An application of convolutional neural network in street image classification

The case study of London

Stephen Law
Alan Turing Institute
London, UK
slaw@turing.ac.uk

Yao Shen
University College London, CASA
London, UK
y.shen.12@ucl.ac.uk

Chanuki Seresinhe
Alan Turing Institute
London, UK
cseresinhe@turing.ac.uk

ABSTRACT

Street frontage quality is an important element in urban design as it contributes to the interest, social life and success of public spaces. To collect the data needed to evaluate street frontage quality at the city or regional level using traditional survey method is both costly and time consuming. As a result, this research proposes a pipeline that uses convolutional neural network to classify the frontage of a street image through the case study of Greater London. A novelty of the research is it uses both Google streetview images and 3D-model generated streetview images for the classification. The benefit of this approach is that it can provide a framework to test different urban parameters to help evaluate future urban design projects. The research finds encouraging results in classifying urban frontage quality using deep learning models. This research also finds that augmenting the baseline model with images produced from a 3D-model can improve slightly the accuracy of the results. However these results should be taken as preliminary, where we acknowledge several limitations such as the lack of adversarial analysis, labeled data, or parameter tuning. Despite these limitations, the results of the proof-of-concept study is positive and carries great potential in the application of urban data analytics.

CCS CONCEPTS

• **Computing methodologies** → **Scene understanding**; *Supervised learning by classification*; • **Applied computing** → **Computer-aided design**; • **General and reference** → *General conference proceedings*;

KEYWORDS

urban design, deep learning, convolutional neural network, machine vision, London

1 INTRODUCTION

The field of urban design primarily concerns the space between buildings and how this can influence the way that pedestrians move, navigate, interact, live and work in cities. In this study, we focus on the design of streets, and in particular, active urban street frontage [18]. In the urban design literature, active frontage is defined as ground floor building frontage having windows and doors as opposed to blank walls, fences and garages [15]. The quality of street frontages is an important element in urban design, as it contributes to the wider interest, social life and success of public space [4][5]. There are multiple benefits in creating successful public spaces through active street frontage. These include social benefits such as safety factors, economic benefits such as the increase in property value, and health benefits such as improved pedestrian access [8]. A particularly important notion in the quality of street frontage is that it can provide natural surveillance at street-level. To put it more simply, the greater the number of doors, staircases and windows fronting a street, the safer and the more inclusive does the street seem. As Jacobs [9] famously said, there are potentially more "eyes-on-the-streets", which brings greater sense of security at the street level. Quantitatively, the concept of active frontage has been expressed through indicators such as the facade evaluation scale, in which a higher grade (A) has a greater frequency of fenestrations and doors than a lower grade (E), which has lower frequency [8][18]. To measure these metric requires many structured interviews with professionals. Therefore, to collect the data required to evaluate street frontage quality at the city scale is both costly and time-consuming. One approach is to cast this problem as a classification problem in machine vision. This research applies deep learning methods to the classification of urban street images. These machine learning techniques have made great advances in image classification [11], object detection [6], image segmentation [1] and edge detection [12]. This research proposes a pipeline and a proof-of-concept that uses a convolutional neural network in classifying the street frontages of a front-facing street image. This research differs from previous research in that it focuses on first of all classifying a street image into four classes of ground-floor street frontages, and secondly in augmenting the training dataset with 3D-modelled street image data. The benefit OF using a mixed-reality approach is that it could provide a framework that can be used to test different urban design parameters.

2 PREVIOUS WORKS

Despite the many benefits of active street frontages [8], there has been limited computational research in the classification of street frontages using street image data [13]. Four studies are here referred by way of illustrating the current status of computation techniques in urban street image analysis.

The first is from Doersch et al. [2], who uses object detection to identify distinctive local architectural features in a case study of Paris. In this study, architectural elements such as cast iron railings, fenestrations and doors have been identified as distinctive features in the Parisian streetscape. The research uses traditional machine vision features such as histograms of oriented gradients in architectural object detection. This research was successful in identifying distinctive architectural features and was one of the earlier studies to use classical machine vision techniques in architecture.

The second is the study by Naik et al. [14], who use streetview images to estimate the perceived quality of streets. A large-scale crowdsource game known as 'Place Pulse' was developed in order to assess the perception quality through a pairwise image comparison. The author called this perception indicator Streetscore[17] and it has been used to identify historic districts in cities, to quantify urban changes and to determine urban perception on a global scale.

The third study is by Seresinhe et al [16], who used an image-database from an online game called 'Scenic-or-not' in which visitors would rate a random outdoor image in the UK on a scale from 1-10. The novelty of the study is that it uses deep learning techniques to estimate the perceived scenicness of outdoor images in the UK. The study also found that places with flat topography such as large areas of flat grass are associated with lower scenicness while places with varying topography such as valleys are associated with greater scenicness.

The final study is by Liu et al [13]. This study uses streetview images to estimate the visual quality of a street facade. It compares ratings collected from the survey to train an image classifier in predicting a facade quality evaluation scale. The results show that the ratings predicted by the machine learning algorithm is comparable to those defined by domain-experts.

The current study extends from these previous works in examining the quality of street facades, and it also differs from the earlier studies in two ways. First of all, this study collects front-viewing street image data from the median of a street rather than image data perpendicular to the street. This allows the streetview image to be classified into four groups; 1. active on both sides of a street, 2. blank frontage on one side of a street, 3. blank frontage on both sides of the street, and 4. non-urban frontage on both side of a street. Secondly, it also compares the baseline model that uses only street images collected from the Google Streetview API [7] to a model that augments that baseline model with street images produced from a 3D model of the city. finally, the model is used to predict both ground truth labels on images from Google streetview as well as the 3D-model streetview separately.

3 METHOD AND MATERIALS

This study proposes the StreetFrontageNet (SFN) model, which classifies the ground floor of a front-facing streetview image into four categories. In order to train the classifier, this study uses Greater London, fig1, as the case study, and collects two ground truth streetview image datasets. The ground truth images are subsequently trained and tested using a CNN image classifier. Figure 2 shows the proposed pipeline, which consists of data collection, ground truth labelling and CNN image classification.

3.1 Data collection

We first describe the data collection of the two image sets. The first set is comprised of street images from Google Streetview API [7](©2017 Google Inc. Google and the Google logo are registered trademarks of Google Inc.). Using the API, one front-facing image has been collected for each streets in the Greater London Area. To collect the dataset, we first constructed a graph from the street network of London (OS Meridian line2 dataset), in which every node is a junction and every edge is a street. We then took the geographic

Figure 1: Greater London study area. Contains Ordnance Survey data ©Crown copyright and database right ©2017.

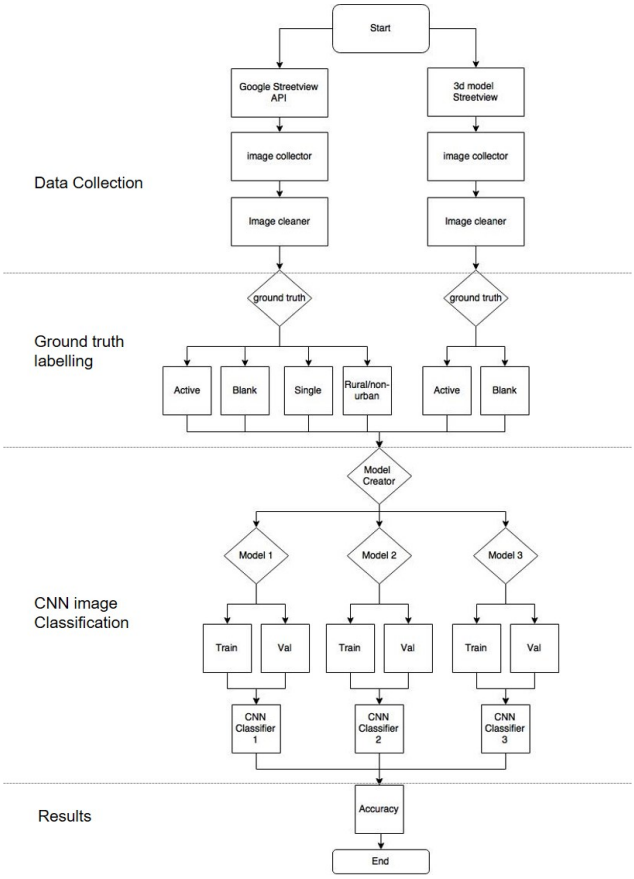

Figure 2: StreetFrontageNet modelling pipeline

median and the azimuth of the street edge between two junctions to give both the location and the bearing of each streetview image. This is to ensure the streetview images are constantly front-facing

and are taken from the centre of the road rather than from near the junction. This reduces the problem of images being too close to the junction. The field of view has been set to 120 degree in order to ensure that both sides of the facades are captured. 110,000 images have been collected from this process.

The second set comprises street images generated from a 3D model of an abstract city produced in ESRI City Engine [3](©1995-2017 ESRI. All rights reserved.). To re-create the effect of a blank surface at ground level, a four metre high blank wall was modeled at the edge of the pavement, representing a blank surface. Two types of surfaces were chosen for the blank wall, namely a brown coloured one representing wood and another the colour of concrete. A simulation of an agent walking along a random path in this abstract city was then recorded at 10 frames per second. The objective was to replicate a similar sequence of images to that of the Google streetviews. A total of 4800 images were collected by this process.

3.2 Ground truth labelling

The research then progressed and the two image datasets mentioned in the previous section were labelled. First, we processed and labelled the Google streetview images, which involved removing invalid images such as the interior of buildings and those in which, facades had been obscured by large vehicles such as buses. We also removed images that were too dark or those not available on Google Streetview. A series of automatic functions and manual processes were also used to identify and remove invalid images. Figure 3 shows examples of the invalid images.

Figure 3: Invalid images. From left to right, not available image, dark image, interior image, interior image. ©2017 Google Inc. Google and the Google logo are registered trademarks of Google Inc.

Following the cleaning process, 10,000 images were randomly selected. This dataset was then resized into a set dimension (256 pixels x 256 pixels) and the ground truth labelling was then manually performed by the author. This study defines active frontages [15] as those in which the ground floor building frontage has windows and doors as opposed to blank walls, fences or garages. For the Google streetview images, four street frontage classes were adopted namely; 0 - active frontage on both sides of the street; 1 - active frontages on one side of the street; 2 - blank frontage on both sides of the street; 3 - rural/non-urban/unclassified images. The four classes of urban street frontages can be seen in figure 4.

Second, we then processed and labelled the 3D-model streetview images. This process included removing invalid images close to intersection, images taken at the end of a road, those that are not facing the street and duplicate images. There were many duplicate images due to the high number of frames per second. Following

Figure 4: Google Streetview urban frontage images. From left to right, active frontage, single-sided active frontage, blank frontage, non-urban/unclassified frontage. ©2017 Google Inc. Google and the Google logo are registered trademarks of Google Inc.

the cleaning process, 1029 images were randomly selected. This dataset was then similarly resized into a set dimension (256 pixels x 256 pixels). The ground truth labeling was performed automatically as we produced two sets of images from the 3D model; one with blank walls representing blank frontages and one without blank walls representing active frontages. For the 3D-model of streetview images, two classes of frontages were adopted, namely; 0 - active frontages on both sides of the street and; 2 - blank frontages on both sides of the street. The two classes of street frontage facade can be seen in figure 5. One limitation is that this study did not produce any 3D-model single-sided active frontage images or any 3D-model non-urban images. Future research will consider augmenting the SFN model through simulated street frontage images of both classes.

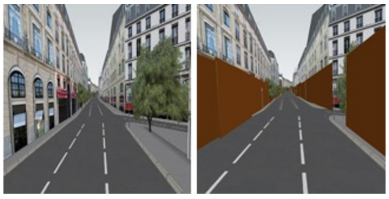

Figure 5: 3D model urban frontage images. Left: Active frontages, Right: blank frontages. ©1995-2017 ESRI. All rights reserved.

3.3 CNN image classifier

These two datasets were then fed into three separate Convolutional Neural Network (CNN) models. The first CNN model used the Google streetview image data, the second model used both the Google streetview image data and the blank frontage images from the 3D streetview model. The third CNN model used the Google streetview image data as well as both the active and blank frontage images produced from the 3D streetview model. This study selected the widely used Alexnet CNN architecture [11] due to its wide use, efficiency and performance. For robustness, future research should consider more advanced image classification architectures such as VGG and Googlenet.

This study uses the Alexnet architecture [11]. The model has five convolution layers and three fully connected layers that detect basic edges in the earlier layers up to more complex shapes in the latter layers. Similar to [16], we also used transfer learning from

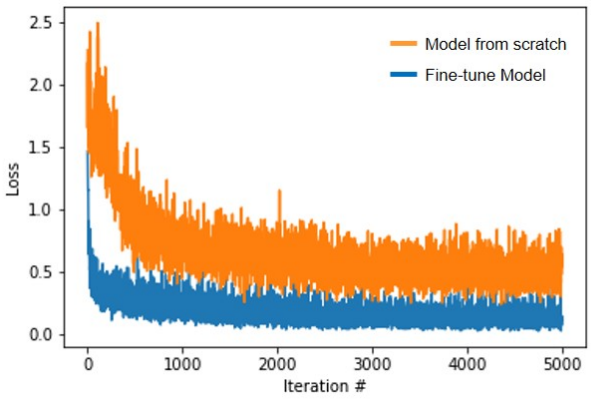

Figure 6: Loss graph comparing a fine-tune model vs a model from scratch after 5000 iterations

the Places Database [19] to leverage knowledge from the weights of the pre-trained model. Figure 6 compares the results of a fine-tuned model (transfer learning model) in blue with a model built from scratch in orange. The result shows that the fine-tuned model produces lower losses and converges more quickly than the same CNN model from scratch. This conforms to the current practice for image classification. To train, we modified the last fully connected layer (fc8 in Caffe) of the eight-layer Alexnet to classify four output units. Softmax function 1 is then used to estimate the probability distribution of an image class.

$$\hat{p}_{nk} = \frac{exp(x_{nk})}{\sum_{k=1}^{K} \exp(x_{nk})} \quad (1)$$

We have split the dataset where 80% is used for training and 20% is used for validation. We train the CNN using stochastic gradient descent (SGD) in the Caffe Library [10]. The learning rate starts at 10^{-4} and it drops to 10^{-5} after the 2500 iterations. The weight decay parameter is set to 5^{-4} with a SGD momentum of 0.9. These parameters achieved a high accuracy for the study. Fine-tuning of the model parameters will be tested in future studies.

4 RESULTS

To test the accuracy of the three models, we first compared the Google streetviews ground truth label with the most likely frontage class predicted in model 1 that uses only Google streetview images, model 2 where the baseline model is augmented with blank-frontage images generated from the 3D streetview model, and also with model 3, in which the baseline model is augmented with both active as well as blank frontage images generated from the 3D streetview model. The result in table 1 shows, first, that all three models achieve a high accuracy when predicting the ground-truth label. Model 1 achieves 75% accuracy, while model 2 achieves 79% accuracy and model 3 achieves 77% accuracy. The results show that the model with the 3D augmented data achieves similar or slightly better results than the model without the 3D augmented data 1.

We then compared the 3D model streetview ground truth label with the most likely frontage class predicted from the three models. The results in table 2 show that model 1 achieves a 43% accuracy while model 2 and model 3 achieves over 95% accuracy. This is to be expected as the training set for both models 2 and 3 contains images from the 3D-model, while model 1 does not. This suggests that real streetview images cannot be used in this case to predict the 3D-model ground truth data. This is somewhat contradictory to the previous results. Further research is needed to validate this outcome.

Table 1: Google Streetviews prediction accuracy

Model	description	accuracy
model 1	base model	75.27
model 2	base + 3D(blank)	79.06
model 3	base + 3D(blank + active)	77.79

Table 2: 3D-model Streetviews prediction accuracy

Model	description	accuracy
model 1	base model	42.66
model 2	base + 3D(blank)	94.46
model 3	base + 3D(blank + active)	98.05

Finally, we used the best performing model to predict the probability of an active frontage on every single street segment in London. The results can be seen in figure 7, in which red represents a greater probability of active frontages and blue represents a lower probability of active frontages. The results show that central London has, as expected, a higher probability of active frontages. This is probable as Central London has a higher urban density. The results also show that areas such as the Isle of Dogs have a higher probability of having blank frontages. This is consistent with the consensus, in which the urban nature of newer areas are generally less active and less pedestrian friendly.

5 DISCUSSION

To end, this study finds encouraging results in classifying urban street frontage quality using deep learning CNN models [16] [13]. This research also finds that augmenting the baseline model with images produced from the 3D-model can improve the overall accuracy of the model. However these results should be taken as preliminary, where we acknowledge several limitations such as the use of a simple CNN architecture, the lack of adversarial analysis, labeled data, or parameter tuning. Secondly, the focus on classifying a simple urban design parameter for a single case study reduces the extent the research results can be generalised. The study, for example, does not differentiate between a wall with a small window or a large one. Thirdly, the images from Google streetviews are not entirely reliable. Concerns can range from visual obstruction to poor lighting condition. A semi-automatic process of image removal was implemented. However, reliability checks and further improvements in the pipeline are necessary to improve the image processing efficiency. Future research is needed to improve the overall accuracy of the results, to better understand what the computer

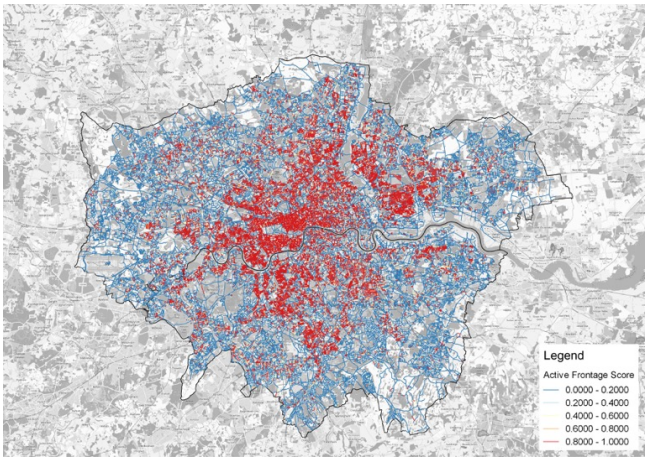

Figure 7: Prediction outcome for the SFN classifier for every single street in London. Contains Ordnance Survey data ©Crown copyright and database right ©2017.

actually sees within the CNN model and the extent augmented 3D-images can be used to replace real ones and vice versa. Despite these limitations, the results are encouraging as a proof-of-concept study. Knowing the geographical distribution of active and blank frontages is an important topic for urban planning. The implication is that these models can be used to reduce the time needed for data collection. These research can help better understand the extent an urban design quality can influence human behaviour, social and economic outcome at the city-wide scale using computational techniques.

ACKNOWLEDGMENTS

This work was supported by The Alan Turing Institute under the UK Engineering and Physical Sciences Research Council (EPSRC) grant no. EP/N510129/1. The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions.

REFERENCES

[1] L Chen, G Papandreou, I Kokkinos, K Murphy, and A.L. Yuille. 2014. Semantic image segmentatiom with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062* (2014).
[2] C Doersch, S Singh, C Wu, and W Hui. 2012. ACM Transactions on Graphics. *What makes Paris look like Paris* (2012).
[3] ESRI. 2013. http://www.esri.com/software/cityengine/. (2013).
[4] J Gehl. 1971. *Life between buildings: Using public space.* The Danish Architectural Press.
[5] J. Gehl. 2010. *Cities for People.* Island Press.
[6] R Girshick. 2015. Fast R-CNN. *IEEE International Conference on Computer Vision* (2015).
[7] Google. 2017. https://www.maps.google.com/. (2017). Accessed: 2017.
[8] E Heffernan, T Heffernan, and W Pan. 2014. The relationship between the quality of active frontages and public perceptions of public spaces. *Urban Design international* (2014).
[9] J. Jacobs. 1961. *The Death and Life of Great American Cities.* Random House Inc.
[10] Y Jia. 2012. Caffe: An open source convolution architecture for fast feature embedding. http://caffe.berkeleyvision.org. (2012). Accessed: 2016.
[11] A Krizhevsky, I Sutskever, and G.E Hinton. 2012. Imagenet classifi- cation with deep convolutional neural networks. *Advances in neural information processing* (2012).
[12] Y Li, M Paluri, J Rehg, and P Dollar. 2016. Unsupervised learning of edges. *CVPR* (2016).
[13] L Liu, E Silva, C Wu, and W Hui. 2017. A machine learning-based method for the large-scale evaluation of the urban environment. *Computers, Environment and Urban Systems* (2017).
[14] N. Naik, J. Philipoom, R. Raskar, and C.A. Hidalgo. 2014. StreetScore - Predicting the Perceived Safety of One Million Streetscapes. In *CVPR Workshop on Web-scale Vision and Social Media.*
[15] Office of the Deputy Prime Minister. 2005. *Safer Places: The Planning System and Crime Prevention.* Home Office.
[16] C Seresinhe, T Preis, and S Moat. 2017. Using deep learning to quantify the beauty of oudoor plaes. *Royal Society Open Science* (2017).
[17] Streetscore. 2014. http://streetscore.media.mit.edu. (2014). Accessed: 2016-04-29.
[18] Llewelyn Davies Yeang and Homes Communities Agency. 2013. *Urban Design Compendium.* English Partnerships and The Housing Corporation.
[19] B Zhou, J Lapedriza, A Xiao, A Torralba, and A Oliva. 2014. Learning deep features for scene recognition using places database. *Advances in neural information processing* (2014).