GUI recordings from end-users, we propose a more advanced approach to capture the spatial features of touch indicators and the temporal features of touch effects, to achieve better performance on user action identification.

To generate video captions, many works [78], [84] started using one single unified deep-learning model (one-fit-all). Recent works infused knowledge about objects in the video by using object detectors to generate more informative captions. For example, Zhang et al. [85] adopted an object detector to augment the object feature to yield object-specific video captioning. Different from the natural scenes, generating action-centric descriptions for GUI recording requires a more complex GUI understanding, as there are many aspects to consider, such as the elements in the GUI, their relationships, the semantics of icons, etc. Therefore, we modeled GUI-specific features by using mature methods, and then proposed a tailored algorithm to automatically generate natural language descriptions for GUI recordings.

VII. Conclusion

The bug recording is trending in bug reports due to its easy creation and rich information. However, watching the bug recordings and understanding the user actions can be time-consuming. In this paper, we present a lightweight approach `CAPdroid` to automatically generate semantic descriptions of user actions in the recordings, without requiring additional app instructions, recording tools, or restrictive video requirements. Our approach proposes image-processing and deep-learning models to segment bug recordings, infer user actions, and generate natural language descriptions. The automated evaluation and user study demonstrate the accuracy and usefulness of `CAPdroid` in boosting developers' productivity.

In the future, we will keep improving our method for better performance in terms of action segmentation and action attribute inference. According to user feedback, we will also improve the understanding of GUI to achieve higher-level semantic descriptions.

References

[1] S. Planning, "The economic impacts of inadequate infrastructure for software testing," *National Institute of Standards and Technology*, 2002.

[2] J. Anvik, L. Hiew, and G. C. Murphy, "Coping with an open bug repository," in *Proceedings of the 2005 OOPSLA workshop on Eclipse technology eXchange*, 2005, pp. 35–39.

[3] J. Aranda and G. Venolia, "The secret life of bugs: Going past the errors and omissions in software repositories," in *2009 IEEE 31st International Conference on Software Engineering*. IEEE, 2009, pp. 298–308.

[4] S. Feng and C. Chen, "Gifdroid: Automated replay of visual bug reports for android apps," in *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*. IEEE, 2022, pp. 1045–1057.

[5] "Record the screen on your iphone, ipad, or ipod touch," https://support.apple.com/en-us/HT207935 2022.

[6] "Take a screenshot or record your screen on your android device," https://support.google.com/android/answer/9075928?hl=en 2021.

[7] S. Feng and C. Chen, "Gifdroid: Automated replay of visual bug reports for android apps," *arXiv preprint arXiv:2112.04128*, 2021.

[8] C. Bernal-Cárdenas, N. Cooper, K. Moran, O. Chaparro, A. Marcus, and D. Poshyvanyk, "Translating video recordings of mobile app usages into replayable scenarios," in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, 2020, pp. 309–321.

[9] M. A. Gernsbacher, "Video captions benefit everyone," *Policy insights from the behavioral and brain sciences*, vol. 2, no. 1, pp. 195–202, 2015.

[10] T. J. Garza, "Evaluating the use of captioned video materials in advanced foreign language learning," *Foreign Language Annals*, vol. 24, no. 3, pp. 239–258, 1991.

[11] J. Chen, C. Chen, Z. Xing, X. Xu, L. Zhut, G. Li, and J. Wang, "Unblind your apps: Predicting natural-language labels for mobile gui components by deep learning," in *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*. IEEE, 2020, pp. 322–334.

[12] K. Moran, A. Yachnes, G. Purnell, J. Mahmud, M. Tufano, C. B. Cardenas, D. Poshyvanyk, and Z. H'Doubler, "An empirical investigation into the use of image captioning for automated software documentation," in *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 2022, pp. 514–525.

[13] C. Chen, T. Su, G. Meng, Z. Xing, and Y. Liu, "From ui design image to gui skeleton: a neural machine translator to bootstrap mobile gui implementation," in *Proceedings of the 40th International Conference on Software Engineering*, 2018, pp. 665–676.

[14] S. Feng, S. Ma, J. Yu, C. Chen, T. Zhou, and Y. Zhen, "Auto-icon: An automated code generation tool for icon designs assisting in ui development," in *26th International Conference on Intelligent User Interfaces*, 2021, pp. 59–69.

[15] S. Feng, M. Jiang, T. Zhou, Y. Zhen, and C. Chen, "Auto-icon+: An automated end-to-end code generation tool for icon designs in ui development," *ACM Transactions on Interactive Intelligent Systems*, vol. 12, no. 4, pp. 1–26, 2022.

[16] J. D. Cintas and A. Remael, *Audiovisual translation: subtitling*. Routledge, 2014.

[17] S. Feng and C. Chen, "Gifdroid: an automated light-weight tool for replaying visual bug reports," in *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings*, 2022, pp. 95–99.

[18] "Github," https://github.com/ 2022.

[19] D. Spencer, *Card sorting: Designing usable categories*. Rosenfeld Media, 2009.

[20] "Du recorder," https://www.du-recorder.com/ 2022.

[21] H. Chen, M. Sun, and E. Steinbach, "Compression of bayer-pattern video sequences using adjusted chroma subsampling," *IEEE transactions on circuits and systems for video technology*, vol. 19, no. 12, pp. 1891–1896, 2009.

[22] R. Sudhir and L. D. S. S. Baboo, "An efficient cbir technique with yuv color space and texture features," *Computer Engineering and Intelligent Systems*, vol. 2, no. 6, pp. 78–85, 2011.

[23] M. Livingstone and D. H. Hubel, *Vision and art: The biology of seeing*. Harry N. Abrams New York, 2002, vol. 2.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[25] Y. Du, C. Li, R. Guo, C. Cui, W. Liu, J. Zhou, B. Lu, Y. Yang, Q. Liu, X. Hu *et al.*, "Pp-ocrv2: Bag of tricks for ultra lightweight ocr system," *arXiv preprint arXiv:2109.03144*, 2021.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[28] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 203–213.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[34] P. Irani, C. Gutwin, and X. D. Yang, "Improving selection of off-screen targets with hopping," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006, pp. 299–308.