For nearly a decade, manual visual question answering systems[1], [2] have been employed where person uses a mobile application to ask a question and obtains answers through human volunteers or crowdsourcing workers. They were used for tasks like locating an object in a complex scene [3], helping the user in grocery shopping[1]. Such systems while valuable, depends on manual human labour. An automated system for such a task would have great benefits like enhanced latency, cost, privacy. In 2018, Gurari et al. [4] described unique challenges facing the visual question answering systems in assisting blind people. They also created a real-wold visual question answering dataset which was obtained from blind people. While constructing the dataset, prior works used dataset gathered from the web [5]–[8]. This was very tedious and required lot of resources. The dataset was not domain specific. Works like [9]–[11] used artificially created dataset. The artificially created dataset might inherently contain human bias which makes it drastically different from the real-world examples. The dataset taken from web might not mimic the images taken by the blind in real time. Hence the current VizWiz dataset solves these problems where the blind themselves have taken the pictures and asked a question. This mapping is saved and used in training.

The second class of model that we use is for automatic image captioning. It deals with image understanding and language description of the image. Now, we review some image captioning techniques used in the current literature. Farhadi et al. [12] used a triplet of scene elements to fill the template slots for generating image captions. Kulkarni et al. [13] used a conditional random field (CRF) to infer objects, attributes and prepositions before filling in the gaps. In our model, template-based method can be used with the help of answer generated by VQA model. Such models are relatively faster. However, template-based captioning follows a pre-defined template and has fixed length captions. In later iterations, parsing-based methods[14]–[18] can be used which are more powerful than simple template-based methods. Captioning can also be done using pre-defined retrieval-based approaches. Here, a set of captions is stored in advance. While captioning, visually similar images are found the captions are retrieved from the training set. These captions are called candidate captions and the captions for query images are retrieved from this caption pool[19]–[22]. Retrieval-based methods can be desirable as it is efficient and it can be assumed that a visually impaired person only encounters objects and scenes from a fixed set of possibilities. Such methods help in reducing the communication latency and also help in lowering the cost of the model. More advanced techniques which generate novel captions can be used [23]–[25]. They analyse the visual content of the image and generate captions using a language model.

A helpful and convenient feature is analysis of facial expressions or mood detection. Such a system can be very helpful for a visually impaired person as they are deprived of these capabilities. They generally rely on instinct and sound cues to make decisions in such cases. An image based automated system for detection of action units or basic and non-basic emotions can be implemented with the whole face as the region of interest. Rigid registration [26]–[29] can be used by detecting facial landmarks and using a distance metric like Euclidean distance or any affine transform that maps a input face to a prototypical face. Other non-rigid registration approaches [30] enable local registration and can supress errors due to facial inactivity. Hence, these systems are more robust. Point registration systems [31]–[33] are needed for shape representations. They involve localization of fiducial points and localization accuracy is crucial for such systems. Systems which depend on localizing the facial features require good quality image. Such a privilege cannot be guaranteed by our system. Hence, we depend on rigid registration methodology for facial expression identification and mood detection.

Visual-based scene change detection is also implemented. This helps provide additional real-time assistance to the visually impaired. Such a system is also helpful in implementing additional features like warning indicators and caution system. The scene change system we are implementing only depends on visual data. No auxiliary textual or audio data is required. Certain rule-based approaches [34] use techniques like rhythm based heuristics. Rule-based approaches only work for certain situations and have low performance. Kender and Yeo [35] introduced one of the first video segmentation algorithms. They calculated shot-to-shot coherence based on colour similarity. Other works [36], [37] also used such shot similarity measures to identify scene change. Hanjalic et al. [38] used block-based similarity to measure in LUV colour space. Kwon et al. [39] used motion-based features and improved overlapping link methods. Zhao et al. [40] and Cheng and Xu [41] considered temporal distance between the two shots. Wang et al. [42] introduced overlapping link method that uses forward and backward search. Such scene changes detection methodologies work well for our implementation as there is limited image data available.

[1] J. P. Bigham *et al.*, “VizWiz: Nearly real-time answers to visual questions,” in *UIST 2010 - 23rd ACM Symposium on User Interface Software and Technology*, 2010.

[2] W. S. Lasecki, P. Thiha, Y. Zhong, E. Brady, and J. P. Bigham, “Answering visual questions with conversational crowd assistants,” in *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2013*, 2013.

[3] J. P. Bigham, C. Jayant, A. Miller, B. White, and T. Yeh, “VizWiz::LocateIt - Enabling blind people to locate objects in their environment,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, 2010.

[4] D. Gurari *et al.*, “VizWiz Grand Challenge: Answering Visual Questions from Blind People,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[5] T. Y. Lin *et al.*, “Microsoft COCO: Common objects in context,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014.

[6] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, “SUN database: Large-scale scene recognition from abbey to zoo,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[7] G. Patterson and J. Hays, “SUN attribute database: Discovering, annotating, and recognizing scene attributes,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012.

[8] O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *Int. J. Comput. Vis.*, 2015.

[9] S. Antol *et al.*, “VQA: Visual question answering,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[10] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, “Neural module networks,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.

[11] J. Johnson, L. Fei-Fei, B. Hariharan, C. L. Zitnick, L. Van Der Maaten, and R. Girshick, “CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[12] A. Farhadi *et al.*, “Every picture tells a story: Generating sentences from images,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010.

[13] G. Kulkarni *et al.*, “Baby talk: Understanding and generating simple image descriptions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013.

[14] M. Mitchell *et al.*, “Midge: Generating image descriptions from computer vision detections,” in *EACL 2012 - 13th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, 2012.

[15] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, “Collective generation of natural image descriptions,” in *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*, 2012.

[16] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, “ T ree T alk : Composition and Compression of Trees for Image Descriptions ,” *Trans. Assoc. Comput. Linguist.*, 2014.

[17] A. Aker and R. Gaizauskas, “Generating image descriptions using dependency relational patterns,” in *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2010.

[18] D. Elliott and F. Keller, “Image description using visual dependency representations,” in *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2013.

[19] C. Sun, C. Gan, and R. Nevatia, “Automatic concept discovery from parallel text and visual corpora,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[20] V. Ordonez, G. Kulkarni, and T. L. Berg, “Im2Text: Describing images using 1 million captioned photographs,” in *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, 2011.

[21] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” in *IJCAI International Joint Conference on Artificial Intelligence*, 2015.

[22] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, “Improving image-sentence embeddings using large weakly annotated photo collections,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014.

[23] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.

[24] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, “Boosting Image Captioning with Attributes,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[25] K. Xu *et al.*, “Show, attend and tell: Neural image caption generation with visual attention,” in *32nd International Conference on Machine Learning, ICML 2015*, 2015.

[26] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, “A dynamic appearance descriptor approach to facial actions temporal modeling,” *IEEE Trans. Cybern.*, 2014.

[27] G. C. Littlewort, M. S. Bartlett, and K. Lee, “Automatic coding of facial expressions displayed during posed and genuine pain,” *Image Vis. Comput.*, 2009.

[28] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1998.

[29] G. Tzimropoulos, V. Argyriou, S. Zafeiriou, and T. Stathaki, “Robust FFT-based scale-invariant image registration with image gradients,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010.

[30] S. Yang and B. Bhanu, “Facial expression recognition using emotion avatar image,” in *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, 2011.

[31] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, “Facial point detection using boosted regression and graph models,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[32] D. Vukadinovic and M. Pantic, “Fully automatic facial feature point detection using gabor feature based boosted classifiers,” in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 2005.

[33] M. F. Valstar and M. Pantic, “Fully automatic recognition of the temporal phases of facial actions,” *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, 2012.

[34] B. Adams, C. Dorai, and S. Venkatesh, “Toward automatic extraction of expressive elements from motion pictures: Tempo,” *IEEE Trans. Multimed.*, 2002.

[35] J. R. Kender and B. L. Yeo, “Video scene segmentation via continuous video coherence,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1998.

[36] Z. Rasheed and M. Shah, “Scene detection in Hollywood movies and TV shows,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003.

[37] Y. Rui, T. S. Huang, and S. Mehrotra, “Constructing table-of-content for videos,” *Multimed. Syst.*, 1999.

[38] A. Hanjalic, R. L. Lagendijk, and J. Biemond, “Automated high-level movie segmentation for advanced video-retrieval systems,” *IEEE Trans. Circuits Syst. Video Technol.*, 1999.

[39] Y. M. Kwon, C. J. Song, and I. J. Kim, “A new approach for high level video structuring,” in *IEEE International Conference on Multi-Media and Expo*, 2000.

[40] L. Zhao, S. Q. Yang, and B. Feng, “Video scene detection using slide windows method based on temporal constrain shot similarity,” in *Proceedings - IEEE International Conference on Multimedia and Expo*, 2001.

[41] W. Cheng and D. Xu, “A novel approach of generating video scene structure,” in *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, 2003.

[42] X. Wang, S. Wang, H. Chen, and M. Gabbouj, “A shot clustering based algorithm for scene segmentation,” in *Proceedings - CIS Workshops 2007, 2007 International Conference on Computational Intelligence and Security Workshops*, 2007.