SimBuild
2024

Eleventh National Conference of IBPSA-USA
Denver, Colorado
May 21 – 23, 2024

Decomposition of Dynamic Window Views Using Semantic Segmentation

Simeon N. Ingabo[1] and Ying-Chieh Chan[1]
[1]National Taiwan University, Taipei, Taiwan

Abstract

Movement in window views impacts the indoor experience and comfort of building occupants. Some building standards therefore stipulate the presence of dynamic content as a key window view quality evaluation criterion. However, there is a scarcity of tools for evaluation of dynamic window views, owing to the complex interactions between elements in the views. Computing the compositional ratios of view elements vis-à-vis the amount of movement demands an integrated methodological framework. This paper therefore addresses the insufficiency of existing literature on dynamic window view evaluation tools and methods. A framework was developed using the *DeepLabV3* semantic segmentation architecture pre-trained on the *Cityscapes* dataset, for decomposition of dynamic content in fifty recorded urban window views. Movement within the views was calculated using functions contained in the *OpenCV* library. The pre-trained model yielded accurate predictions of twenty *Cityscapes* urban object classes, thus facilitating calculation of compositional ratios in the window views. A case study was also discussed to illustrate the practical application of the framework in determination of preferred amount of movement in office window views. This study affirms the suitability of semantic segmentation as a dynamic window view content evaluation tool.

Introduction

Window views are a key determinant of indoor comfort since they affect the well-being and productivity of building occupants (van Esch et al., 2019). Access to window views has been linked to psychological comfort (Aries et al., 2010), stress recovery (Ulrich et al., 1991), and hastened recovery from surgery (Ulrich, 1984). Movement within the window view is a desirable trait and it has a significant impact on the overall quality of the view. It has also been included as one of the window view evaluation criteria by the Leadership in Energy and Environmental Design (LEED) rating system (USGBC, 2018).

Dynamic window views are preferred by occupants over static views. Besides providing the occupants with a general awareness of their surroundings, they also relay information such as weather, time and season (Collins, 1975). However, excessive movement can have a negative impact on occupants' comfort as noted by Fisher et al. (2014). The appropriate balance of movement within window views is a subject of particular interest as stated by Ko et al. (2022).

Studies on dynamic window view content have lately gained traction. However, the deficiency of effective dynamic window view evaluation tools coupled with the lack of a unified dynamic window view evaluation framework, still pose methodological obstacles (Ko, Schiavon, et al., 2022). This has triggered interest in the use of various computer vision techniques to evaluate dynamic attributes in window views (Rodriguez et al., 2021).

Examining the complex interactions between different dynamic and static elements within the window view requires a complete decomposition of the scene. Classification of window view contents is essential for better window scene understanding. Semantic segmentation can be applied to classify and investigate the interrelationships and contributions of different constituent components of window views to the overall quality of the window views.

Aikoh et al. (2023) used semantic segmentation of street view photographs to determine the green view index, which measures the amount of greenery. Their study mainly focused on greenery at ground level and did not examine dynamic attributes in the windows. Another study by Xia et al. (2021) used a similar approach to evaluate the sky view factor using street view images. However, application of artificial intelligence in window view studies remains limited, due to the lack of frameworks and metrics specifically formulated for window view analysis. A study by Laovisutthichaia et al.

(2021) outlined a quantitative window view assessment model that applied semantic segmentation and City Information Modeling to evaluate the nature view index. These authors did not consider any dynamism in the window views.

This paper describes a methodological framework for the classification and evaluation of elements in dynamic window views using computer vision techniques, with an emphasis on semantic segmentation.

Methodology

Collection of window scenes

Fifty videos of various urban window views were recorded from different buildings, mostly using a Sony AXP55 4K Handycam. The recording procedure was such that the camera was mounted at the window in a horizontal line of sight facing the outdoor view. The windows did not to have shading devices and window frames did not to appear within the video, to ensure a clear and unobstructed view out. Each video was at least 10 minutes long. Factors considered to ensure variability of window view content included floor numbers, window orientations, time of day and sky condition.

The frame rate and total pixels per frame were inspected. Videos with a frame rate higher than 30fps were converted to 30fps. This frame rate was selected since human eyes can clearly perceive motion at 30fps, which is commonly used in most visual media such as films (Pueo, 2016). Each video was trimmed using python's *MoviePy* library to ensure that it contained exactly 18000 frames, thus each video had a duration of exactly 10 minutes.

Calculation of motion in the window scenes

Movement in the scenes was calculated using *OpenCV* by iterating through each video frame, converting each frame to grayscale, and denoising the frames by applying Gaussian blur using the *cv2.GaussianBlur* function. The *cv2.absdiff* function was used to calculate absolute difference between pixels in successive frames. Binary thresholding was performed using *cv2.threshold* to convert pixels that exhibited any change in intensity to white. Pixels that did not exhibit any change were converted to black. The change in each successive frame was calculated by counting the number of white pixels using the *cv2.countNonZero* function. Amount of movement in the scene was quantified by expressing the total changed pixels as a percentage of the total number of pixels. Average movement in each video across the 10-minute duration was expressed as the average percentage change per frame. The video processing procedure is illustrated in Fig. 1.

Semantic segmentation of window scenes

This study utilized a pretrained *DeepLabV3* semantic segmentation model provided by *TensorFlow*. The model was pretrained on the *Cityscapes* semantic segmentation dataset (Cordts et al., 2015), which contains twenty classes of urban elements initially presumed to be appropriate for this study. The classes are listed in Table 1 alongside their representative color labels in the *Cityscapes* dataset. The *Cityscapes* classes were redefined based on the prediction results from the *DeepLabV3* model to represent objects that appeared in the window views under five compositional ratios as shown in Table 1. Object classes associated with human movement (person, car, rider, truck, bus, motorcycle, bicycle and train) were categorized under the Ratio of Human-Associated Dynamic Objects (RHDO).

Table 1 Cityscapes dataset classes

Color	Cityscapes class	Class description in window view	Window view compositional ratio
	Sky	Sky	Sky ratio (SR)
	Vegetation	Vegetation	Greenery ratio (GR)
	Terrain	Grass	
	Building	Building	Building ratio (BR)
	Wall	Wall	
	Fence	Fence and balcony grills	
	Road	Road	Road ratio (RR)
	Sidewalk	Sidewalk	
	Person	Person	Ratio of Human-Associated Dynamic Objects (RHDO)
	Rider	Rider	
	Car	Car	
	Bus	Bus	
	Truck	Truck	
	Motorcycle	Motorcycle	
	Bicycle	Bicycle	
	Train	Train	
	Pole	Pole	
	Traffic light	Traffic light	
	Traffic sign	Traffic or advertisement sign	
	Void	-	-

Another aim of the study was to evaluate the adequacy of this dataset for window view studies and determine whether it may be necessary to develop a new task-specific dataset.

Semantic segmentation was thereafter performed on the extracted frames using the pretrained model to yield prediction maps. The ground truths and predictions were compared to identify visible differences. A similar approach was used to segment every frame of each video and recombine the prediction masks of the frames back into videos. The segmentation output videos were then used to calculate class areas and compositional ratios in each scene.

Calculation of class areas and compositional ratios

The videos were recorded using different cameras, hence they contained varying pixels counts per frame. Total number of pixels in each video was calculated by counting the number of pixels in one frame and multiplying it by 18000 frames. The area of each class was determined by counting the total number of pixels of each class color code across all frames in each video. The five compositional ratios were obtained by finding the combined percentage of pixels in the scene representing its constituent classes listed in Table 1.

Case study on application of the framework

In order to illustrate a practical application of the developed framework, a survey was designed to investigate the amount of movement in window views that is preferred for office tasks. The window scene videos were divided into ten-second segments. The segments containing the frames with highest pixel change in each scene were selected, with reference to the movement profiles. This ensured that selected clips had the highest movement observed within the scene. Thirty occupants of an office building were recruited to participate in the survey. The 50 ten-second video segments of the scenes were grouped into two to reduce survey time. Each subject was simply asked to observe 25 scenes and rate the amount of movement suitable for office tasks under three levels; insufficient (too little), satisfactory (enough), and excessive (too much). The subjects were provided with a link to the online questionnaire and asked to watch the 10-second videos of the scenes one by one, while rating the amount of movement in the questionnaire. Study subjects alternated between the two groups based on their order of participation. The first subject viewed the first group of 25 videos while the second subject viewed the second group. This alternating format was preserved and each group of videos was viewed by 15 subjects. The survey lasted 5 to 10 minutes and no personal information was collected from the subjects.

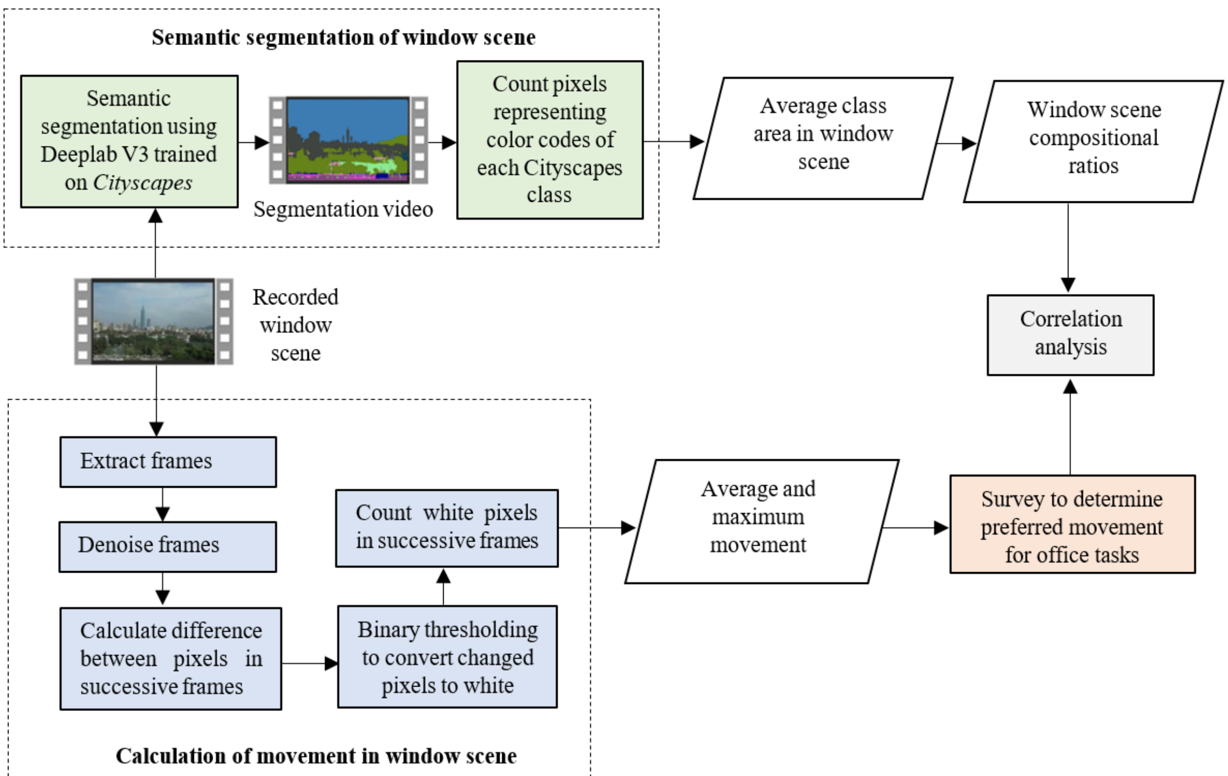


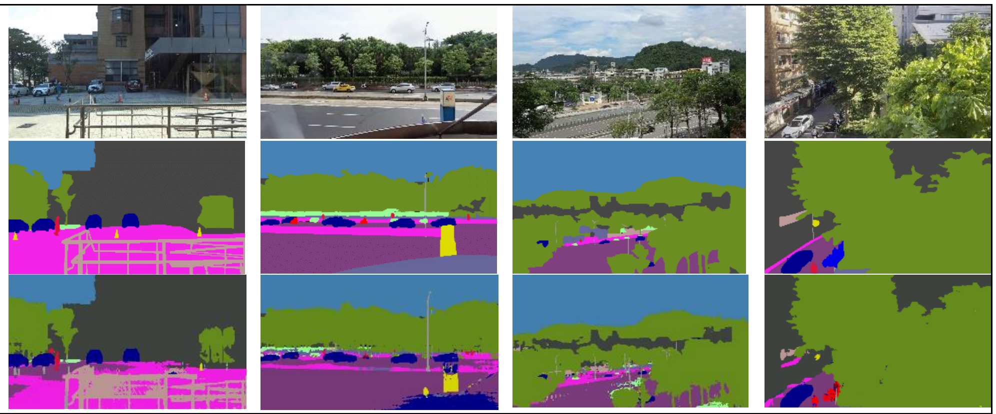
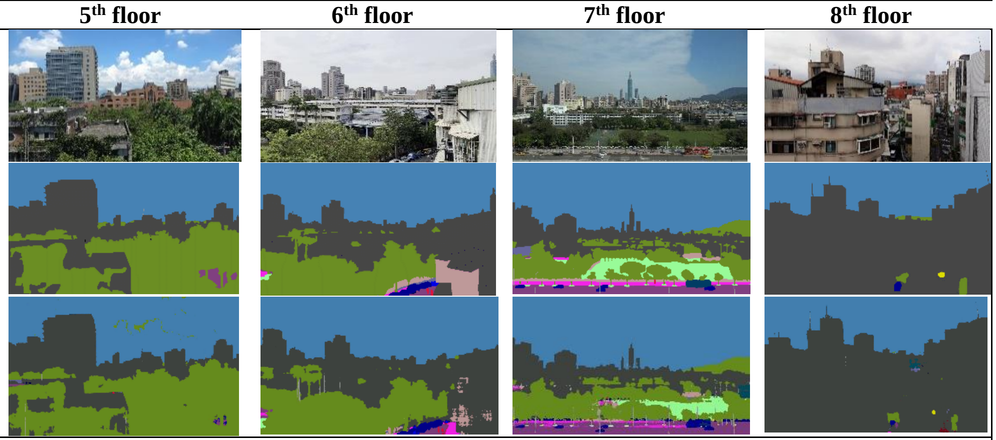
Figure 1 Flow chart of research framework

Results and discussion

Performance of the semantic segmentation model

One random frame was extracted from selected videos recorded on different floors and labeled to create a ground truth based on the *Cityscapes* classes and their color labels. Each video contained 18000 frames and labeling them in their entirety would be extremely tedious. The Intersection Over Union (IoU) metric was used to evaluate class accuracy of the *DeepLabV3* model. IoU was calculated using the ground truth image against the prediction mask of the same image from the model. However, this is not representative of the performance of the model across the entire video. It serves to provide a general overview of the model's accuracy. Table 2 shows the performance of the model on 6 prominent classes constituting the 5 compositional ratios. It was observed that class IoU decreases with increase in distance from the viewing position. However, this does not imply that the model does not detect human-related dynamic content, rather the intersection between predicted and labelled masks is less accurate with increasing distance from the window as the objects progressively decrease in size.

Table 2 DeepLabV3 class accuracy on selected window scenes recorded on different building floors

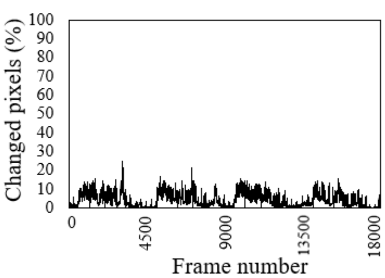
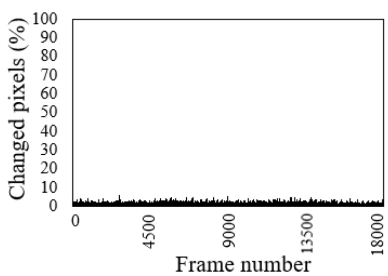
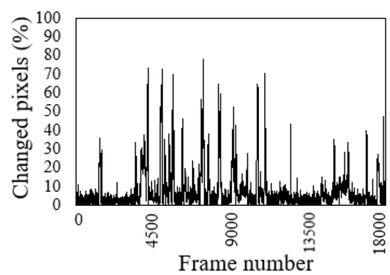
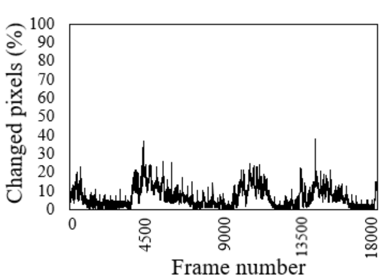
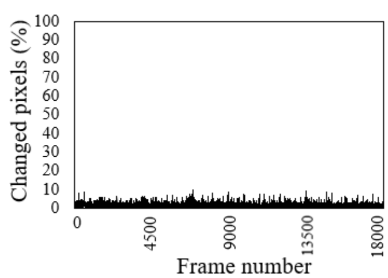
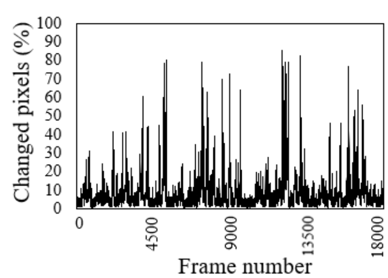
		1st floor	**2nd floor**	**3rd floor**	**4th floor**
Window views	View Ground truth Prediction				
Class IoU (%)	Sky	86.5	95.2	98.3	-
	Vegetation	77.5	87.7	88.2	94.4
	Building	94.2	6.1	64.5	79.1
	Road	-	79.6	70.9	73.3
	Car	86.1	30.1	33.8	89.8
	Person	53.3	18	25.9	4.3
		5th floor	**6th floor**	**7th floor**	**8th floor**
Window views	View Ground truth Prediction				
Class IoU (%)	Sky	95.8	98.6	99.0	95.2
	Vegetation	86.3	90.0	77.7	36.2
	Building	80.0	82.9	81.7	97.0
	Road	4.0	60.1	74.4	-
	Car	-	80.2	21.3	24.4
	Person	-	28.6	4.4	-

The six window views, their ground truths and predictions are shown in Table 2. The model was observed to provide accurate predictions of the sky, building and vegetation classes, which constitute a significant area of each window view. However, the accuracy of human-associated dynamic elements was notably observed to decrease significantly with increasing distance from the window. Objects that partially appear in the views such as parked cars and motorcycles in the 4th floor scene and building section in the 2nd floor scene are sometimes wrongly classified. This was principally observed to occur between human-associated dynamic objects. However, these objects had already been categorized under RHDO, thus reducing the effects of inter-object misclassification on the calculated compositional ratios.

Movement profiles

Table 3 shows the movement profiles of selected views. The graphs indicate the percentage of pixels that change between successive frames in the videos. It was observed that scenes with similar or comparable compositional ratios tend to have a similar movement profile. Scenes in column A contained controlled traffic flow, thus their movement profiles displayed a pattern. Scenes in column B had hardly any movement, but the marginal change observed in the movement profiles can be attributed to daylight changes in the scene, which may not be immediately noticeable. The window scenes in column C were captured in very close proximity to roads, hence they had high random movement.

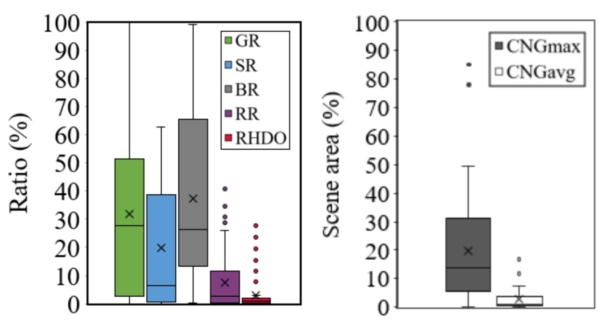
Table 3 Movement profiles of selected window scenes

A	B	C
		
		
		
		

Compositional ratios from semantic segmentation of recorded window scenes

Figure 2(a) shows the distribution of compositional ratios in the fifty recorded window scene videos. The building ratio was the most prominent due to the built up nature of urban environments, with buildings appearing in most window views. The mean BR, GR, SR, RR and RHDO were 38%, 31%, 18%, 8% and 4% respectively. Some window scenes were almost entirely composed of greenery and buildings. Maximum sky ratio was 63%. The RHDO was below 10% since most human-associated dynamic objects enter and leave the scene, thus they do not appear in all frames. Some outlier RHDO points are observed in Fig. 2(a), resulting from scenes containing parking lots and cars parked by the roadside.

Figure 2(b) shows the distribution of average percentage change between successive frames (CNGavg) in the 50 window scenes, calculated by averaging the change between successive frames in the entire duration of each video. The figure also shows the maximum percentage change between succesive frames (CNGmax), selected by establishing the highest change observed between any two successive frames in each video. The mean maximum change was 20% with most scenes exhibiting less than 50% change between successive frames. Two outlier scenes experienced a maximum change of 77% and 86%. The change between successive frames was not deployed as a dynamic window view evaluation metric. It was specifically used to provide a quantitative measure of movement in the recorded videos.



a) Compositional ratios *b) Movement*

Figure 2 Distribution of compositional ratios and amount of movement in recorded videos

Case study

This case study illustrates the application of compositional ratios computed using semantic segmentation to estimate the likelihood of experiencing satisfactory, excessive or insufficient movement in an office window view. Compositional ratios in each scene and corresponding responses of the subjects from their subjective evaluation of preferable movement in the scenes are shown in Fig. 3. and Fig. 4 They are arranged in decreasing order from the scene that was perceived to contain satisfactory movement by the highest number of subjects to the lowest.

Table 4 shows the Spearman correlation coefficients between the compositional ratios in the 10-second survey videos and movement preference. The p-values of the correlations are also listed. Greenery ratio was observed to have a significant positive correlation with satisfactory movement and a significant negative correlation with insufficient movement. This implies that window scenes with high amounts of greenery are less likely to contain insufficient movement. Sky ratio had no significant correlation with all the three levels of movement. Building ratio had a significant negative correlation with satisfactory movement and a significant positive correlation with insufficient movement, indicating that window views with high building ratios are less likely to experience satisfactory movement and more likely to experience insufficient movement.

The road ratio had a significant positive correlation with excessive movement and a significant negative correlation with insufficient movement. This indicates that scenes with high road ratios are likely to experience excessive movement and less likely to experience insufficient movement. On the other hand, the ratio of human-associated dynamic objects was observed to have a significant positive correlation with excessive movement and a significant negative correlation with insufficient movement. Both road ratio and ratio of human-associated dynamic objects had no significant correlation with satisfactory movement. Based on the sample utilized in this study, greenery ratio appears to be the best indicator of satisfactory movement in window views.

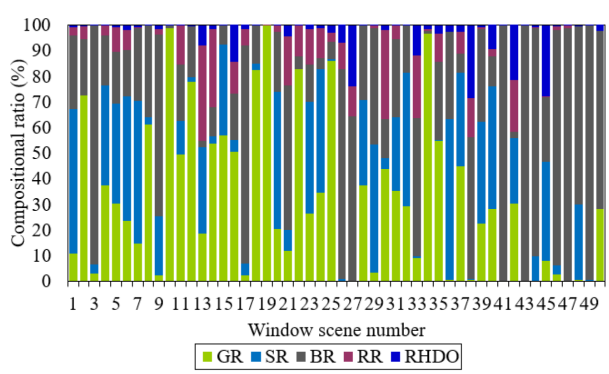


Figure 3 Compositional ratios in each scene

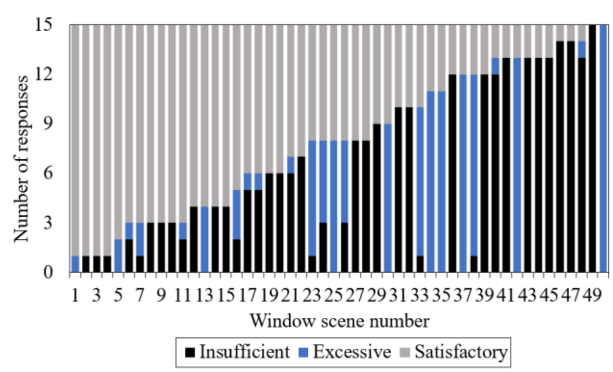


Figure 4 Survey responses

Table 4 Spearman correlation coefficients in ten-second scenes used in the survey

	GR	SR	BR	RR	RHDO	CNG max	CNG avg
Satisfactory	0.431	0.180	-0.334	0.128	-0.205	0.062	0.069
Excessive	0.135	0.007	-0.277	0.498	0.618	0.397	0.514
Insufficient	-0.450	-0.012	0.457	-0.431	-0.373	-0.459	-0.567
p-values of correlations (**p <= 0.05*)							
Satisfactory	0.002*	0.211	0.018*	0.375	0.152	0.667	0.634
Excessive	0.349	0.959	0.051	0.000*	0.000*	0.004*	0.000*
Insufficient	0.001*	0.936	0.001*	0.002*	0.008	0.001*	0.000*

Potential applications in window view studies

Semantic segmentation of dynamic window views enables several potential applications. It can be used to evaluate the impact of movement from different object classes in the scene on occupants. It can also be used as a foundation for methodological frameworks formulated to evaluate window view content using other metrics, owing to the additional visual information additional visual information provided by the *Cityscapes* classes.

Besides compositional ratios, semantic segmentation can also facilitate automated application of existing view evaluation methods such as evaluation of view layers present, rule of thirds and presence of far elements. It can also be used to estimate how much movement is in a scene from static images of the scene, thus enabling the use of static images for evaluation of dynamic views.

In the context of building simulation, the automated framework built around semantic segmentation can be integrated into window view evaluation plugins in 3D modelling software.

Limitations of the framework

The semantic segmentation framework outlined in this study was effective for dynamic window view evaluation. However, notable limitations were identified. *Cityscapes* is a robust dataset based on ground views of urban scenes, thus the pretrained model was observed to be more accurate in semantic segmentation of first floor window views. The dataset however lacks annotations of rooftops. They are classified as road or sidewalk in scenes containing large sections of rooftops, as observed in Fig. 5. This decreases the accuracy of the model with increasing floor height. Distant objects that are not clearly identifiable were also erroneously classified, but they generally cover small portions of the scenes.

The framework used is also unsuitable for window scenes that contain special architectural and landscape features such as mountains, water bodies and landmark buildings, which significantly influence the occupants' subjective preference of the views. These features are classified under normal classes available in the *Cityscapes* dataset, thus neglecting their compositional contribution to the views. *Cityscapes* contains one general class for all sky conditions, thus it cannot be used to evaluate cloud movement and daylight changes. These are integral elements in dynamic window views. Reflections on glazed building facades and water bodies also pose a challenge. Water bodies appearing within the window view are classified based on the class that is reflected by the water surface.

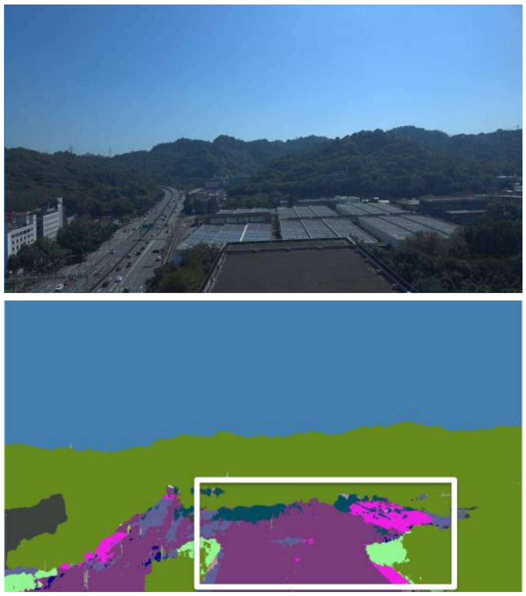


Figure 5 Misclassification of rooftops

This study only focused on short-term movement within window views. The framework may not be effective for evaluation of long-term movement such as diurnal daylight and weather changes, as well as seasonal changes in window view content.

Conclusion

Quantitative understanding of window views is essential in conducting window view studies, particularly the development of window view evaluation metrics and indices. These studies are largely dependent on an aggregated assessment of all elements within the view. However, dynamic window views present a challenge since the elements also exhibit movement. The novel framework presented in this study addressed both composition and movement.

A *DeeplabV3* semantic segmentation model pretrained on the robust *Cityscapes* dataset was used to classify the content in 50 videos of urban dynamic window scenes. This enabled the computation of compositional ratios. Amount of movement in the scenes and movement profiles were determined by counting the number of changed pixels in successive frames using *OpenCV* functions. Despite few limitations resulting from the dataset used, the model made accurate class predictions. The framework can be applied in examination of the association between movement and composition of window views, as evidenced by its application in a practical case study. It is therefore suitable for decomposition and evaluation of dynamic window views.

Future work

We are developing a window views dataset that addresses the limitations identified in the *Cityscapes* dataset. A new semantic segmentation model will be trained using the dataset developed to enhance the accuracy of predictions.

Acknowledgment

This research was supported by the National Science and Technology Council, Taiwan, under Grant no. NSTC 112-2221-E-002 -123 -. The authors declare that they have no known competing interests.

References

Aikoh, T., Homma, R., & Abe, Y. 2023. Comparing conventional manual measurement of the green view index with modern automatic methods using google street view and semantic segmentation. *Urban Forestry & Urban Greening*, 80, 127845.

Aries, M. B., Veitch, J. A., & Newsham, G. R. 2010. Windows, view, and office characteristics predict physical and psychological discomfort. *Journal of environmental psychology*, 30(4), 533-541.

Collins, B.L., 1975. Windows and people: A literature survey. Psychological reaction to environments with and without windows. (No. NBS-BSS-70). National Bureau of Standards, Washington, DC (USA); US Dept. of Commerce, National Bureau of Standards. *Institute for Applied Technology, Washington*.

Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. 2015. The cityscapes dataset. In *CVPR Workshop on the Future of Datasets in Vision* (Vol. 2). sn.

Fisher, A. V., Godwin, K. E., & Seltman, H. 2014. Visual environment, attention allocation, and learning in young children: When too much of a good thing may be bad. *Psychological science*, 25(7), 1362-1370.

Ko, W. H., Kent, M. G., Schiavon, S., Levitt, B., & Betti, G. 2022. A window view quality assessment framework. *Leukos*, 18(3), 268-293.

Laovisutthichai, V., Li, M., Xue, F., Lu, W., Tam, K. and Yeh, A.G., 2021, November. CIM-enabled quantitative view assessment in architectural design and space planning. In *Proceedings of 38th International Symposium on Automation and Robotics in Construction* (ISARC 2021) (pp. 65-72).

Pueo, B. 2016. High speed cameras for motion analysis in sports science. *Journal of Human Sport and Exercise*, 11(1), 53-73.

Rodriguez, F., Garcia-Hansen, V., Allan, A., & Isoardi, G. 2021. Testing the adequacy of luminous change descriptors to represent dynamic attributes in outdoor views. *Building and Environment*, 191, 107591.

Ulrich, R. S. 1984. View through a window may influence recovery from surgery. *Science*, 224(4647), 420-421.

Ulrich, R. S., Simons, R. F., Losito, B. D., Fiorito, E., Miles, M. A., & Zelson, M. 1991. Stress recovery during exposure to natural and urban environments. *Journal of environmental psychology*, 11(3), 201-230.

USGBC. 2018. LEED v4.1 Interior Design and Construction. EQ Credit: Quality Views. *U.S Green Building Council*: 145.

van Esch, E., Minjock, R., Colarelli, S. M., & Hirsch, S. 2019. Office window views: View features trump nature in predicting employee well-being. *Journal of environmental psychology*, 64, 56-64.

Xia, Y., Yabuki, N., & Fukuda, T. 2021. Sky view factor estimation from street view images based on semantic segmentation. *Urban Climate*, 40, 100999.