Clustering Spatial Sequence

- A Deep Learning Investigation into Encoding and Clustering Spatial Sequences in Greek Temples using Depth Panoramas-

Tracy Miao

Final project for the MIT class 4.550/4.570 Computation Design Lab
Development: February-May, 2023
Instructor: Prof. Takehiko Nagakura, Dr. Daniel Tsai

Project Overview:

This project undertakes a deep learning investigation into encoding and clustering spatial sequences in three Greek temples. It aims to quantify and compare spatial experiences in different buildings.

The project initiates this investigation by simulating plausible pedestrian paths in the temples using the Grasshopper plugin Pedsim and employing a 3D isovist method to sample depth panoramas along these paths. This approach aims to represent sequential spatial experiences.

Two experiments are then conducted to cluster these panorama sequences. The first experiment uses a CNN network trained on predefined spatial typology labels, inspired by Peng's work on Machine's Perception of Space. The labels as a numeric array are reduced to 2 dimension through PCA, enabling a spatial cluster plot. However, this approach oversimplifies the visual features and treats the sequence as a static collection, neglecting its temporal aspect.

The second experiment remedies these limitations by using a ConvLSTM Autoencoder network that allows automatic label generation considering both visual and motion features. Unsupervised clustering algorithms is then applied to these latent features to generate labels and PCA algorithm is used as in the first experiment to observe clustering results. This data-driven approach allows for a more comprehensive, quantitative comparison of spatial sequences in different temples.

Videos:

Video1: Circulation Simulation
Description: The video demonstrates a circulation simulation using the PedSim plugin in Grasshopper. PedSim applies the social force model to guide pedestrians, considering target force, person repulsion force, and

obstacle repulsion force. They navigate from a start gate to a destination gate, avoiding obstacles and other individuals while optionally exploring points of interest. The simulation starts near the steps and terminates in the interior, with points of interest including columns, walls, and steps.

Video2: Sampling Process
Description: The video shows the process of sampling depth panorama along simulated path. Each path is sampled at 1 meter interval, which creates flexibility for resampling to adapt to different network structures. At each view point, a customized 3D isovist sampling script is deployed to map the depth values within a sphere onto a 60*30 depth panorama.

Figures:

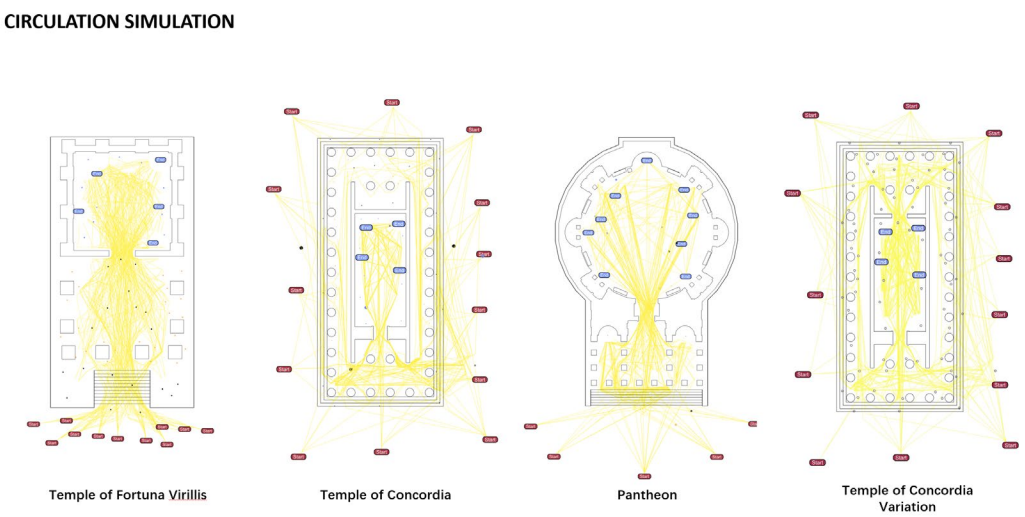

Fig1: Simulated Pedestrian Paths

Description:
In the case study, I choose 3 Greek temples – Temple of Fortuna Virillis, Temple of Concordia and Pantheon for their possessing both similarity and difference in typology. To further investigate the influence of small modification on plan, I created an opening in the back side of Cella in Temple of Concordia, thus resulting in a variation.

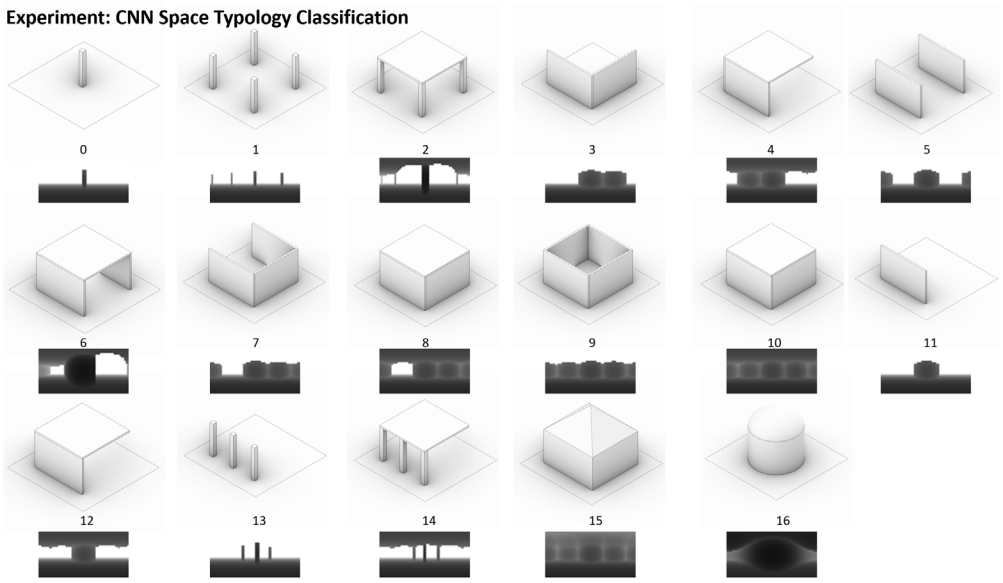

Fig2: 17 Spatial Typology Labels for CNN Classification and Corresponding Depth Panorama

Description:
In the first experiment, 17 predefined labels are used for CNN model to predict each frame's label along one sequence. 15 labels are inherited from Peng's paper Machine's Perception of Space, the last 2 labels are added to further differentiate the roof forms.

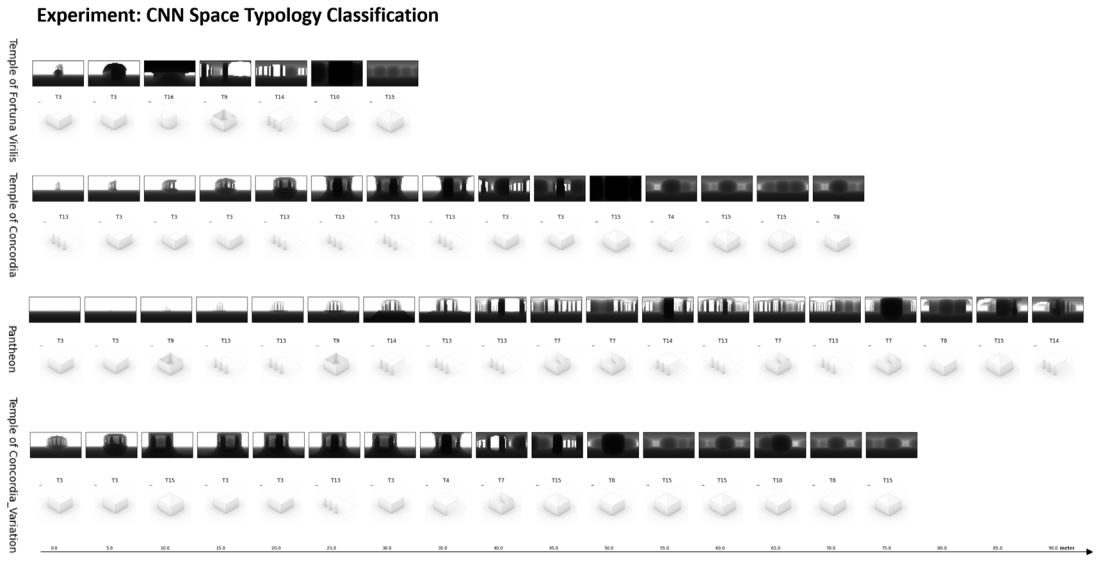

Fig3: Predicted Typology Labels Along One Sequence at 5-meter Interval

Description:

After multiple training iterations, the CNN model achieved a high prediction accuracy of 96.37% on the test set. Utilizing this trained CNN, I made predictions of typology labels along a sequence at 5-meter intervals and generated a comparative plot. The X-axis represents the walking distance from the start point, while each column corresponds to a specific distance and displays the depth panorama and predicted label. Additionally, the plot illustrates the variation in temples' scale.

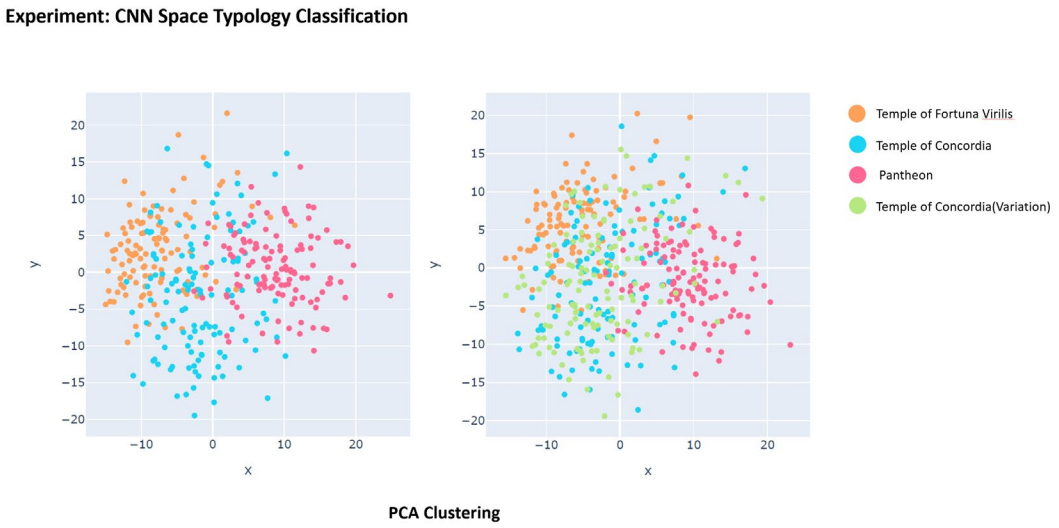

Fig4: PCA Clustering Result from Predicted Label Sequence

Description:
Using the trained CNN model, I predicted typology labels for all simulated sequences. Each sequence was resampled to 10 frames with equal intervals, resulting in a 10-label array. To visualize the data and observe clustering, I employed PCA to reduce the dimensionality from 10 to 2. The plot revealed three distinct clusters. The Temple of Concordia and its variations exhibited overlap, reflecting diverse circulation due to visitors approaching from different directions. The clustering result confirmed the preliminary hypothesis.

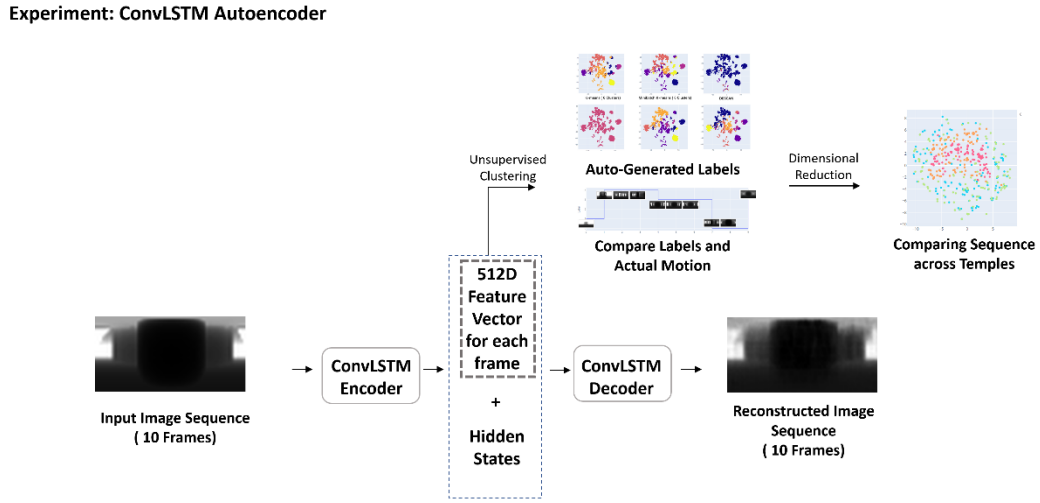

Fig5: Workflow of Second Experiment

Description:

In the second experiment, I addressed the limitations of the first by using a ConvLSTM Autoencoder structure. This allowed me to capture both visual and motion features from one frame in an encoded feature vector. I then applied unsupervised clustering to auto-generate labels for each frame. These generated labels were used to cluster spatial sequences across temples, improving the analysis compared to the first experiment.

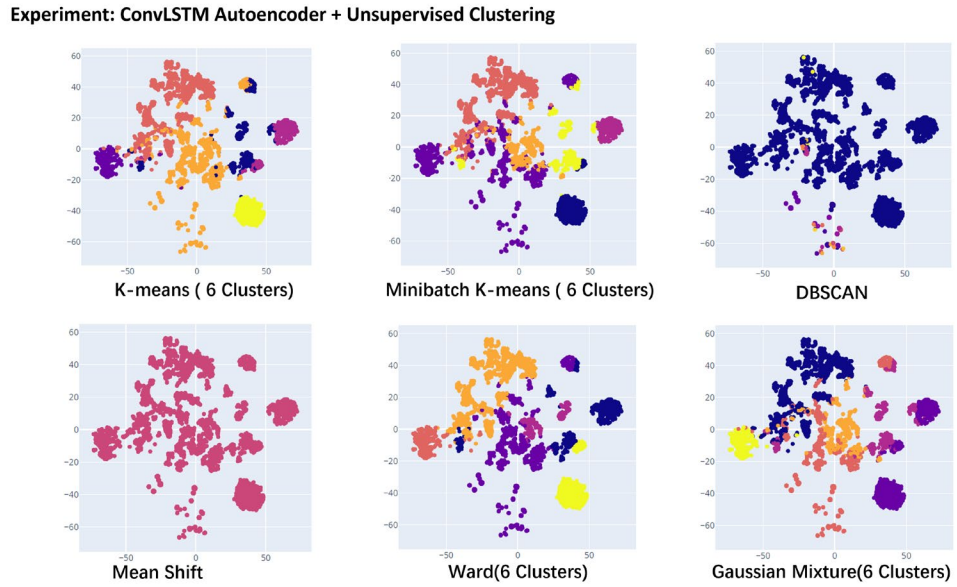

Fig6: Unsupervised Clustering Result of Encoded Feature Vector

Description:

In the second experiment, I replaced predefined labels with unsupervised

clustering algorithms to automatically cluster feature vectors derived from the ConvLSTM Autoencoder network. I tested 6 different clustering algorithms, using 6 clusters as input. Ultimately, I selected the result of the Ward algorithm for its clear outcome compared to the others.

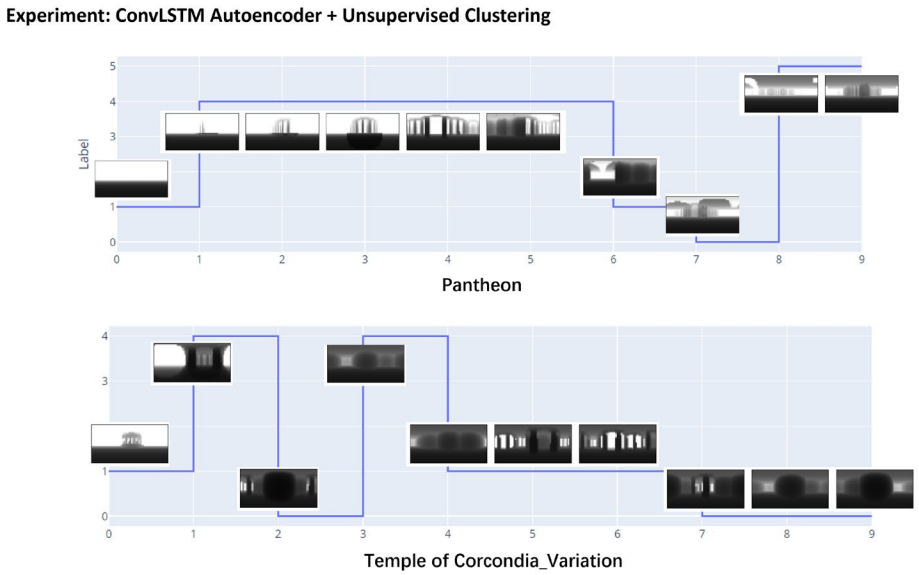

Fig7: Investigation of Auto-generated Labels and Corresponding Frame

Description:
The plot depicts how different labels relate to specific movements in the path, such as approaching the temple, entering the interior, and staying inside.

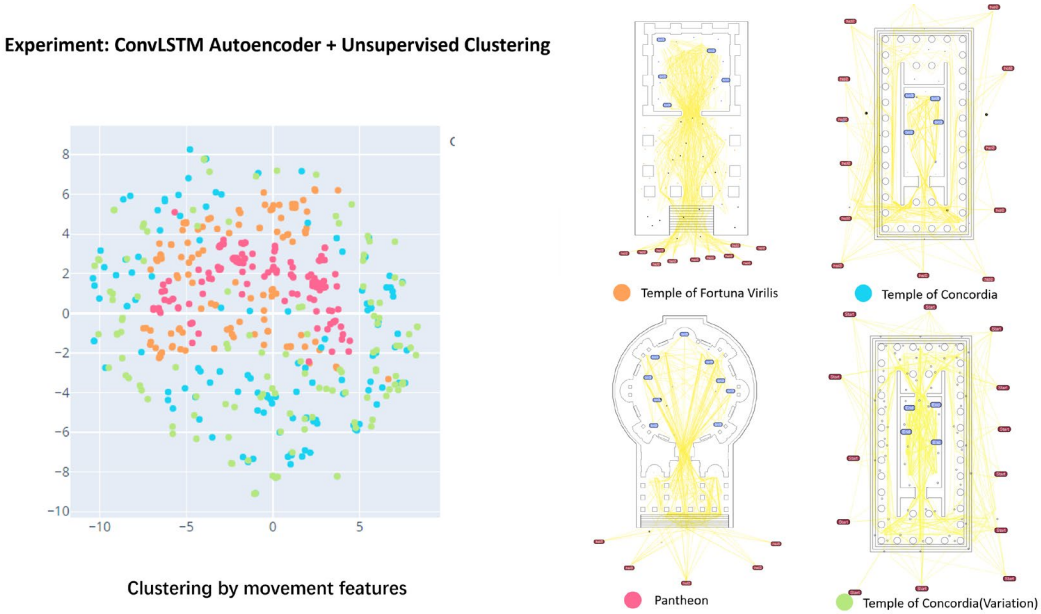

Fig8: Clustering Result of Second Experiment

Description:
By utilizing machine-generated labels instead of CNN predictions in the

first experiment, the clustering algorithm showcased the spatial sequence similarity between the Temple of Fortuna Virillis and the Pantheon, highlighting their homogeneous patterns. Moreover, it revealed the diverse spatial sequence of the Temple of Concordia.

Limits:

1. The dataset, currently comprising three temples and one variation, should be expanded to facilitate a broader comparison.
2. In the initial experiment, the predefined labels are based on modern architectural typologies. To accurately capture Greek temple features, I need to create a new set of labels tailored to this architectural genre.
3. The ConvLSTM autoencoder network architecture used in the second experiment primarily aims at high-quality reconstruction and prediction rather than efficient representation of image sequences. Further research into unsupervised representation learning networks for image sequences is required.
4. The machine-generated labels, while useful, lack legibility in architectural interpretation. They need to be further investigated and linked to the original plans or views for better understanding.
5. A valid evaluation method is necessary to assess the quality and relevance of the clustering results.

References:

Julianriise. “GitHub - Julianriise/Pedsim: Populate a Plan with People Using PedSim Plugin for Grasshopper, Rhino.” GitHub, n.d. https://github.com/julianriise/pedsim.

Lee, Jisun, and Hyunsoo Lee. 2019. “Agent-Driven Accessibility and Visibility Analysis in Nursing Units.” *Proceedings of the 24th Conference on Computer Aided Architectural Design Research in Asia (CAADRIA)*, 351–60. doi:10.52842/conf.caadria.2019.1.351.

Peng, Wenzhe. 2018. “Machines’ Perception of Space.” Massachusetts Institute of Technology.

Shi, Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. 2015. “Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting.” *ArXiv*. doi:10.48550/arxiv.1506.04214.

Shuuchen. “GitHub - Shuuchen/Video_autoencoder: Video Lstm Auto Encoder Built with Pytorch. Https://Arxiv.Org/Pdf/1502.04681.Pdf.” GitHub, n.d.https://github.com/shuuchen/video_autoencoder.

Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. 2015. “Unsupervised Learning of Video Representations Using LSTMs.” *ArXiv*. doi:10.48550/arxiv.1502.04681.