Visual Computing Magazine

RGB-D Segment Captioning

K. Delloul, Prof. S. Larabii., Computer Science Faculty, USTHB University

1. Introduction

In recent years, image captioning and image segmentation have emerged as crucial tasks in computer vision, with applications ranging from autonomous driving to content analysis. While several solutions have emerged to textually describe the content of a given image, few AI models are capable of generating fully detailed descriptions of each panoramic segment within an image.

In our conducted research, we propose an approach based on a deep learning model that generates a descriptive phrase for each segment present in the image. The uniqueness of our research lies in the fact that the generated captions are enriched with region positions relative to the user (left, right, front) and position relationships between the regions. This solution is applied to our TS-RGBD dataset, consisting of images collected in the auditorium using the Kinect.

Therefore, the goal is to create a deep learning model for translating images into text, where the output takes the form of textual descriptions of all segments in the input image. The textual descriptions of the segments are enhanced by their positioning relative to the user (left, right, in front) and relative to other segments. The solution is intended to assist visually impaired individuals.

2. Method

Before embarking on any scientific work, a state-of-the-art study is essential. We first collected a set of publications related to topics relevant to our subject, such as Image Captioning, Image Segmentation and Depth Estimation. On the other hand, a state-of-the-art study also includes studying datasets that could be used to complete our research.

Analyzing different published papers and proposed solutions from laboratories and firms from all around the globe allowed us to build a survey and to present the findings of our analysis in [1]. We classified different image captioning models according to their architectures, from single sentence to paragraph captioning. We presented different existing datasets and their lack of diversity as well as their shortcomings regarding depth information and detailed annotations. We also highlighted the limitations of existing solutions regarding our research objectives and thus the necessity to start building our own model.

RGB-D Segment Captioning

K. Delloul, Prof. S. Larabii., Computer Science Faculty, USTHB University

One of the main pillars to our research is a the DenseCap model for dense captioning proposed by a Ph.D. student of Prof. Fei Fei Li, a pionneer in Computer Vision. It takes an RGB image as input and returns a number of regions of interest with their descriptive sentences. We generated dense descriptions for the Visual Genome Dataset on which DenseCap was trained, then we applied another model for depth estimation of said images. Both depth and RoI boxes were used to generated egocentric descriptions of images, which constitute the topic of our second paper [2].

For each image from the VG dataset, we used the AdaBins model to generate a depth map as seen in the figure 1. Note that depth is the distance between the pixel and the camera. AdaBins was trained to estimate this depth from monocular images.

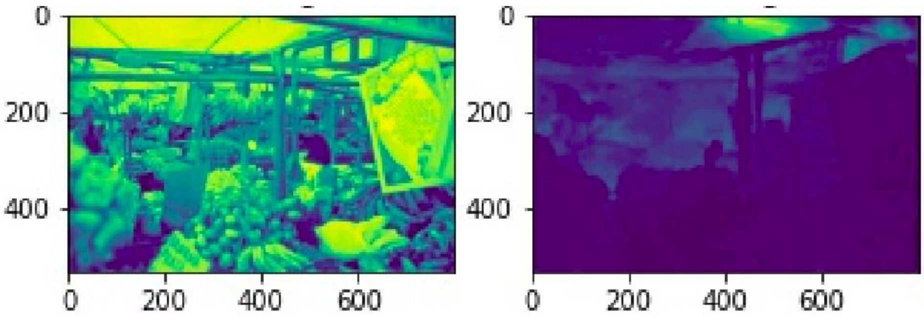


Figure 1. Image from VG Dataset and the computed depth map.

Then using DenseCap model we extracted bounding boxes of regions of interest with generated captions (see figure 2). The same process was applied on frames we extracted from theatre shows that were available on YouTube for free (see figure 3).

RGB-D Segment Captioning

K. Delloul, Prof. S. Larabii., Computer Science Faculty, USTHB University

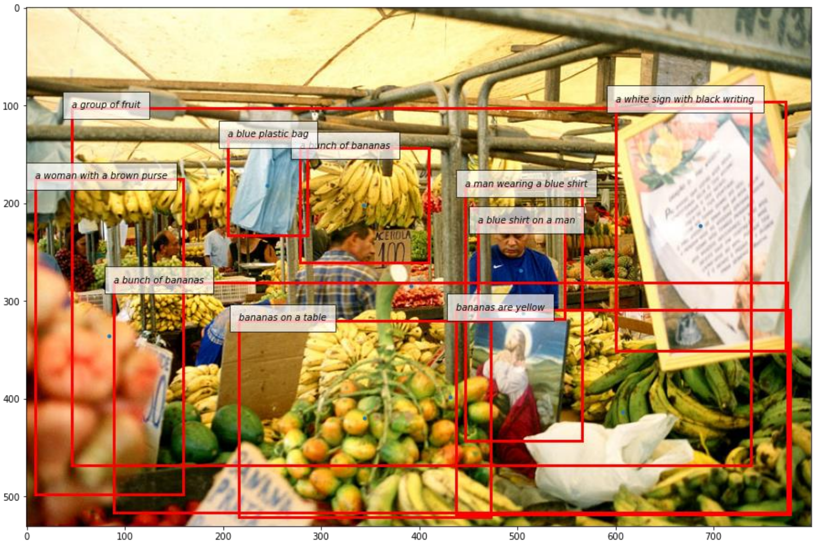


Figure 2. Extracted bounding boxes.

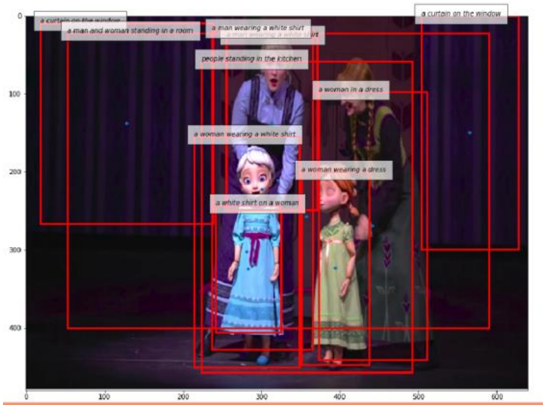
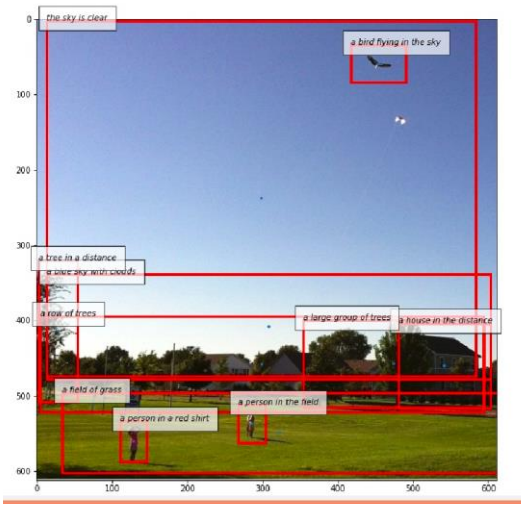

a curtain on the window
a man and woman standing in a room
a man wearing a white shirt
a man wearing a white shirt
a curtain on the window
people standing in the kitchen
a woman in a dress
a woman wearing a white shirt
a woman wearing a dress
a white shirt on a woman
0
100
200
300
400
0
100
200
300
400
500
600

Figure 3. Image from theatre shows .

RGB-D Segment Captioning

K. Delloul, Prof. S. Larabi., Computer Science Faculty, USTHB University

Then we passed results through our proposed algorithm [4] to get egocentric descriptions.



In front of you there is a row of trees, a field of grass, the sky is clear, a person in the field, a blue sky with clouds

On your right a bird flying in the sky, a house in the distance, a large group of trees

And on your left a person in a red shirt, a tree in a distance

Figure 4. Labelling with egocentric description.

Evaluation of the achieved results on 20 images from VG and theatre scenes is presented in the table below:

Dataset	Captions N°	Correct Directions	Incorrect Directions	Accuracy
VG	196	175	21	89%
Theatre	200	174	26	87%

RGB-D Segment Captioning

K. Delloul, Prof. S. Larabii., Computer Science Faculty, USTHB University

The proposed method for image captioning with a focus on blind guidance and scene understanding has identified several limitations. Firstly, the method's applicability is limited to specific places due to the lack of diversity in the VG dataset, particularly affecting visually impaired individuals in varied environments. The depth estimation achieved good metrics, but the generated map lacks real depth values, hindering the output of actual distances and angles between objects and users. The AdaBins model also has restrictions on image sizes and formats, suggesting the use of RGB-D image sensors for better performance. Additionally, reliance on the DenseCap model introduces redundancies and inaccuracies in captioning, especially that it was not trained on theatre images. To address these issues, we emphasize the necessity of collecting and annotating RGB-D images of theatre scenes, recognizing it as a crucial step for further research.

So in the next step we proceeded to collect our own dataset using RGB-D images and theatre-like scenarios. The image capturing took place in the university amphitheater using two Microsoft Kinects v1, with volunteer students.

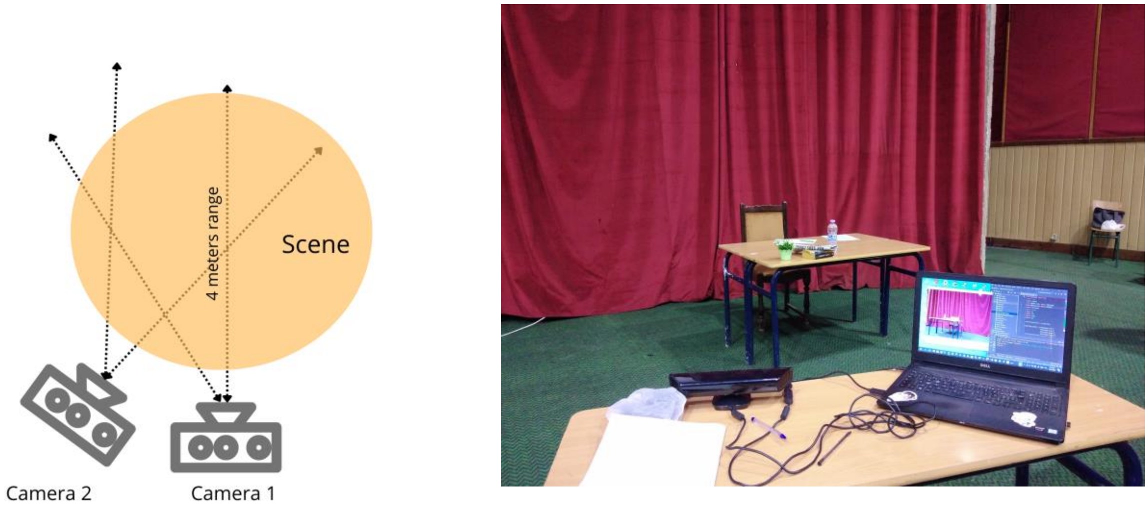


Figure 5. Collect of RGB-D images.

RGB-D Segment Captioning

K. Delloul, Prof. S. Larabii., Computer Science Faculty, USTHB University

Images from the dataset were annotated in such a way that for each panoptic segment of the image, there is a descriptive phrase. Annotations were made using the open source program LabelMe.

Finally, DenseCap model was modified to be capable of providing multiple sentences per image, each corresponding to a segment, instead of extracting the regions of interest itself, the model is fed with panoptic segments that were generated by a segmentation model OneFormer.

The architecture of our proposed solution can be illustrated as follows [4]:

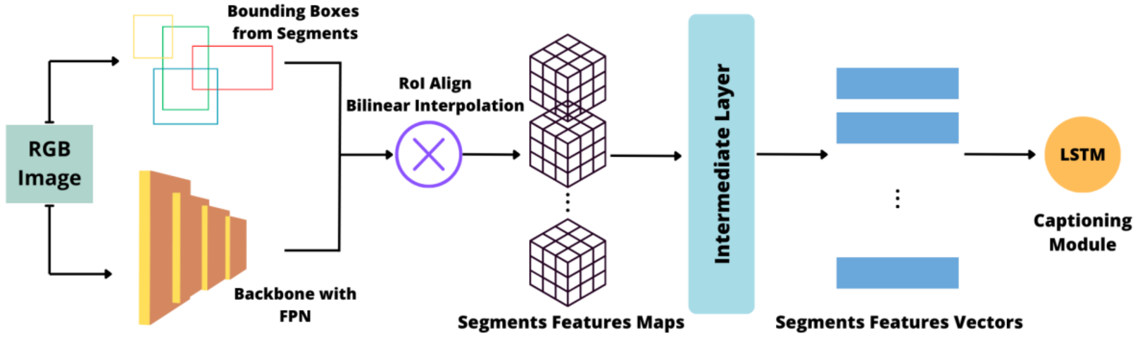


Figure 6. The used model.

The sentences are enriched by directions extracted from the point cloud, as well as the positioning relationships between different segments. The point cloud is generated using depth values that were collected using the Kinect.

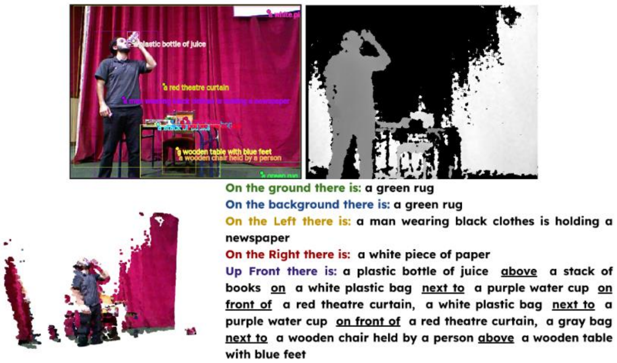
The solution is tested on our new TS-RGBD dataset of theater scenes. The solution is fast compared to dense textual description models and effective in terms of directions and relationships between segments, thanks to the depth information.

RGB-D Segment Captioning

K. Delloul, Prof. S. Larabii., Computer Science Faculty, USTHB University

Conclusion

We provided a comprehensive review of recent advancements in AI technologies for visual scene understanding, covering image captioning, image segmentation, and scene understanding. Despite notable progress, challenges such as handling occlusions, non salient regions, and generalizing to unseen scenarios persist. Our study addresses these challenges by developing a framework that enhances scene understanding through textual descriptions of image segments. The solution, applied to our novel TS-RGBD dataset of theatre scenes, outperforms other image captioning models in terms of captions per image and execution time. The approach successfully processes positional relationships using depth information, and future work includes refining ground truth captions, expanding the dataset, and incorporating more sophisticated sensors for wider applications, particularly in actual theatre plays, with plans for evaluation by blind and visually impaired users.


a plastic bottle of juice
a red theatre curtain
a wooden table with blue feet
a wooden chair held by a person

On the ground there is: a green rug
On the background there is: a green rug
On the Left there is: a man wearing black clothes is holding a newspaper
On the Right there is: a white piece of paper
Up Front there is: a plastic bottle of juice above a stack of books on a white plastic bag next to a purple water cup on front of a red theatre curtain, a white plastic bag next to a purple water cup on front of a red theatre curtain, a gray bag next to a wooden chair held by a person above a wooden table with blue feet

Figure 7. Result of labelling image from TS-RGBD dataset

References

[1] Delloul Khadidja, Slimane Larabi.
Image Captioning State-of-the-Art: Is It Enough for the Guidance of Visually Impaired in an Environment? .
In: Senouci, M.R., Boulahia, S.Y., Benatia, M.A. (eds) Advances in Computing Systems and Applications. 17-18 May, CSA 2022. Lecture Notes in Networks and Systems, vol 513. Springer, Cham.
[2] Delloul Khadidja, Slimane Larabi.
Egocentric Scene Description for the Blind and Visually Impaired .
5th International Symposium on Informatics and its Applications (ISIA), M'Sila University, November 29-30, 2022
[3] Leyla Benhamida, Khadidja Delloul, Slimane Larabi.
TS-RGBD Dataset: a Novel Dataset for Theatre Scenes Description for People with Visual Impairments.
arXiv:2308.01035 [cs.CV], 2 Aug 2023
[4] Khadidja Delloul, Slimane Larabi.
Towards Real Time Egocentric Segment Captioning for The Blind and Visually Impaired in RGB-D Theatre Images.
arXiv:2308.13892 [cs.CV], 26 Aug 2023