workshop on Attention/Intention Understanding*.

Feng, L., Z. Li, Z. Kuang, and W. Zhang. (2018). "Extractive video summarizer with memory augmented neural networks". In: *ACM International Conference on Multimedia*. 976–983.

Geisler, G. and G. Marchionini. (2000). "The open video project: research-oriented digital video repository". In: *ACM Conference on Digital Libraries*. 258–259.

Goldman, D. B., B. Curless, D. Salesin, and S. M. Seitz. (2006). "Schematic storyboarding for video visualization and editing". In: *ACM Transactions on Graphics*. Vol. 25. No. 3. 862–871.

Gong, Y. and X. Liu. (2000). "Video summarization using singular value decomposition". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 174–180.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. (2014). "Generative adversarial nets". *Advances in Neural Information Processing Systems*. 27.

Gygli, M., H. Grabner, H. Riemenschneider, and L. van Gool. (2014). "Creating summaries from user videos". In: *European Conference on Computer Vision*. Vol. 8695. 505–520.

Gygli, M., Y. Song, and L. Cao. (2016). "Video2GIF: automatic generation of animated GIFs from video". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1001–1009.

Ha, D., A. Dai, and Q. V. Le. (2017). "HyperNetworks". In: *International Conference on Learning Representations*.

Han, H.-K., Y.-C. Huang, and C. C. Chen. (2019). "A deep learning model for extracting live streaming video highlights using audience messages". In: *Artificial Intelligence and Cloud Computing Conference*. 75–81.

He, L., E. Sanocki, A. Gupta, and J. Grudin. (1999). "Auto-summarization of audio-video presentations". In: *ACM International Conference on Multimedia*. 489–498.

Jaimes, A., T. Echigo, M. Teraguchi, and F. Satoh. (2002). "Learning personalized video highlights from detailed MPEG-7 metadata". In: *IEEE International Conference on Image Processing*.

Jin, H., Y. Song, and K. Yatani. (2017). "Elasticplay: interactive video summarization with dynamic time budgets". In: *ACM International Conference on Multimedia*. 1164–1172.

Kaushal, V., S. Kothawade, A. Tomar, R. Iyer, and G. Ramakrishnan. (2021). "How good is a video summary? A new benchmarking dataset and evaluation framework towards realistic video summarization".

Khosla, A., R. Hamid, C.-j. Lin, and N. Sundaresan. (2013). "Large-scale video summarization using web-image priors". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2698–2705.

Kitani, K. M., T. Okabe, Y. Sato, and A. Sugimoto. (2011). "Fast unsupervised ego-action learning for first-person sports videos". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 3241–3248.

Laganière, R., R. Bacco, A. Hocevar, P. Lambert, G. Païs, and B. E. Ionescu. (2008). "Video summarization from spatio-temporal features". In: *ACM TRECVid Video Summarization Workshop*. 144–148.

Lee, Y. J., J. Ghosh, and K. Grauman. (2012). "Discovering important people and objects for egocentric video summarization". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1346–1353.

Lei, J., T. L. Berg, and M. Bansal. (2021). "QVHightlights: detecting moments and highlights in videos via natural language queries". In: *Advances in Neural Information Processing Systems*.

Li, B., H. Pan, and M. Sezan. (2003). "A general framework for sports video summarization with its application to soccer". *IEEE International Conference on Acoustics, Speech, and Signal Processing*. 3: III–169.

Liu, C., W. Zhang, S. Jiang, and Q. Huang. (2012). "Cross community news event summary generation based on collaborative ranking". In: *International Conference on Internet Multimedia Computing and Service*. 38–41.

Liu, W., T. Mei, Y. Zhang, C. Che, and J. Luo. (2015). "Multi-task deep visual-semantic embedding for video thumbnail selection". In: *IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, Y.-T., Y.-J. Li, and Y.-C. F. Wang. (2020). "Transforming multi-concept attention into video summarization". In: *Asian Conference on Computer Vision*.

Lu, Z. and K. Grauman. (2013). "Story-driven summarization for egocentric video". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2714–2721.

Lugaresi, C., J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. (2019). "MediaPipe: a framework for perceiving and processing reality". In: *CVPR Workshop*.

Ma, Y., L. Lu, H. Zhang, and M. Li. (2002). "A user attention model for video summarization". In: *ACM International Conference on Multimedia*. 533–542.

Mahmoud, K. M., N. M. Ghanem, and M. A. Ismail. (2013). "Unsupervised video summarization via dynamic modeling-based hierarchical clustering". In: *International Conference on Machine Learning and Applications*. Vol. 2. 303–308.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. (2021). "A survey on bias and fairness in machine learning". *ACM Computing Surveys*. 54(6).

Molino, A. G. del, X. Boix, J.-H. Lim, and A.-H. Tan. (2017). "Active video summarization: customized summaries via on-line interaction with the user". In: *Association for the Advancement of Artificial Intelligence Conference*.

Narasimhan, M., A. Rohrbach, and T. Darrell. (2021). "CLIP-It! language-guided video summarization". In: *Advances in Neural Information Processing Systems*.

Ngo, C.-W., Y.-F. Ma, and H.-J. Zhang. (2005). "Video summarization and scene detection by graph modeling". *IEEE Transactions on Circuits and Systems for Video Technology*. 15(2): 296–305.

Nguyen, C., Y. Niu, F. Liu, A. G. Money, and H. Agius. (2012). "Video summagator: an interface for video summarization and navigation". In: *SIGCHI Conference on Human Factors in Computing Systems*. Vol. 19. No. 2. 3–6.

Otani, M., Y. Nakahima, E. Rahtu, and J. Heikkilä. (2019). "Rethinking the evaluation of video summaries". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Panda, R. and A. K. Roy-Chowdhury. (2017a). "Multi-view surveillance video summarization via joint embedding and sparse optimization". *IEEE Transactions on Multimedia*. 19(9): 2010–2021.

Panda, R. and A. K. Roy-Chowdhury. (2017b). "Collaborative summarization of topic-related videos". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 4274–4283.

Park, J., J. Lee, I. Kim, and K. Sohn. (2020). "SumGraph: video summarization via recursive graph modeling". In: *European Conference on Computer Vision*. 647–663.

Pavel, A., D. B. Goldman, B. Hartmann, and M. Agrawala. (2015). "SceneSkim: searching and browsing movies using synchronized captions, scripts and plot summaries". In: 181–190.

Pavel, A., C. Reed, B. Hartmann, and M. Agrawala. (2014). "Video digests: a browsable, skimmable format for informational lecture videos". In: *Annual ACM Symposium on User Interface Software and Technology*. 573–582.

Peng, W.-T., W.-T. Chu, C.-H. Chang, C.-N. Chou, W.-J. Huang, W.-Y. Chang, and Y.-P. Hung. (2011). "Editing by viewing: automatic home video summarization by viewing behavior analysis". *IEEE Transactions on Multimediaa*.

Pfeiffer, S., R. Lienhart, S. Fischer, and W. Effelsberg. (1996). "Abstracting digital movies automatically". *Journal of Visual Communication and Image Representation*. 7(4): 345–353.

Poms, A., W. Crichton, P. Hanrahan, and K. Fatahalian. (2018). "Scanner: efficient video analysis at scale". *ACM Transactions on Graphics*. 37(4): 138:1–138:13.

Potapov, D., M. Douze, Z. Harchaoui, and C. Schmid. (2014). "Category-specific video summarization". en. In: *European Conference on Computer Vision*. 540–555.

Rochan, M., M. K. K. Reddy, and Y. Wang. (2020a). "Sentence guided temporal modulation for dynamic video thumbnail generation". In: *British Machine Vision Conference*.

Rochan, M., M. K. K. Reddy, L. Ye, and Y. Wang. (2020b). "Adaptive video highlight detection by learning from user history". In: *European Conference on Computer Vision*.

Rochan, M. and Y. Wang. (2019). "Video summarization by learning from unpaired data". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Rochan, M., L. Ye, and Y. Wang. (2018). "Video summarization using fully convolutional sequence networks". In: *European Conference on Computer Vision*. Ed. by V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss. 358–374.

Sang, J. and C. Xu. (2010). "Character-based movie summarization". In: *ACM International Conference on Multimedia*. 855–858.

Saquil, Y., D. Chen, Y. He, C. Li, and Y.-L. Yang. (2021). "Multiple pairwise ranking networks for personalized video summarization". In: *IEEE International Conference on Computer Vision*. 1718–1727.

Sawahata, Y. and K. Aizawa. (2003). "Wearable imaging system for summarizing personal experiences". In: *International Conference on Multimedia and Expo*. Vol. 1. 45–48.

Sharghi, A., B. Gong, and M. Shah. (2016). "Query-focused extractive video summarization". In: *European Conference on Computer Vision*.

Sharghi, A., J. S. Laurel, and B. Gong. (2017). "Query-focused video summarization: dataset, evaluation, and a memory network based approach". In: *IEEE Conference on Computer Vision and Pattern Recognition*.

Singla, A., S. Tschiatschek, and A. Krause. (2016). "Noisy submodular maximization via adaptive samplingwith applications to crowdsourced image collection summarization". In: *Association for the Advancement of Artificial Intelligence Conference*.

Smith, M. and T. Kanade. (1997). "Video skimming and characterization through the combination of image and language understanding techniques". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 775–781.

Song, Y. (2016). "Real-time video highlights for Yahoo Esports". In: *Advances in Neural Information Processing Systems*. 5 pages.

Song, Y., M. Redi, J. Vallmitjana, and A. Jaimes. (2016). "To click or not to click: automatic selection of beautiful thumbnails from videos". In: *ACM International on Conference on Information and Knowledge Management*. 659–668.

Song, Y., J. Vallmitjana, A. Stent, and A. Jaimes. (2015). "TVSum: summarizing web videos using titles". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 5179–5187.

Su, Y.-C. and K. Grauman. (2017). "Making 360° video watchable in 2D: learning videography for click free viewing". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1368–1376.

Su, Y.-C., D. Jayaraman, and K. Grauman. (2017). "Pano2Vid: automatic cinematography for watching 360° videos". In: *Asian Conference on Computer Vision*. 154–171.

Taskiran, C. M., Z. Pizlo, A. Amir, D. Ponceleon, and E. J. E. J. Delp. (2006). "Automated video program summarization using speech transcripts". *IEEE Transactions on Multimedia*. 8(4): 775–790.

Tiwari, V. and C. Bhatnagar. (2021). "A survey of recent work on video summarization: approaches and techniques". *Multimedia Tools and Applications*.

Uchihashi, S., J. Foote, A. Girgensohn, and J. Boreczky. (1999). "Video manga: generating semantically meaningful video summaries". In: *ACM International Conference on Multimedia*. 383–392.

Varini, P., G. Serra, and R. Cucchiara. (2015). "Egocentric video summarization of cultural tour based on user preferences". In: *ACM International Conference on Multimedia*.

Vasudevan, A. B., M. Gygli, A. Volokitin, and L. V. Gool. (2017). "Query-adaptive video summarization via quality-aware relevance estimation". In: *ACM International Conference on Multimedia*.

Wang, J., W. Wang, Z. Wang, L. Wang, D. Feng, and T. Tan. (2019). "Stacked memory network for video summarization". In: *ACM International Conference on Multimedia*. 836–844.

Xiong, B., Y. Kalantidis, D. Ghadiyaram, and K. Grauman. (2019). "Less is more: learning highlight detection from video duration". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Xiong, B., G. Kim, and L. Sigal. (2015). "Storyline representation of eocentric videos with an applications to story-based search". In: *IEEE International Conference on Computer Vision*. 4525–4533.

Xu, J., L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh. (2015). "Gaze-enabled egocentric video summarization via constrained submodular maximization". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2235–2244.

Yang, H., L. Chaisorn, Y. Zhao, S.-Y. Neo, and T.-S. Chua. (2003). "VideoQA: question answering on news video". In: *ACM Multimedia*.

Yang, S., J. Yim, J. Kim, and H. V. Shin. (2022). "CatchLive: real-time summarization of live streams with stream content and interaction data". In: *CHI Conference on Human Factors in Computing Systems*.

Yao, T., T. Mei, and Y. Rui. (2016). "Highlight detection with pairwise deep ranking for first-person video summarization". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 982–990.

Yeung, S., A. Fathi, and L. Fei-Fei. (2014). "VideoSET: video summary evaluation through text". In: *CVPR Workshop on Egocentric (First-person) Vision*. 15.

Yuan, Y., L. Ma, and W. Zhu. (2019). "Sentence specified dynamic video thumbnail generation". In: *ACM International Conference on Multimedia*.

Zen, G., P. de Juan, Y. Song, and A. Jaimes. (2016). "Mouse activity as an indicator of interestingness in video". In: *ACM on International Conference on Multimedia Retrieval*. 47–54.

Zhang, K., W.-L. Chao, F. Sha, and K. Grauman. (2016). "Video summarization with long short-term memory". In: *European Conference on Computer Vision*. 766–782.

Zhang, K., K. Grauman, and F. Sha. (2018). "Retrospective encoders for video summarization". In: *European Conference on Computer Vision*. 383–399.

Zhao, B., X. Li, and X. Lu. (2017). "Hierarchical recurrent neural network for video summarization". In: *ACM International Conference on Multimedia*. 863–871.

Zhao, B. and E. P. Xing. (2014). "Quasi real-time summarization for consumer videos". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2513–2520.