A Multi-Model Approach for Comprehending Human Behavior and Stories in Space

Chi-Li Cheng

Video name: Demonstration of Real-Time Scene Estimation Tool in Rhino Grasshopper

Descriptions:

The tool integrates with 3D modeling software, Rhino Grasshopper, and Python. It allows users to tag tokens representing the activities of people in 3D models and provides real-time behavior estimations that update as the environment is modified. This facilitates iterative design refinement while considering the impact on people's behavior.

Abstract

Understanding human social interactions in open spaces is crucial for architecture, urban design, and sociology. Videos provide valuable insights into human behavior by capturing spatial sense, events, and immersive experiences. They inspire designers by offering a comprehensive understanding of spatial elements, events, and people's activities in open spaces. However, traditional methods of behavioral analysis using observers are expensive and time-consuming. AI and computer vision-based approaches show promise, but they require high-quality video footage and have limitations in analyzing behavior over time. To fully leverage computer vision for promoting human socialization in design, a computer tool capable of comprehending and interpreting complex visual data is necessary. This tool should mimic human abilities in understanding behavior from imagery, bridging the gap between video analysis and design applications. This project proposes a multi-model approach to gain essential information for analyzing video in terms of human behavior in a space. By using high-level information, the approach can comprehend and estimate human behavior and scenarios effectively while avoiding data redundancy. This approach is suitable for creating a fine-tunable computational design tool (See Figure 1) that facilitates architectural, urban, and plot scene design.

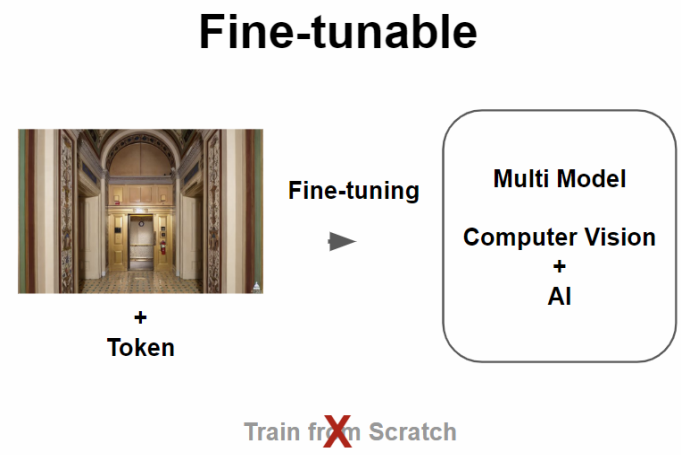


Figure 1

This research demonstrates the Multi-Model Approach by utilizing clips from William H. Whyte's documentary "The Social Life of Small Urban Spaces" (1980) and the movie "The Grand Budapest Hotel" (Wes Anderson, 2014) (see Figure 1). The proposed method will process the videos and categorize scenes based on spatial and social behavior features. A scene-searching tool will be developed to demonstrate the proposed multi-model method, assessing its ability to find similar scenes suitable for given spaces. Additionally, a Real-Time Scene Estimation tool in Rhino Grasshopper will be proposed to showcase the potential of the Multi-Model Approach in inspiring architectural design, facilitating social behavior studies, and advancing AI generative design.

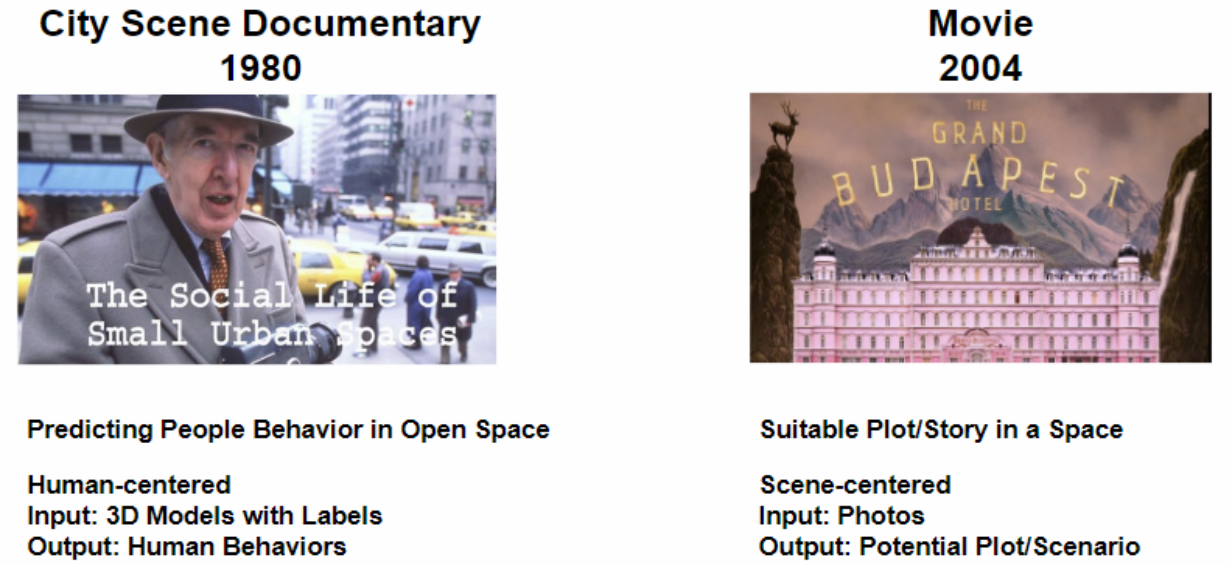


Figure 1

We utilize video from YouTube, extracting frames and captions, as depicted in the following image (See Figure 2).

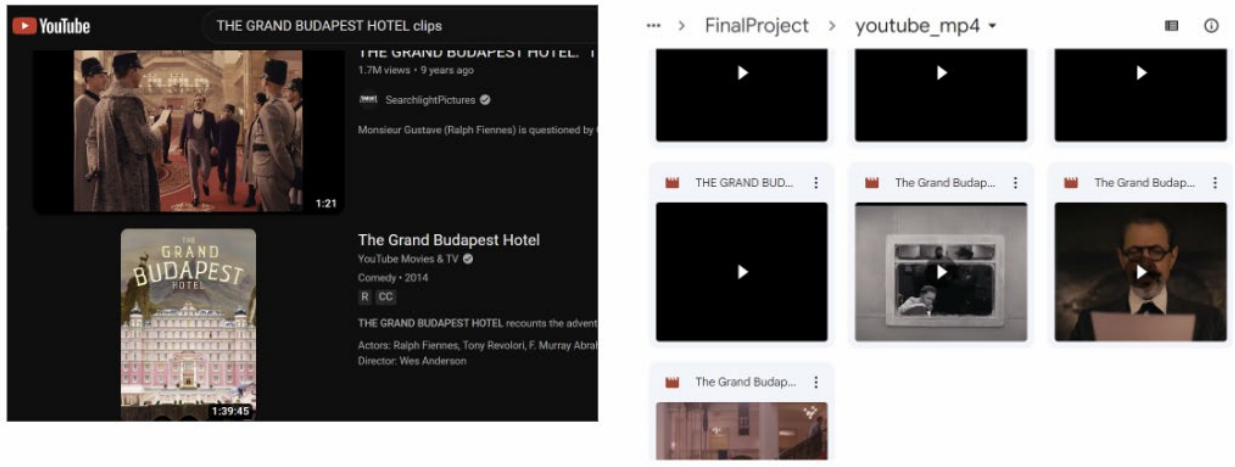


Figure 2

Multi-Model Pipeline

This paper proposes a pipeline for extracting essential information about spatial components and individuals in open environments from video frames and captions. The pipeline consists of several phases of data processing, where the system identifies, groups, and recognizes individuals in a scene using pre-trained models such as depth estimation (Ranftl et al. 2021), action estimation (Monfort et al. 2020), semantic segmentation (Zhou et al. 2017), and object detection (Redmon et al. 2016) models. Additionally, the pipeline recognizes and measures neighboring spatial components based on their proximity to the groupings.

These aforementioned data will be normalized and formed as a dataset for training a neural network. The training goal is to predict the activities of the group of people; therefore, the activity data is used for the output part. Since most of the work for comprehending high-level information is processed by the pre-trained models, the last task is relatively simple. The Model consists of only two hidden layers; as a result, it can run efficiently, which is suitable for being a module for developing design tools. In the following paragraphs, two usages are proposed.

Estimation of Human Behavior in Open Space & Estimation of Scene

This project introduces a user-friendly tool that estimates the behavior and social interactions of individuals within a design space. It seamlessly integrates with Rhino Grasshopper and Python, allowing users to tag tokens representing potential activities of people as agents within 3D models. Real-time behavior estimations continuously update as users modify the 3D environment, enabling them to refine their designs while considering the impact on people's behavior. The following image (Figure 3) illustrates distinct predictions resulting from different designs. Additionally, the tool can be used to estimate the best-fit plot and scene for a given space using an input image, showcasing its versatility.

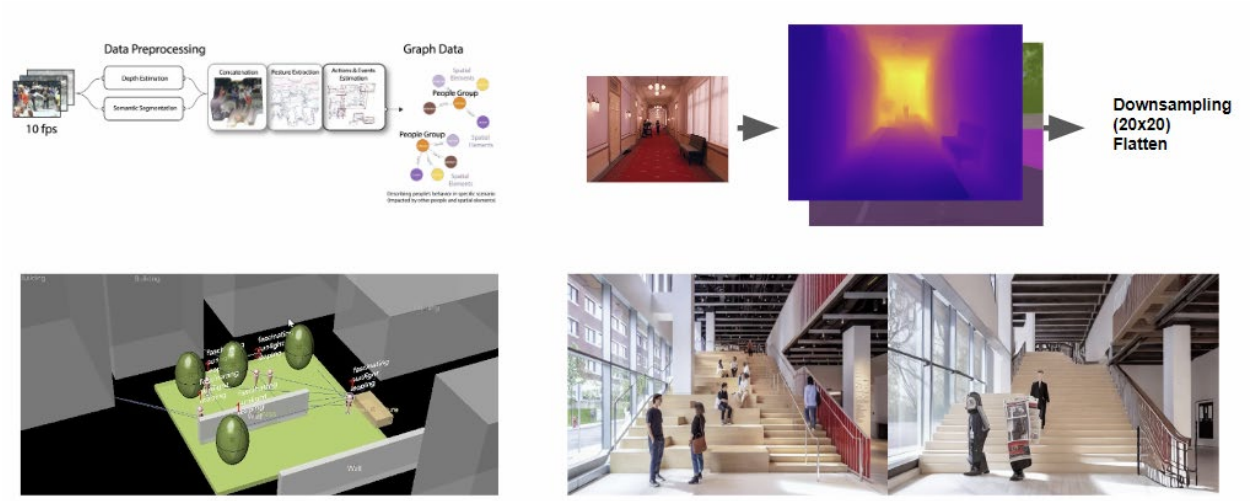


Figure 3

Estimation of Human Behavior in Open Space

The system's objective is to comprehend human behavior in a specific space and utilize this understanding to predict behavior in a new design through the analysis of 3D models. It comprises two primary components (refer to Figure 4):

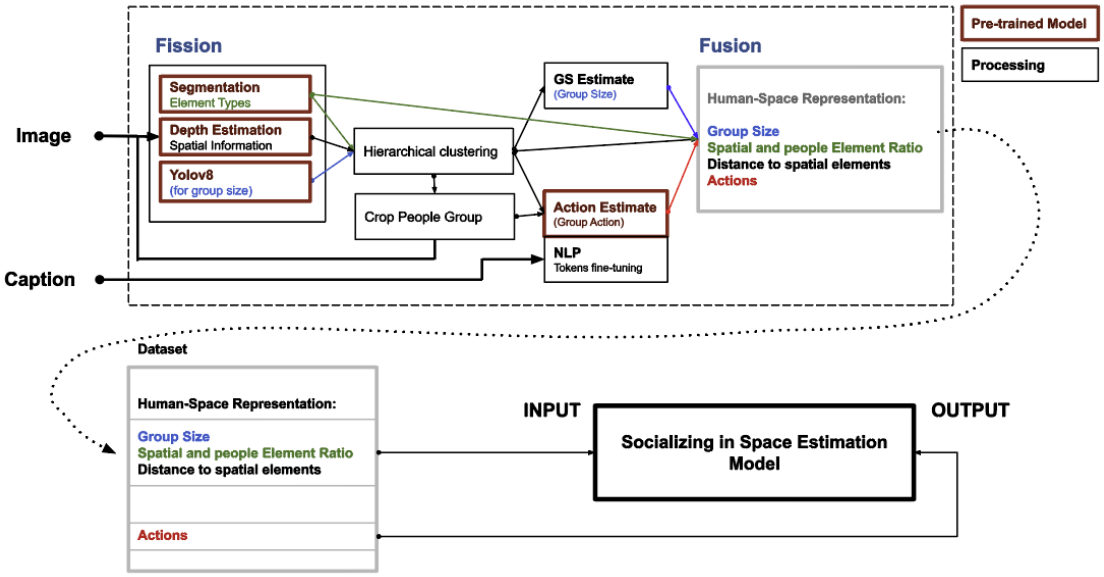


Figure 4

The first component adopts a multi-model approach to extract essential and high-level information from provided images and video captions, forming a dataset. This dataset serves as input for the second component. The second component involves training a simple neural network that estimates people's behavior within a 3D modeling environment (refer to Figure 5). The neural network utilizes the dataset created in the first part to make accurate predictions about human behavior in the given space. By combining these two components, the system effectively leverages visual and textual information to gain insights into human behavior and apply that knowledge to predict behavior in new design scenarios.

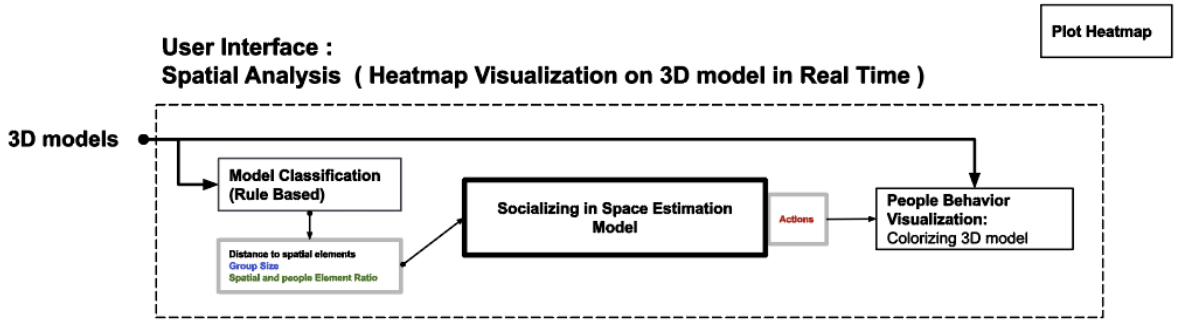


Figure 5

The system incorporates depth estimation to reconstruct the 3D scene from images, as it plays a crucial role in estimating people's behavior within a 3D environment. Additionally, a semantic segmentation model is employed to recognize objects and provide contextual information. This integration allows for the identification and classification of various scene elements, including people, buildings, trees, and street furniture. By capturing correlations between people's behavior and their surrounding spatial elements, a more comprehensive understanding is achieved. Combining the information gathered from the depth estimation and semantic segmentation steps, a 3D point cloud is generated, with each point assigned a corresponding label. To effectively organize and group these points, a hierarchical clustering process is applied. Once clustering is performed on group individuals, an action estimation model (specifically, Moments in Time-based) is employed to predict the actions of each group. The bounding box of the clustered points labeled as people is used for cropping the image, defining the region of interest for further processing. Subsequently, action estimation is performed on each cropped image (refer to Figure 6).

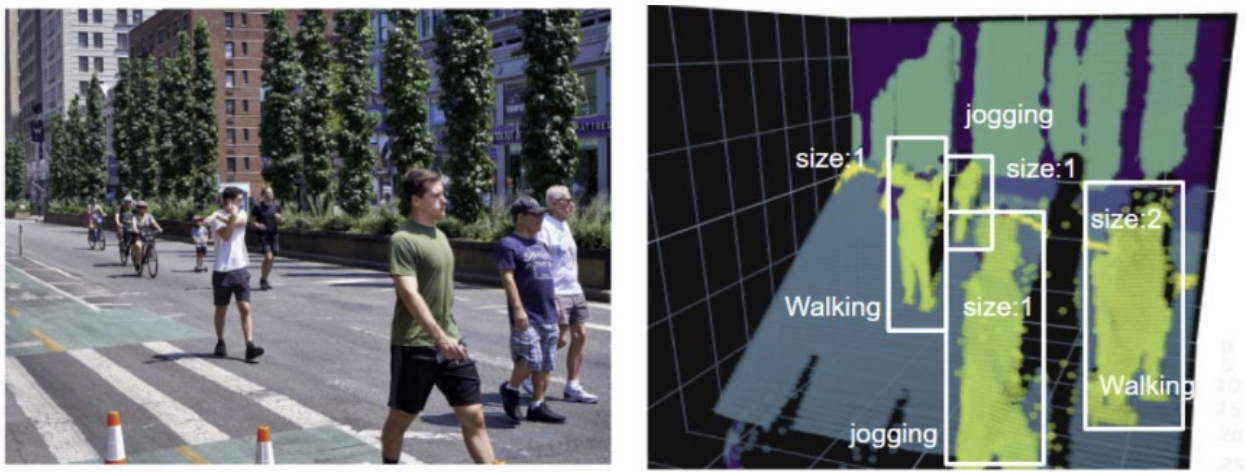


Figure 6

Furthermore, the system incorporates the YOLO (You Only Look Once) object detection algorithm to count the number of people within the cropped images, facilitating the determination of group size. To enhance the alignment between predicted tokens and the captions of the original video, dimension reduction using T-SNE and word embedding with GloVe 100D representation is employed. This approach enables the identification of the closest matching token from the video's captions, which aids in predicting actions. The captions for the video are generated using the Google Speech-to-Text API.

After the aforementioned steps, the system calculates the correlation between people groups and spatial elements by measuring the distance between them using their center points. For example, by observing a group of people engaged in a picnic activity at a short distance from a nearby tree, the system can learn the tendency of people to picnic near trees. The human-centered dataset (as shown in Figure 7) obtained from the project captures the physical environment surrounding a group of people, including their behavior (actions), group size, and the composition of the scene in terms of the percentages of each element present in the image. This comprehensive dataset recognizes that people's behavior can be influenced by factors such as group size and the spatial

elements within the scene. To train the model effectively, a dataset of 1000 data points was generated, ensuring a robust foundation for accurate behavior estimation.

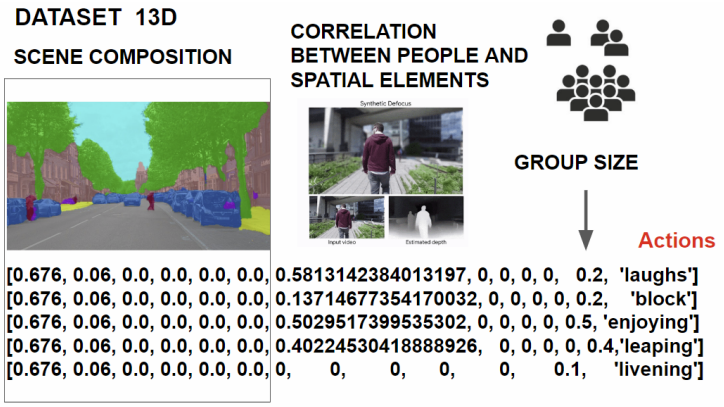


Figure 7

By following this process, the system successfully leverages depth estimation, semantic segmentation, hierarchical clustering, and action estimation to train a model for estimating people's behavior in a 3D environment accurately(see Figures 8 and 9).

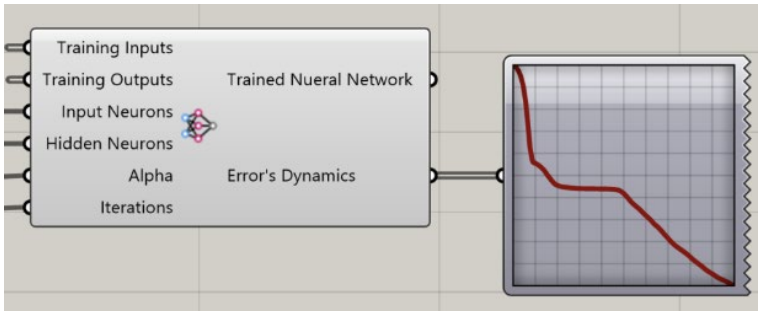


Figure 8

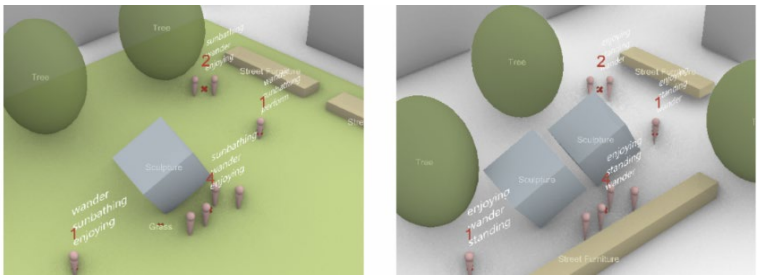


Figure 9

Estimation of Scene

In addition to estimating human behavior in open spaces, this project extends the multi-model approach to estimate spaces for specific movie plots and scenes. By utilizing clips from the movie and its screenplay document, the project aims to analyze the movie's image features and label the scenes, resulting in a dataset. Users can input images depicting particular spaces, and the system will identify similar spaces within the movie scene. The system then generates a description of the scene and employs a stable diffusion model to simulate how the plot unfolds within the inputted images. This enables visualization of how the plot is portrayed in those specific images, enhancing the understanding and analysis of the movie's spatial context.

In the Estimation-of-Scene system, the initial task is to label the scene automatically. This is accomplished by utilizing the Google-speech-to-text library to generate captions from the recorded footage. These captions are then used to identify specific lines and their corresponding scene descriptions(See Figure 10).

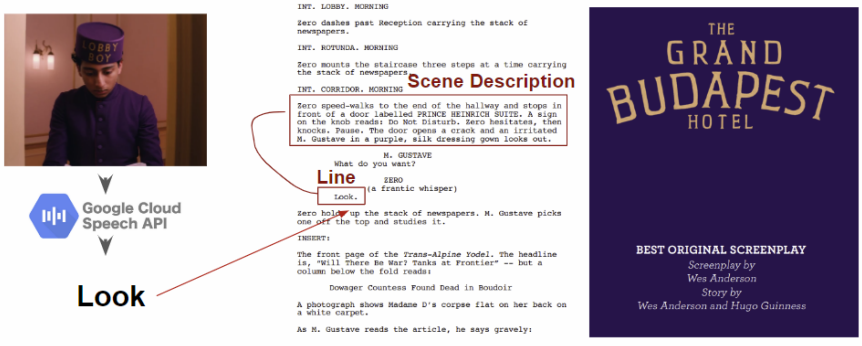


Figure 10

While the Estimation-of-Scene system utilizes the same set of pre-trained models, there is a variation in the representation of the space. The main distinction lies in enhancing the system's sensitivity to the scene's composition. To achieve this, the depth estimation is downsampled and transformed into a flattened 2D matrix, which is then utilized as the features. This approach enables a comprehensive representation of the movie scene's composition, as depicted in the accompanying Figures 11 and 12.

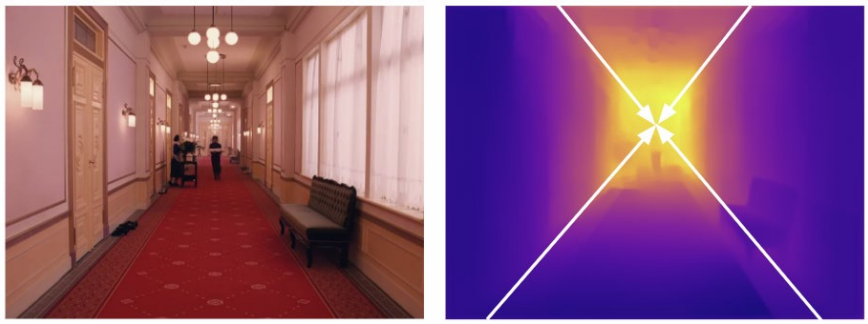


Figure 11

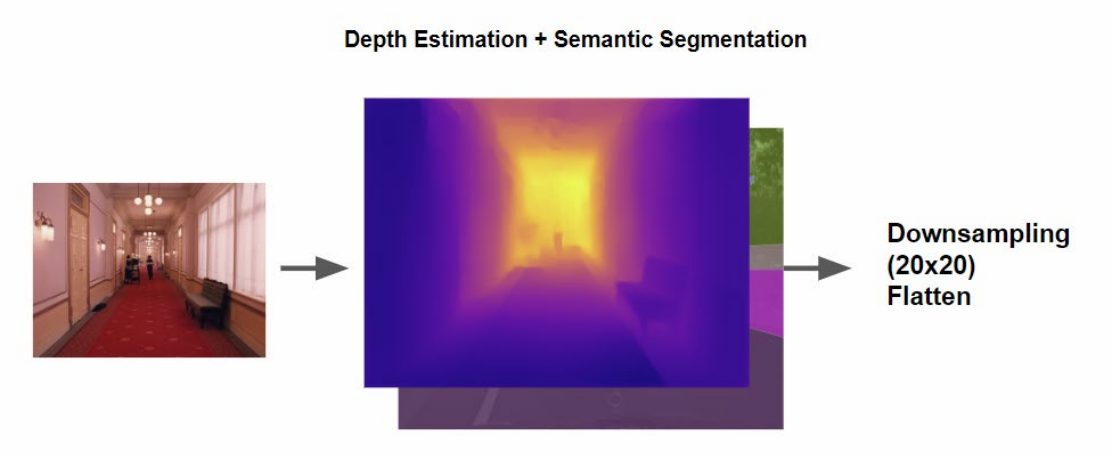


Figure 12

To identify a suitable scene, the system employs the T-SNE algorithm to reduce the dimensionality of the input image. Subsequently, it calculates the similarity between the reduced image representation and the dataset of scenes (See Figure 13).

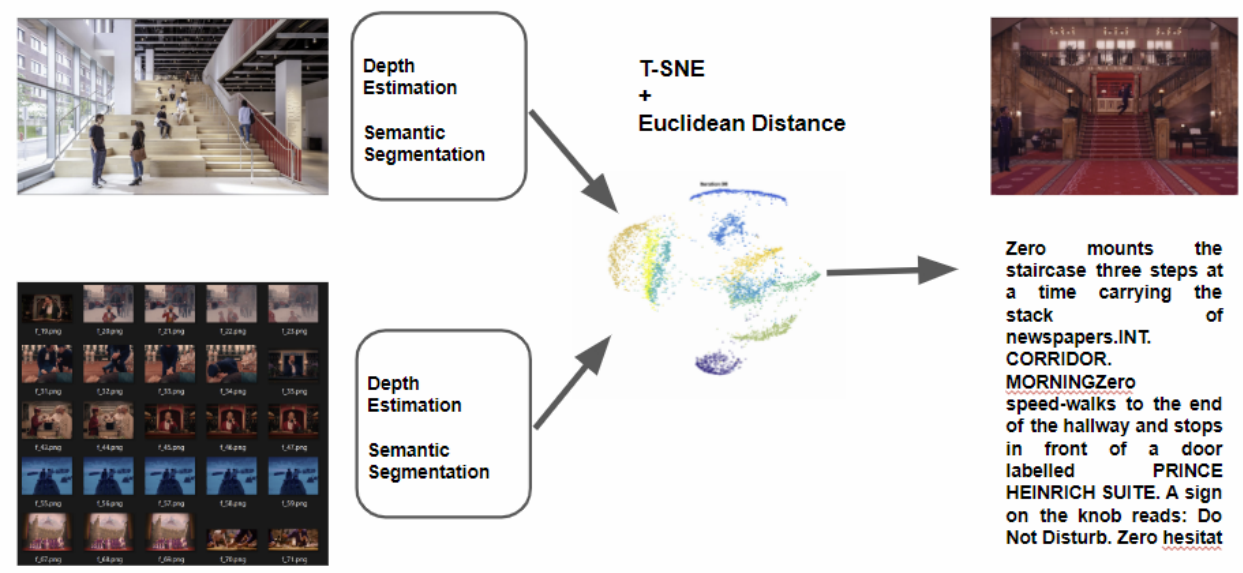


Figure 13

Testing the Estimation-of-Scene System

The initial case involves the input of an image depicting a hall featuring a large staircase within the premises of the MIT museum. A visual representation of the data processing steps corresponding to this case is illustrated in Figure 13.

Result

The subsequent result, as depicted in Figure 14, is produced utilizing a robust diffusion model. This model operates on the system's suggested scene description to generate the outcome.

Figure 14

Case of Frank Lloyd Wright's Ennis House

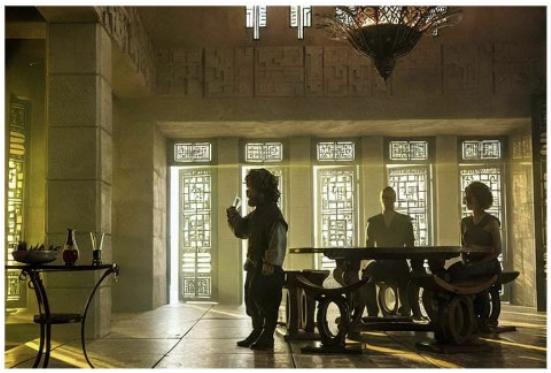
Processing

Frank Lloyd Wright's Ennis House

Dmitri and Jopling confer at a tense whisper. The sisters grumble, dismayed. The group as a whole sits up to attention. Deputy Kovacs slides a hand-written letter on pale-pink paper out of an envelope and reads:

Figure 15

Result



Dmitri and Jopling confer at a tense whisper. The sisters grumble, dismayed. The group as a whole sits up to attention. Deputy Kovacs slides a hand-written letter on pale-pink paper out of an envelope and reads:

Figure 16

Case of a Hotel Lobby

Processing

Dmitri fires three more times into the service elevator. The officer fires back at Dmitri. Dmitri ducks behind a room-service cart and quickly re-loads. More doors open up and down the corridor, and more armed officers in various states of dress/undress look ou

Figure 17

Result

Dmitri fires three more times into the service elevator. The officer fires back at Dmitri. Dmitri ducks behind a room-service cart and quickly re-loads. More doors open up and down the corridor, and more armed officers in various states of dress/undress look ou

Figure 18

Conclusion

This paper introduces an intelligent multi-model approach to develop a design tool that utilizes video data to comprehend human interactions within specific spaces. The potential of this intelligent multi-model strategy is demonstrated through two primary functionalities: real-time estimation of human behavior in open spaces and a scene estimation tool. However, the current version of the tool is still in its early stages and requires further improvements. Firstly, the system presently processes frames individually, which may pose challenges in detecting certain complex human behaviors that span multiple frames. Therefore, future development will focus on enhancing the capability to process sequential frames.

Additionally, an issue arises from the fact that the contents of frames within a video can vary significantly due to variations in the filming techniques employed by the cameraman. To address this issue, a data imputation mechanism can be introduced. This mechanism would enable the system to impute spatial information from adjacent frames when certain frames lack spatial information due to the presence of a crowded scene obstructing the background. By addressing these areas for improvement, the tool will become more robust and effective in its analysis and understanding of human behavior within different spatial contexts.

Bibliography

Anderson, Wes. (2014). The Grand Budapest Hotel.

Borah, C. 2021. What is the Social Impact of Architecture? RTF | Rethinking The Future. https://www.re-thinkingthefuture.com/2021/10/24/a5616-what-is-the-social-impact-of-architecture/

Monfort, M., Vondrick, C., Oliva, A., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., & Gutfreund, D. 2020. Moments in Time Dataset: One Million Videos for Event Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 502–508.

Ranftl, R., Bochkovskiy, A., & Koltun, V. 2021. Vision Transformers for Dense Prediction. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 12159–12168. https://doi.org/10.1109/ICCV48922.2021.01196

Whyte, W. H. (1980). The Social Life of Small Urban Spaces. Conservation Foundation.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. 2017. Scene Parsing through ADE20K Dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5122–5130.