Architecture and Planning Journal (APJ)

Volume 28 Issue 3 *ASCAAD 2022 - Architecture in the Age of the Metaverse – Opportunities and Potentials*
ISSN: 2789-8547

Article 5

March 2023

COMPUTER VISION AIDED HOTSPOT CREATION IN VIRTUAL ENVIRONMENTS

LAMA A. AFFARA
Department of Mathematics & Computer Science, Faculty of Science, Beirut Arab University, Lebanon,
l.affara@bau.edu.lb

BILAL E. NAKHAL
Department of Mathematics & Computer Science, Faculty of Science, Beirut Arab University, Lebanon,
b.nakhal@bau.edu.lb

Recommended Citation

AFFARA, LAMA A. and NAKHAL, BILAL E. (2023) "COMPUTER VISION AIDED HOTSPOT CREATION IN VIRTUAL ENVIRONMENTS," *Architecture and Planning Journal (APJ)*: Vol. 28: Iss. 3, Article 5.
DOI: https://doi.org/10.54729/2789-8547.1200

COMPUTER VISION AIDED HOTSPOT CREATION IN VIRTUAL ENVIRONMENTS

Abstract

Hotspot creation is one of the most important modules within virtual environments which helps show the navigators of these environments some information about semantic elements within it and facilitate the navigation between the virtual spaces. In this paper, a system for automatic hotspot proposals and creation in virtual environments is proposed. The system uses computer vision modules to automatically propose hotspot locations in addition to identifying and creating these hotspots with candidate labels. Two main modules used in the system are object detection and scene segmentation. The scene segmentation helps give candidate hotspot areas and provides an overall understanding of the semantics of the virtual environment. The object detection module also uses pretrained deep networks for automatic hotspot creation over these objects. The system helps speed up the hotspot creation process and offers a tool for virtual environment users and creators.

Keywords

hybrid metaverse, virtual environments, object detection, deep learning, computer vision

COMPUTER VISION AIDED HOTSPOT CREATION IN VIRTUAL ENVIRONMENTS

LAMA A. AFFARA AND BILAL E. NAKHAL

Assistant Professors, Department of Mathematics & Computer Science, Faculty of Science,
Beirut Arab University, Lebanon
l.affara@bau.edu.lb
b.nakhal@bau.edu.lb

ABSTRACT

Hotspot creation is one of the most important modules within virtual environments which helps show the navigators of these environments some information about semantic elements within it and facilitate the navigation between the virtual spaces. In this paper, a system for automatic hotspot proposals and creation in virtual environments is proposed. The system uses computer vision modules to automatically propose hotspot locations in addition to identifying and creating these hotspots with candidate labels. Two main modules used in the system are object detection and scene segmentation. The scene segmentation helps give candidate hotspot areas and provides an overall understanding of the semantics of the virtual environment. The object detection module also uses pretrained deep networks for automatic hotspot creation over these objects. The system helps speed up the hotspot creation process and offers a tool for virtual environment users and creators.

Keywords: hybrid metaverse, virtual environments, object detection, deep learning, computer vision

ملخص

يعد إنشاء النقاط الفعالة أحد أهم الوحدات النمطية في البيئات الافتراضية والتي تساعد في إظهار ملاحي هذه البيئات بعض المعلومات حول العناصر الدلالية بداخلها وتسهيل التنقل بين المساحات الافتراضية. في هذه الورقة البحثية، تم تقديم نظام لاقتراح وإنشاء النقاط الفعالة التلقائية في بيئات افتراضية. يستخدم النظام تقنية رؤية الكمبيوتر لاقتراح مواقع النقاط الفعالة تلقائيًا بالإضافة إلى تحديد وإنشاء هذه النقاط الفعالة باستخدام الملصقات المرشحة. هناك وحدتان أساسيتان مستخدمتان في النظام هما كشف الأشياء وتجزئة المشهد. يساعد تقسيم المشهد في إعطاء مناطق النقاط الفعالة المرشحة ويوفر فهمًا شاملاً لدلالات البيئة الافتراضية. تستخدم وحدة كشف الأشياء أيضًا شبكات عميقة مُدرَّبة مسبقًا لإنشاء نقطة اتصال تلقائية فوق هذه الأشياء. يساعد النظام في تسريع عملية إنشاء النقاط الفعالة ويوفر أداة لمستخدمي البيئة الافتراضية ومنشئيها.

الكلمات المفتاحية: الميتافرس المختلط، البيئات الافتراضية، كشف الأشياء، التعلم العميق، تقنيات رؤية الكمبيوتر.

1. INTRODUCTION

The metaverse is a new era in technology and social communication which offers a parallel universe and provides access to multitudes of 3D virtual spaces and environments created by users. In fact, virtual environments are vital components in the metaverse, with an effect over various aspects ranging from environment architectures to personal interactions.

The latest advancements in Virtual Reality tools are convincing users to increasingly be immersed in virtual worlds. It takes control of their vision and provides them with hotspots to interact with virtual objects or to navigate in the virtual environment. The hotspots can also ruin the visual display or cause some distraction when not implemented correctly.

Accordingly, one of the most important modules within virtual environments development is hotspot creation. Among this process, the creator of the virtual environment can set interactive mediums for visitors to interact with the virtual environment and launch predefined events and actions. These hotspots attract the navigators of these environments and show them some information and definitions about semantic elements within it and facilitate their navigation between the virtual spaces. (Napolitano, Blyth and Glisic, 2018; Eiris, Wen and Gheisari, 2020).

Current techniques for hotspot creation are implemented using manual localization by the users during the creation of the environment and are updated upon need. This can be considered a highly hectic process for the users and might even result in missing some important hotspots.

Certain virtual spaces within virtual environments contain similar hotspots which the virtual environment creator must separately define. For example, when the virtual environment creator is working on a 360^{O} virtual classroom that contains 20 chairs, 20 hotspots must be localized and defined by the user. However, the virtual environment creator could miss defining some hotspots on some objects in the environment which might cause visitors to wonder, such as a certain class chair in the virtual classroom we mentioned, or any other object (i.e. door) in a virtual room.

In this paper, a fully automatic hotspot creation tool is proposed. The virtual spaces in the virtual environments are represented by 360^{O} panoramic images. This tool scans these images when the virtual environment creator is building the virtual environment so that required hotspots can be automatically localized and set, or at least proposed. This will facilitate the development phase of the virtual environment, and aid in mitigating the negative impact or unintended consequences from missing essential hotspots. However, architects can then be immersed in these environments and benefit from this tool to design and manage the assets in their intended virtual spaces (Zhang *et al.*, 2011; Schnabel and Kvan, 2003).

Most of the Integrated Development Environments (IDE) that are used to build VR scenes provide the feature of manually adding interactive hotspots. The focus in our work is to propose such automated hotspots creation and proposition tool that could be integrated in any of the interesting VR-IDEs.

The rest of the paper is organized as follows. In section 2, the work related to the hotspot creation system is summarized. The pipeline used is then explained in section 3 where details about each component used are shown. Experimental results follow in section 3 where the dataset used to verify the hotspot creation approach is discussed and sample results are shown. Finally, the conclusion and future work are included in section 4.

2. RELATED WORK

The virtual environment is composed of a set of scenes having defined entities. The creator of this environment works then on connecting these scenes and enabling interactions on these entities through hotspots (Cantatore, Lasorella and Fatiguso, 2020; Lanzieri *et al.*, 2021; Shah *et al.*, 2021). Additional techniques are currently being used for adding hotspots, like in https://kuula.co/. They provide creators of virtual environments with some tools that allow them to select multiple hotspots for updating or duplication abilities.

One of the modules used in the hotspot creation is object detection. In general object detection is a widely studied area in computer vision and it started with basic hand-crafted image features such as HOG (Wang, Han and Yan, 2009) coupled with supervised learning and basic classifiers. Advances in deep learning and convolutional neural networks (Girshick, 2015; He et al., 2017; Huang et al., 2018) in specific gave rise to a leap in object detection where more advanced detectors are able now to give state of the art performance and succeed at identifying a huge range of objects. Recently, object detection in 360º images i.e. panoramic object detection is now being studied where Deng, Zhu and Ren (2017) use a convolutional neural networks on panoramic images. In addition, another relevant work by Zhang et al. (2014) includes geometric understanding of the scene to allow object detection in such images.

In general, object detection performance is highly affected by the training data. Datasets (Deng, 2009; Everingham, 2010; Veit, 2016) found in the computer vision community are based on two dimensional images and most object detection frameworks are built on such images. The Sun360 dataset (Xiao, 2012) handles this issue and offers a collection of panoramic images with ground truth object detection labels which would be most relevant to the meta verse environment. An extension to this dataset, which is considered highly relevant to hotspot creation, was also proposed in (Guerrero-Viu, 2020) where scene segmentations are added including tight outlines of objects and scene background such as floor and wall.

Another main module in the hotspot creation pipeline is image segmentation. Early methods in image segmentation use basic image processing techniques such as thresholding (Otsu, 1979), region growing (Nock and Nielson, 2004), k-means clustering (Dhanachandra, Manglem, and Chanu, 2015). More advanced techniques such as graph cuts (Vicente, Kolmogorov, and Rother, 2008) were also used in the liteature. Recent advances in image segmentation used deep learning models (Chen et al., 2017; Long, Shelhamer and Darrell, 2015; Noh, Hong and Han, 2015) which resulted in remarkable performance improvements. A recent survey on deep learning methods for image segmentation by Minaee et al. (2021) gives a detailed summary of these methods.

3. HOTSPOT CREATION APPROACH

The approach that is used for the creation and localization of hotspots can be summarized in Figure 1. As can be shown in the image, the input to the pipeline is a 360º image panorama taken at a specific location within the virtual environment. The panoramic image is first projected into perspective view. Each view is then input into a deep network for object detection where various semantic elements are localized. In addition, the panoramic image is segmented. Finally, the candidate objects and segmentation map are combined to localize and create hotspots in the panoramic 360 image. These hotspots can thus be integrated into the virtual environment. The following sections describe in more details the various components used within the pipeline.

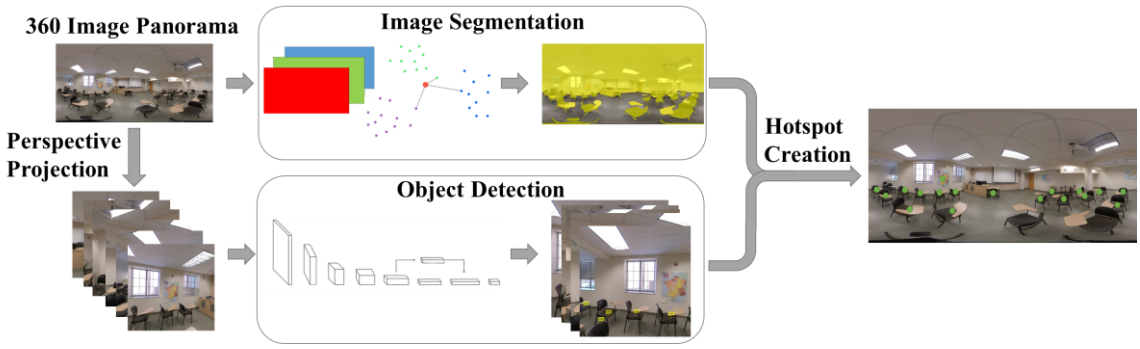

Fig.1: Hotspot Creation Pipeline

3.1. Perspective Projection

The initial step in the hotspot creation is perspective projection of the input 360 image panoramas. Most generic object detection approaches apply object detection on images taken by a normal camera. Thus, one critical step to be implemented is applying equirectangular projection on these images to obtain the perspective view. The equirectangular projection is applied as a transformation that warps the input panorama by identifying specific tilt, roll and pan angles in addition to field of view. The transformation is produced by computing the camera calibration matrix K and the rotation matrices referring to the pan θ and tilt φ respectively (roll is ignored here since it's constant for panoramic images).

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = K^{-1} R_\theta R_\phi \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

The above equation shows how each pixel (x, y, z) is transformed into perspective view as a multiplication of the calibration and the rotation matrices. Finally, the longitude and latitude are calculated from the projected homogeneous coordinates. For each panoramic image, P+T perspective images are created referring to P different pan angles and T different tilt angles. Each of these images are input into the object detection step described in the next section.

3.2. Object Detection

The object detection step in the hotspot creation pipeline is applied on the perspective view images extracted from the panoramas. Each perspective image is input to a pre-trained deep network that allows the detection and localization of various semantic elements such as chairs, sofas, people, etc... The object detection framework used is the YOLO (You Look Only Once) detector (Redmon and Farhadi, 2017) which is considered the current state-of-the-art object detector.

The YOLO object detector applies a single deep network to an input image by dividing the image into regions. It uses a convolution neural network consisting of a total of 24 convolutional layers with different functionalities and followed by 2 fully connected layers. The first 20 layer are used for pre-training and the last 4 layers are specified for object detection. The final layer predicts bounding boxes that are weighted by predicted probabilities.

Figure 2 below shows example object detection output on two perspective images. The above images produce (x', y') locations of detected objects which are mapped to their corresponding locations in the panoramic 360 image for further processing in the hotspot creation step. In addition to the object locations, a confidence score is also given which gives a probability value for the certainty of the identified object.

Fig.2: Object detection on two perspective images. Left image shows the detected chair and sofa objects. Right image shows the detected TV monitor.

3.3. Image Segmentation

The image segmentation step gives an overall understanding of the scene within the virtual environment. It basically divides the input image into a set of segments which gives a simple representation of the main areas that have similar color and texture within the image. Thus, the whole panorama is input into the segmentation algorithm and a mask that maps each pixel to a segment is produced. The image segmentation technique chosen is the k-means clustering technique due to its simplicity and speed. The *k* in this technique refers to the number of segments to be formed. K-means clustering is an iterative approach for image segmentation that takes the R-G-B values of each pixel in the image and assigns them to cluster centers. The segmentation is done in two main steps which are formed of cluster assignment and cluster center update steps. Once the pixels are mapped to the cluster centers a segmentation mask is created.

The segmentation of the scene gives rise to identifying various segments within the panoramic image and thus allows the identification of semantic regions referring to the floor, ceiling, and walls in the scene. These regions are considered vital for adding navigation hotspots in panoramic images especially at location with doors and entrances are located.

Figure 3 below shows an example panoramic image with the corresponding segmentation. As shown in the figure the panoramic scene is divided into 4 segments. These segments can be identified from the scene as wall, floor, chairs and ceiling. It is important to note here that the floor part is in general located in the bottom part of the image assuming that panoramas are taking at an eye-level latitude. The floor segment is considered primarily vital for hotspot localization which gives candidate location for the navigation of the virtual environment and the transition between the scenes.

3.4. Hotspot Localization

The final step in the hotspot creation approach is hotspot localization. The detected object locations, scores, and labels in addition to the scene segmentations are combined to output hotspot locations.

The object detection step outputs the labels and locations of candidate hotspots (refer to bottom left images in Figure 4). These labels when combined will refer to multiple locations within the panoramic image. In fact, the same semantic element might be present in multiple perspective images which results in duplicated overlapping detections when mapped to the input panoramic image.

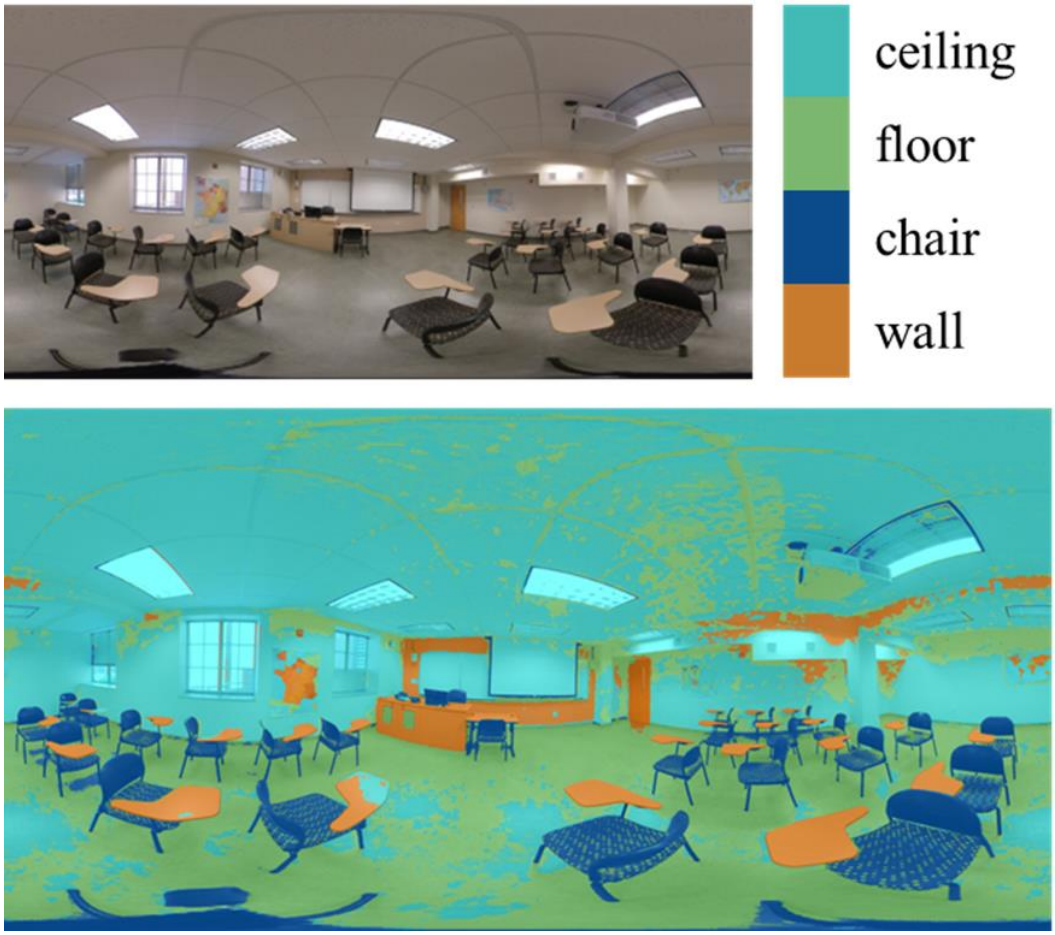

Fig.3: Image segmentation with k-means clustering algorithm using k=4 segments. Each segment refers to different semantic elements in the scene.

In order to handle this issue, the score of each detection is considered and then non-maximum suppression is applied to keep the detection with the highest score within overlapping detections. This results in one candidate hotspot for each semantic element as shown in Figure 4 bottom right image.

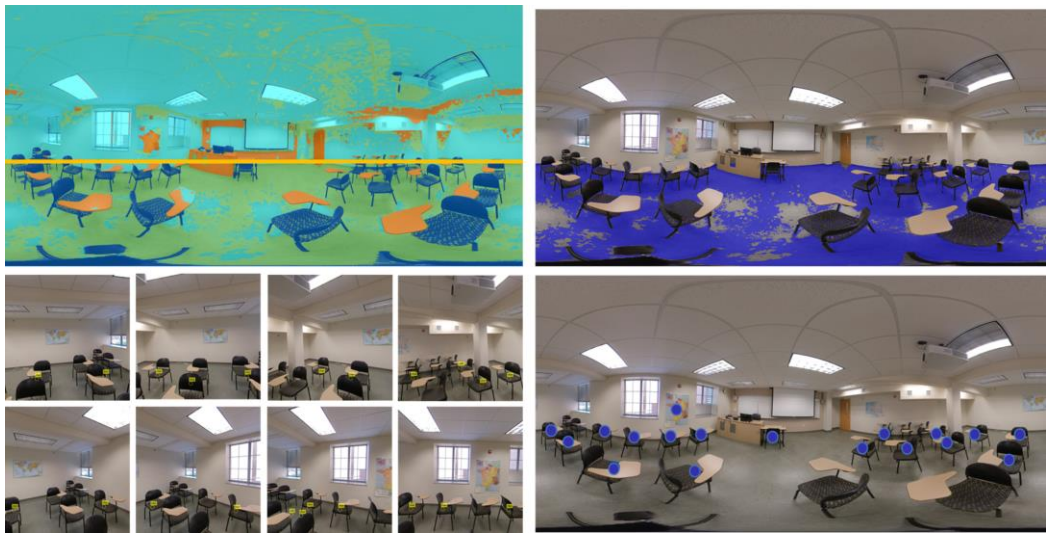

Fig.4: Floor segmentation for navigation space and hotspot localization on detected chair objects

The image segmentation also provides vital information about the semantic elements in the scene. One semantic element vital for navigability is the floor. In order to detect the

floor, the segmentation map is further processed by filtering out pixels that fall in the bottom part of the image. Out of these pixels, the segment with the majority number of pixels is picked as floor and is thus segmented out of the image. The top row in Figure 4 shows the segmented floor in the sample input scene. As shown in the figure the yellow line separates the segmentation and the figure on the right shows the segment with majority of pixels. This segment maps to the floor of the detected segments. The localization of the separation line is computed by identifying the y-value that achieves the maximum difference between the segmentation labels in the y-direction and in the same time being in the bottom half of the image.

4. EXPERIMENTAL RESULTS

4.1. Dataset

The metaverse is formed of various virtual environments which the users can navigate. The environments can thus either be indoor or outdoor with completely different semantic elements. For example, chairs, sofas and TV monitors are usually seen more often within indoor scenes while trees, cars and road signs are seen in the outdoor. In addition, the scenes navigated by the users can vary from realistic to synthesized scenes where a realistic scene in general is formed of panoramic images usually taken by a 360^o camera. We test our framework on both indoor and outdoor realistic images. The indoor images are taken from the Sun360 dataset while the outdoor images are taken from Street View images. Figure 5 shows the detected results on sample images from these datasets.

4.2. Hotspot creation results

The hotspot creation approach contains some parameters related to number of perspective angles P and T, number of clusters k, and separation line localization. Upon extensive validation, the chosen values for each parameter are listed below.

- $P=36$: pan angles start at 0^o and end at 360^o with 10^o increments
- $T=3$: tilt angles are -20^o, 0^o, 20^o
- $k=4$: the number of segments chosen is 4 which offers enough variety for segment types
- $y0 = 0.6 \times h$: the separation line falls in the bottom 40% part of the scene

In addition, the object classes that we use uses the pretrained YOLO network on the Pascal VOC data set, which contains images from 20 different classes. These classes include chair, sofa, TV monitor, person, car, dining table labels in addition to others. Of course, additional object classes can be incorporated to be trained on the YOLO network which would result in more sematic elements identified such as doors and road signs. We show how the results from the 20 classes are obtained and this could be easily generalized on more class labels.

Figure 5 shows the output results of the hotspot creation pipeline on 4 indoor images and 1 outdoor image. As shown in the figure, various hotspot locations are added into the panoramic image based on the object detection results. The locations identified refer to chairs, sofas and TV monitors in the indoor images since these are the labels on which the network is trained. As for the outdoor image, the hotspot locations refer to cars and people. These localizations suggest that the user navigating the scene can interact with these hotspots. For example, they could sit on one of the chairs and get offered a different view of the classroom. The user could also turn on or off the TV monitor. In addition, these hotspots could be vital for the creators of the virtual environment where they could add more panoramic images at the identified locations since they are considered locations of interest. Moreover, segmentations of floor maps are identified. These segmentations allow the automatic navigability of the environments by showing candidate regions for navigation hotspots.

5. CONCLUSION AND FUTURE WORK

In this paper, an automatic hotspot creation approach is proposed which combines two main computer vision tools: object detection and scene segmentation. The proposed pipeline is fully automatic, where for a given 360º panoramic image, candidate hotspots are identified and can be incorporated within the virtual environment to be later used by either the navigator or the creator of the environment. Automated hotspot creation has not been studied in the literature and thus this tool can be considered the first to handle such a problem. In the future, various enhancements can be added to the proposed pipeline. Training deep networks on 360º images with candidate hotspot locations can lead to an end-to-end approach to hotspot creation. In addition, the proposed hotspots can be incorporated within advanced tools in virtual environments which can help in automatically creating spaces and localizing additional objects within architectural scenes.

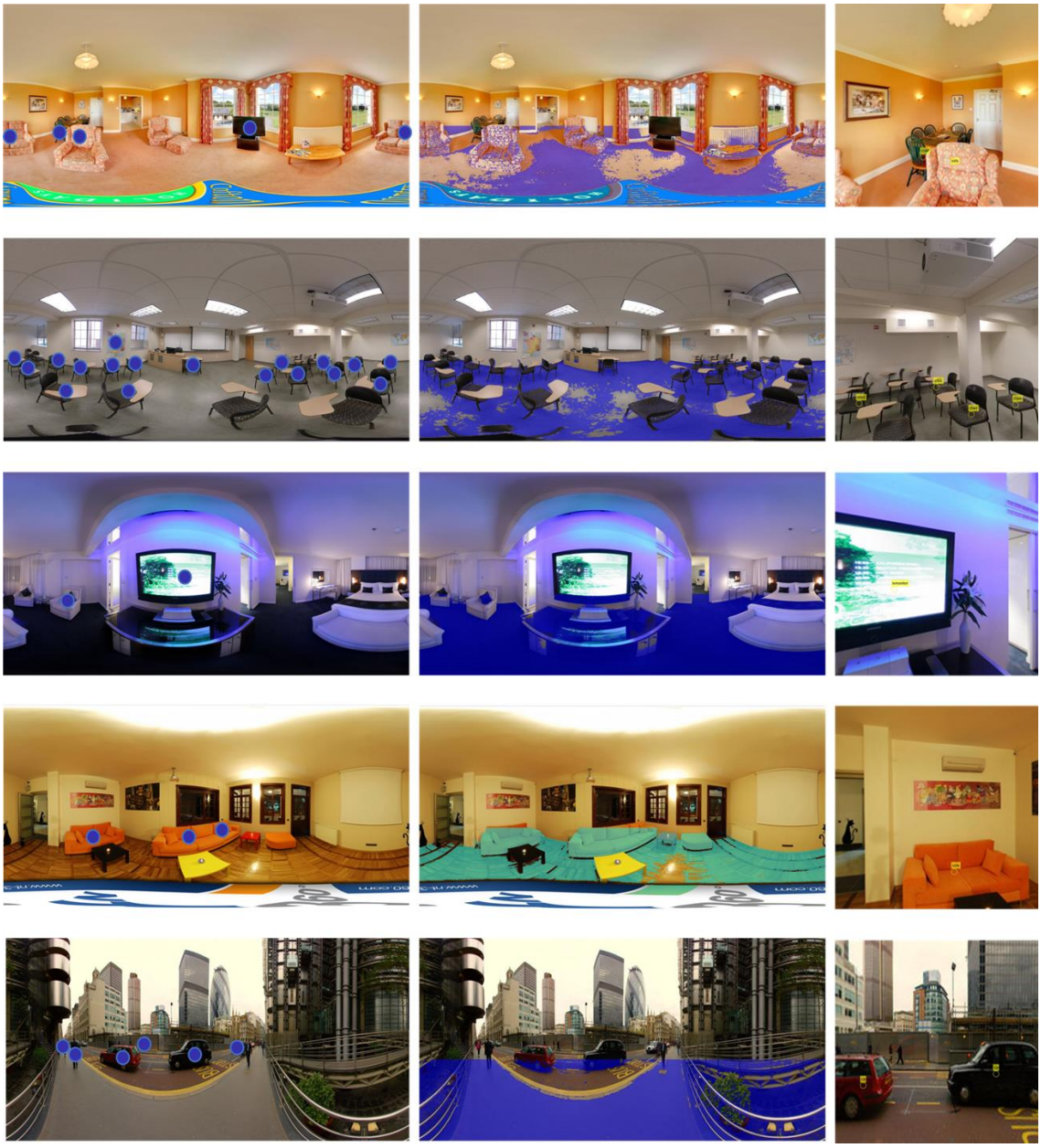

Fig.5: Hotspot creation results on sample indoor and outdoor images. The left column shows the localized hotspots. The middle column shows the segmented navigation space. The right image shows sample object detection(s) from a projected perspective image.

REFERENCES

- CANTATORE, E., LASORELLA, M. AND FATIGUSO, F., 2020. Virtual reality to support technical knowledge in cultural heritage. The case study of cryptoporticus in the archaeological site of Egnatia (Italy). *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *44*, pp.465-472.
- CHEN, L.C., PAPANDREOU, G., SCHROFF, F. AND ADAM, H., 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.
- DENG, F., ZHU, X. AND REN, J., 2017, April. Object detection on panoramic images based on deep learning. In *2017 3rd international conference on control, automation and robotics (iccar)* (pp. 375-380). IEEE.
- DENG, J., DONG, W., SOCHER, R., LI, L.J., LI, K. AND FEI-FEI, L., 2009, June. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee.
- DHANACHANDRA, N., MANGLEM, K. AND CHANU, Y.J., 2015. Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, *54*, pp.764-771.
- EIRIS, R., WEN, J. AND GHEISARI, M., 2020, November. iVisit: Digital interactive construction site visits using 360-degree panoramas and virtual humans. In *2020 ASCE Construction Research Congress (CRC)*.
- EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C.K., WINN, J. AND ZISSERMAN, A., 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, *88*(2), pp.303-338.
- GIRSHICK, R., 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).
- GUERRERO-VIU, J., FERNANDEZ-LABRADOR, C., DEMONCEAUX, C. AND GUERRERO, J.J., 2020, May. What's in my room? object recognition on indoor panoramic images. In *2020 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 567-573). IEEE.
- HE, K., GKIOXARI, G., DOLLÁR, P. AND GIRSHICK, R., 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961-2969).
- HUANG, R., PEDOEEM, J. AND CHEN, C., 2018, December. YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 2503-2510). IEEE.
- LANZIERI, N., MCALPIN, E., SHILANE, D. AND SAMELSON, H., 2021. Virtual reality: an immersive tool for social work students to interact with community environments. *Clinical Social Work Journal*, *49*(2), pp.207-219.
- LONG, J., SHELHAMER, E. AND DARRELL, T., 2015. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).
- MINAEE, S., BOYKOV, Y.Y., PORIKLI, F., PLAZA, A.J., KEHTARNAVAZ, N. AND TERZOPOULOS, D., 2021. Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence.
- NAPOLITANO, R., BLYTH, A. AND GLISIC, B., 2018. Virtual environments for visualizing structural health monitoring sensor networks, data, and metadata. *Sensors*, *18*(1), p.243.
- NOCK, R. AND NIELSEN, F., 2004. Statistical region merging. *IEEE Transactions on pattern analysis and machine intelligence*, *26*(11), pp.1452-1458.
- NOH, H., HONG, S. AND HAN, B., 2015. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE international conference on computer vision (pp. 1520-1528).
- OTSU, N., 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, *9*(1), pp.62-66.
- REDMON, J. AND FARHADI, A., 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7263-7271).
- SCHNABEL, M.A. AND KVAN, T., 2003. Spatial understanding in immersive virtual environments. *International Journal of Architectural Computing*, *1*(4), pp.435-448.
- SHAH, H., SHAIKH, S., TUPE, V., RATHOD, A. AND UKE, N., 2021. Development of Virtual Environment for Educational Campus. *Pensee Int. J.*, *51*(4), pp.1415-1421.
- VEIT, A., MATERA, T., NEUMANN, L., MATAS, J. AND BELONGIE, S., 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*.
- VICENTE, S., KOLMOGOROV, V. AND ROTHER, C., 2008, June. Graph cut based image segmentation with connectivity priors. In *2008 IEEE conference on computer vision and pattern recognition* (pp. 1-8). IEEE.
- WANG, X., HAN, T.X. AND YAN, S., 2009, September. An HOG-LBP human detector with partial occlusion handling. In *2009 IEEE 12th international conference on computer vision* (pp. 32-39). IEEE.
- XIAO, J., EHINGER, K.A., OLIVA, A. AND TORRALBA, A., 2012, June. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2695-2702). IEEE.
- ZHANG, L., GOSSMANN, J., STEVENSON, C., CHI, M., CAUWENBERGHS, G., GRAMANN, K., SCHULZE, J., OTTO, P., JUNG, T.P., PETERSON, R. AND EDELSTEIN, E., 2011, November. Spatial cognition and architectural design in 4d immersive virtual reality: Testing cognition with a novel audiovisual cave-cad tool. In *Proceedings of the Spatial Cognition for Architectural Design Conference*.
- ZHANG, Y., SONG, S., TAN, P. AND XIAO, J., 2014, September. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *European conference on computer vision* (pp. 668-686). Springer, Cham.