7 CONCLUSION

Identifying the need of the community, we presented VISIOCITY, a large benchmarking dataset and demonstrated its effectiveness in real world setting. To the best of our knowledge, it is the first of its kind in the scale, diversity and rich concept annotations. We introduce a recipe to automatically create ground truth summaries typically needed by the supervised techniques. We also extensively discuss and demonstrate the issue behind multiple right answers making the evaluation of video summaries a challenging task. Motivated by the fact that different good summaries have different characteristics and are not necessarily better or worse than the other, we propose an evaluation framework better geared at modeling human judgment through a suite of measures than having to overly depend on one measure. Finally through extensive and rigorous experiments we report the strengths and weaknesses of some representative state of the art techniques when tested on this new benchmark. Motivated by a fundamental problem in current supervised approaches, of learning from a single combined ground truth summary and/or learning from a single loss function tailored and optimizing *one* characteristic, our attempt to make simple extension to an existing mixture model technique gives encouraging results. We hope our attempt to address the multiple issues currently surrounding video summarization as highlighted in this work, will help the community advance the state of the art in video summarization.

ACKNOWLEDGEMENTS

This work is supported in part by the Ekal Fellowship (www.ekal.org) and National Center of Excellence in Technology for Internal Security, IIT Bombay (NCETIS, https://rnd.iitb.ac.in/node/101506)

REFERENCES

[1] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. 2020. Unsupervised Video Summarization via Attention-Driven Adversarial Learning. In *International Conference on Multimedia Modeling*. Springer, 492–504.

[2] Sandra Eliza Fontes De Avila, Ana Paula Brandão Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. 2011. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters* 32, 1 (2011), 56–68.

[3] David Doermann and David Mihalcik. 2000. Tools and techniques for video performance evaluation. In *icpr*. IEEE, 4167.

[4] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2018. Summarizing Videos with Attention. In *Asian Conference on Computer Vision*. Springer, 39–54.

[5] Cheng-Yang Fu, Joon Lee, Mohit Bansal, and Alexander C Berg. 2017. Video highlight prediction using audience chat reactions. *arXiv preprint arXiv:1707.08559* (2017).

[6] Tsu-Jui Fu, Shao-Heng Tai, and Hwann-Tzong Chen. 2019. Attentive and Adversarial Learning for Video Summarization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1579–1587.

[7] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. 2014. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems*. 2069–2077.

[8] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating Summaries from User Videos. In *ECCV*.

[9] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *European conference on computer vision*. Springer, 505–520.

[10] Michael Gygli, Helmut Grabner, and Luc Van Gool. 2015. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3090–3098.

[11] Mei Huang, Ayesh B Mahajan, and Daniel F DeMenthon. 2004. *Automatic performance evaluation for video summarization*. Technical Report. MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES.

[12] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. 2019. Video summarization with attention-based encoder-decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology* (2019).

[13] Sivapriyaa Kannappan, Yonghuai Liu, and Bernie Tiddeman. 2019. Human consistency evaluation of static video summaries. *Multimedia Tools and Applications* 78, 9 (2019), 12281–12306.

[14] Vishal Kaushal, Sandeep Subramanian, Suraj Kothawade, Rishabh Iyer, and Ganesh Ramakrishnan. 2019. A Framework towards Domain Specific Video Summarization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 666–675.

[15] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. 2013. Large-scale video summarization using web-image priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2698–2705.

[16] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 1346–1353.

[17] Zhuo Lei, Chao Zhang, Qian Zhang, and Guoping Qiu. 2019. FrameRank: A Text Processing Approach to Video Summarization. *arXiv preprint arXiv:1904.05544* (2019).

[18] Yingbo Li and Bernard Merialdo. 2010. VERT: automatic evaluation of video summaries. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 851–854.

[19] Yandong Li, Liqiang Wang, Tianbao Yang, and Boqing Gong. 2018. How local is the local diversity? Reinforcing sequential determinantal point processes with dynamic ground sets for supervised video summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 151–167.

[20] Zheng Lu and Kristen Grauman. 2013. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2714–2721.

[21] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. 2002. A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*. ACM, 533–542.

[22] Michel Minoux. 1978. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques*. Springer, 234–243.

[23] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila. 2019. Rethinking the Evaluation of Video Summaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7596–7604.

[24] Rameswar Panda, Niluthpol Chowdhury Mithun, and Amit K Roy-Chowdhury. 2017. Diversity-aware multi-video summarization. *IEEE Transactions on Image Processing* 26, 10 (2017), 4712–4724.

[25] Bryan A Plummer, Matthew Brown, and Svetlana Lazebnik. 2017. Enhancing video summarization via vision-language embedding. In *Computer Vision and Pattern Recognition*, Vol. 2.

[26] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. 2014. Category-specific video summarization. In *European conference on computer vision*. Springer, 540–555.

[27] Aidean Sharghi, Ali Borji, Chengtao Li, Tianbao Yang, and Boqing Gong. 2018. Improving sequential determinantal point processes for supervised video summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 517–533.

[28] Aidean Sharghi, Jacob S Laurel, and Boqing Gong. 2017. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2127–2136.

[29] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. TVSum: Summarizing web videos using titles.. In *CVPR*. IEEE Computer Society, 5179–5187. http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#SongVSJ15

[30] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5179–5187.

[31] Ba Tu Truong and Svetha Venkatesh. 2007. Video abstraction: A systematic review and classification. *ACM transactions on multimedia computing, communications, and applications (TOMM)* 3, 1 (2007), 3.

[32] Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool. 2017. Query-adaptive video summarization via quality-aware relevance estimation. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 582–590.

[33] Shuwen Xiao, Zhou Zhao, Zijian Zhang, Xiaohui Yan, and Min Yang. 2020. Convolutional Hierarchical Attention Network for Query-Focused Video Summarization. *arXiv preprint arXiv:2002.03740* (2020).

[34] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. 2019. Less is More: Learning Highlight Detection from Video Duration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1258–1267.

[35] Serena Yeung, Alireza Fathi, and Li Fei-Fei. 2014. Videoset: Video summary evaluation through text. *arXiv preprint arXiv:1406.5824* (2014).

[36] Li Yuan, Francis EH Tay, Ping Li, Li Zhou, and Jiashi Feng. 2019. Cycle-SUM: Cycle-consistent Adversarial LSTM Networks for Unsupervised Video Summarization. *arXiv preprint arXiv:1904.08265* (2019).

[37] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Summary transfer: Exemplar-based subset selection for video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1059–1067.

[38] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *European Conference on Computer Vision*. Springer, 766–782.

[39] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.

[40] Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2018. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Thirty-Second AAAI Conference on Artificial Intelligence*.