BETWEEN IMAGES AND BUILT FORM: AUTOMATING THE RECOGNITION OF STANDARDISED BUILDING COMPONENTS USING DEEP LEARNING

C. Pezzica [1, a*], J. Schroeter [2, a], O. E. Prizeman [1], C. B. Jones [2], P. L. Rosin [2]
[1] Welsh School of Architecture, Cardiff University, King Edward VII Avenue, Cardiff, UK - (pezzicac, prizemano)@cardiff.ac.uk
[2] School of Computer Science & Informatics, Cardiff University, The Parade, Cardiff, UK – (schroeterj1, jonescb2, rosinpl)@cardiff.ac.uk

Commission II, WG II/8

KEY WORDS: Object Recognition, HBIM, Image Classification, Deep Learning, CNN, Modern Heritage Conservation, Carnegie Libraries, Specifications

ABSTRACT:

Building on the richness of recent contributions in the field, this paper presents a state-of-the-art CNN analysis method for automating the recognition of standardised building components in modern heritage buildings. At the turn of the twentieth century manufactured building components became widely advertised for specification by architects. Consequently, a form of standardisation across various typologies began to take place. During this era of rapid economic and industrialised growth, many forms of public building were erected. This paper seeks to demonstrate a method for informing the recognition of such elements using deep learning to recognise 'families' of elements across a range of buildings in order to retrieve and recognise their technical specifications from the contemporary trade literature. The method is illustrated through the case of Carnegie Public Libraries in the UK, which provides a unique but ubiquitous platform from which to explore the potential for the automated recognition of manufactured standard architectural components. The aim of enhancing this knowledge base is to use the degree to which these were standardised originally as a means to inform and so support their ongoing care but also that of many other contemporary buildings. Although these libraries are numerous, they are maintained at a local level and as such, their shared challenges for maintenance remain unknown to one another. Additionally, this paper presents a methodology to indirectly retrieve useful indicators and semantics, relating to emerging HBIM families, by applying deep learning to a varied range of architectural imagery.

1. INTRODUCTION

1.1 Standardisation in modern heritage and the case of Carnegie Public Libraries

The philanthropic contribution of Andrew Carnegie to fund the erection of over 2000 public library buildings across Britain and America at the turn of the Twentieth century outnumbered any other single interest in the commissioning of a single public building type at the time. Despite the final library buildings resulting from competitions among different architects, Carnegie's common influence over the design of these buildings had a significant impact upon their standardisation (Van Slyck, 1998, Prizeman 2012). The design of public libraries themselves had become a highly refined and systematic field in which multiple modular elements of furniture would dictate standard dimensions for window sills etc. Indeed by 1911, in his 'Notes on Library Bilding' [sic] Carnegie's private secretary, James Bertram, recommended just 5 different standard library plans (Bertram,1911).

These buildings, along with all other public buildings of the time, were designed to maximise natural light and ventilation at a time when electric light was expensive and coal, as fuel for heating, was cheap. As a result, they responded to specific needs in engineering aspects whose priorities are now reversed. Given such peculiarities, these libraries therefore seem to provide a suitable and sufficiently large platform from which to explore the potential for the automated recognition of standard architectural components in early twentieth century buildings. This study is part of an Arts and Humanities Research Council (AHRC) funded research project aiming to investigate the potential for standard elements of Carnegie libraries to be adequately understood and suitably rehabilitated where necessary, so developing efficient methods for conservation practice.

The identification of such elements using deep learning is proposed here to recognise families of elements across a range of buildings. This will then facilitate subsequent matching of such families to their manufacturers through pairing them to illustrations in their contemporary trade literature. Future work will also address the indirect retrieval of useful indicators to relate to emerging HBIM families, such as geometric parameters and material specifications. This will include the creation of a series of HBIM elements particular to the era and therefore potentially relevant to a much wider range of buildings. To this end, patterns in the use of architectural elements in this more defined set are first found and, at a later stage, matched with a benchmark imagery dataset taken from the technical literature of the time. This way the degree to which elements of these and other similar buildings were originally specified is used to inform and support their ongoing care.

Specifically, this paper presents a methodology to apply deep Convolutional Neural Networks (CNNs) to a varied range of architectural images (e.g. standard frames, 360-degree views, scanned reproductions) to automatically recognise architectural components directly related to the engineering of ventilation systems in Carnegie library buildings. Hygienic design strategies with respect to ventilation for public interiors were iteratively refined at the time of their construction. The results will disclose how the most advanced tools in the fields of Computer Vision and Machine Learning can support the retrieval of relevant data

https://doi.org/10.5194/isprs-annals-IV-2-W6-123-2019

for the informed conservation of these and similar modern heritage buildings.
The paper first discusses precedents in the literature for the use of these techniques in Cultural Heritage (CH) conservation. Secondly, the workflow is presented and explained step by step. The proposed pipeline is illustrated through its implementation in the case of Carnegie libraries in the UK. The results show that the method enhances an understanding of the shared formal features and functional properties of these buildings such as common ventilation principles and lighting properties. This, in turn, could foster the adoption of fine-tuned sustainable development strategies both for public libraries and for other public buildings of the time. Suggestions about how it can be integrated in conservation practice, including surveying (sensor based or photogrammetric) and design (e.g. through CAD modelling) are presented in the final discussion.

1.2 Computer Vision and Machine Learning in CH studies

Machine learning and computer vision are fields of scientific research concerned with the development of algorithms that allow machines to respectively see and take decisions based on empirical training data. The use of computer vision techniques and machine learning in CH conservation is not unexplored territory. Precedents in the literature show applications in relation to the creation of multiple HBIM libraries (Murphy et al., 2013; Bruno et al., 2018). In addition, automated feature recognition for existing and historic buildings has a significant body of literature (Wang et al., 2015; Ochmann et al., 2016).

An active area of research in which Computer Vision has been successfully applied to the documentation of Cultural Heritage is in the analysis of the facades of historical buildings. Whilst early work required human experts to write grammars defining, for example, building styles, machine learning was subsequently applied, requiring only a set of images with pixel-wise annotations of defined architectural elements such as windows, doors, chimneys, etc. In their pioneering work, Martinovic and Van Gool (2013) explicitly learnt a two-dimensional attributed, stochastic, context-free grammar, which could then be used to either segment facade images or alternatively, synthesise new ones. Despite this, most of the literature to date focuses on image segmentation methods rather than on generative models. For instance, Teboul et al. (2013) apply reinforcement learning to parse images according to binary split grammars and Li &Yang (2016) perform image classification using the well-established conditional random field approach. The increase in machine learning activity has been facilitated by the creation of publicly available databases of annotated building facades, such as: the Ecole Centrale Paris (ECP) Facades dataset (Teboul et al., 2010); eTRIMS (Korc & Forstner, 2009); CMP Facade Database (Tyleček & Šára, 2013). Consequently, in recent years deep learning techniques have been explored in many studies. Liu et al. (2017), propose a deep learning method to segment building facades in semantic categories using symmetry rules and region proposal to refine the segmentation results. Fathalla & Vogiatzis (2017) suggest a novel method for the semantic segmentation of building facades integrating appearance and layout cues in a single framework. Schmitz & Mayer (2016) use CNN and transfer learning to enable the use of smaller datasets for deep learning applications in facade segmentation and interpretation.

1.2.1 Related work

Machine learning techniques developed in the field of computer vision, have quickly become key drivers for the many recent technical advances in the recording, digitisation and (data) mining of heritage buildings and monuments worldwide. A relevant work by Amato et al. (2015) has exploited a simple supervised machine learning technique based on the k-Nearest Neighbour (kNN) algorithm to rapidly classify Pisa's monuments and landmarks. Their algorithm processes local features, extracted from the images using SIFT (Lowe, 2004) and SURF (Bay, 2008) descriptors. Oses et al. (2014) perform an image-based delineation of masonry walls using 5 machine learning classifiers (kNN, Support Vector Machines, Probabilistic Classifiers, and Classification Trees). Grilli et al. (2017), in their review of point cloud segmentation and classification algorithms, highlight the following as suitable machine learning methods: K-means clustering, hierarchical clustering and mean shift. Recently, Grilli et al. (2018) proposed a supervised machine learning method to classify 3D heritage models by segmenting 2D textures using traditional Random Forests (RF).

1.2.2 Deep Learning

Among machine learning methods, deep learning refers exclusively to a sub-class of end-to-end machine learning techniques involving the use of Deep Neural Networks (e.g. CNN, deep reinforcement learning, GAN etc.); whose integration in computer vision applications has started only in the last few years. Specifically, the passage from "shallow" to "deep" methods happened in 2013, after the success obtained in a computer vision challenge of the first CNN model. In previous machine learning methods features such as vectors of shape measures, edge and colours distributions, feature points etc. were not automatically learned and the trainable classifiers, such as SVM kernels or Decision Forests, were often generic. In contrast, deep learning methods allow the machine to learn feature hierarchies all the way from pixels to classifiers. Each layer - there are many stratified ones with a similar structure performing different transformation functions - extracts features from the output of the layers below and above in a directly connected way.

Specifically, in CNN the multiple layers (the higher the number the deeper the model) are trained jointly, to provide a larger parametrisation space which is useful for retrieving complex relationships between inputs and outputs. To this end internal layers either: learn how to approximate the results of many classical feature extraction and image pre-filtering methods (convolution), provide a non-linear input to output mapping through a layer's activation function (non-linearity) or pool input layers into intermediate ones by applying filters at different locations (pooling). Given its capacity to optimise results towards a given problem, deep learning has rapidly gained the attention of the scientific community, including using it in applications concerning the built environment. For example, Lotte et al. (2018) address the issue of transferring labels of rendered images back to their 3D urban models combining CNN and Structure from Motion (SfM). Similarly, Kelly et al. (2017) use images and 3D models of urban scenes in combination with deep learning techniques to derive structured models of city blocks, addressing the automatic fusion of street-level imagery, polygonal meshes and GIS building footprints.

This paper aims to provide an alternative to the use of more classic machine learning methods previously proposed for CH classification such as traditional RF (not deep RF as in Zhou et. al., 2017). This enables to overcome the issue that RF needs features as input which are generated independently of the RF training process. In contrast, deep learning, with its end-to-end learning architecture, trains both representation learning and classification simultaneously. Hence, feature learning is implicitly tailored to the needs of the classification task, which is not possible when adopting a two-step process such as, for example, the classic RF.

https://doi.org/10.5194/isprs-annals-IV-2-W6-123-2019

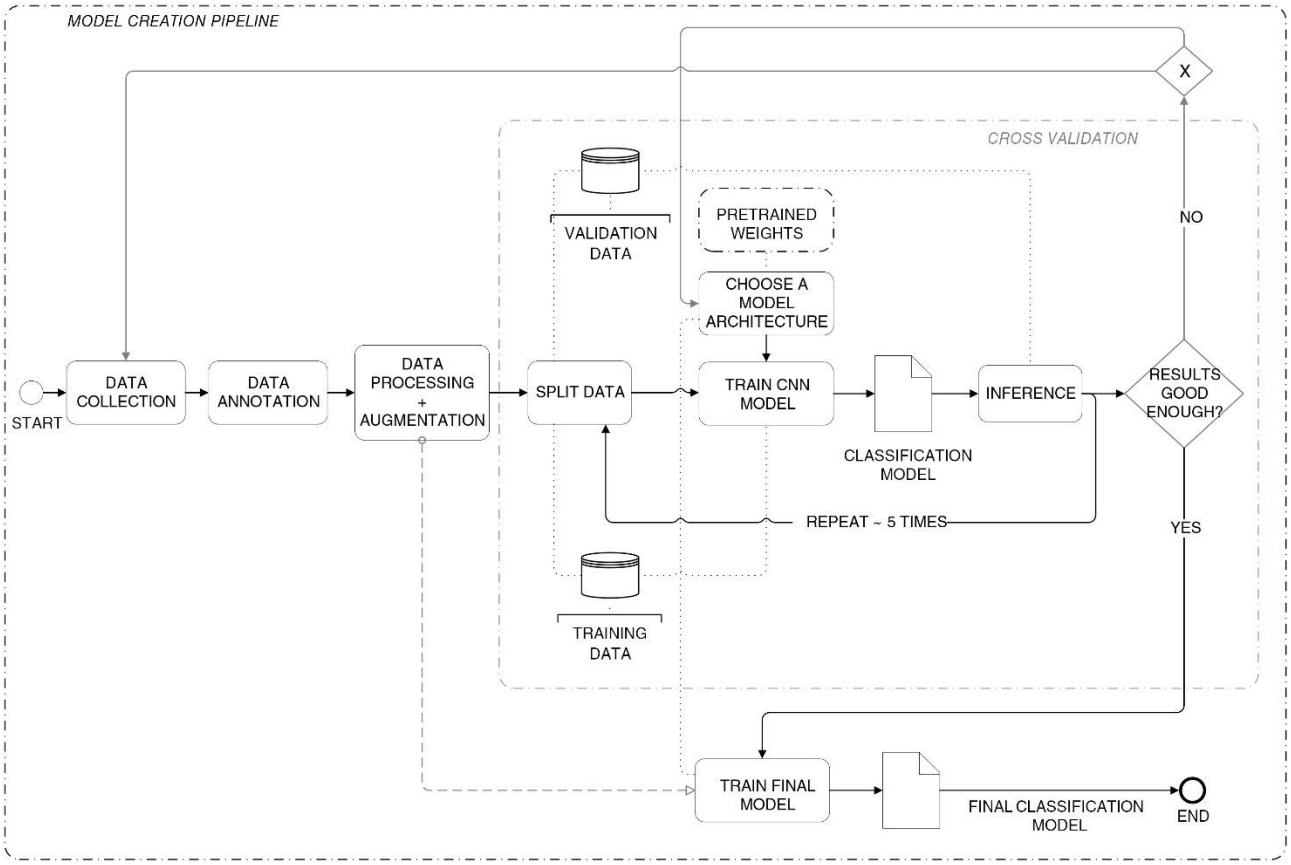

Figure 1. Workflow for creating the deep classifier for architectural components recognition, BPMN notation, ©Microsoft Visio.

2. CLASSIFICATION WORKFLOW

The general workflow for the deployment of the proposed deep learning classification method encompasses the following 5 steps: data collection; data preparation (annotation, processing and augmentation); CNN model construction (model architecture, training and cross-validation); final training; and ultimately, model exploitation with external data.

The model implementation phase covers the first 4 steps, which are the specific focus of this paper. A more detailed explanation of the pipeline as in the diagram above (fig. 1) shows the micro-steps involved in the implementation of the proposed supervised deep learning classification model. The adopted method is supervised, as previous knowledge is associated to the row input data (image) in the form of labels (class) prior to training with the aim to predict classes for new data entries. Conversely, an unsupervised method would work without prior knowledge of output class, producing internal self-evaluations to determine data patterns or groupings, which may or may not have direct correspondence with concepts in use in CH documentation applications. In the following paragraph the relevant sub-steps are explained in detail to allow an easy replication of the method.

2.1 The image dataset and data annotation process

CNN methods enable effective data-driven learning provided that a sufficient amount of training data is available for training the model. It is therefore common practice to exploit existing large-scale annotated image datasets with millions of categorised photographic images and adapt a model trained on one task to use it for another related one. However, similar datasets specifically built for CH imaging are not currently available and the specific complexities of tasks emerging in CH studies require the adoption of a fine-tuned, customised classifier.

2.1.1 Carnegie library dataset

Over the last two years, a comprehensive campaign of documentation has been undertaken across the United Kingdom (UK) to be able to provide high-quality data to record the condition of the Carnegie library building stock in the country today. It included building recording through laser scanning and photogrammetric surveys. Among other data, we collected a set of more than 13,000 images of the almost 600 Carnegie library buildings still standing in the UK. The photos present varied features in terms of lenses types (standard, zoom, wide-angle, 360), image dimension and resolution (made with crop, 360, full frame cameras and smartphones). This paper presents, for the first time, some preliminary results obtained using this new architectural imagery dataset. During the process of its creation, a GIS map was created, which set out the spatial distribution of these buildings across England (62%), Scotland (27%), Wales (8%) and Northern Ireland (3%).

2.1.2 Image annotation

The dataset retrieved from this survey was annotated using around 20 classes such as: wooden panelling, internal glazed partitions, internal and external skylights, barrel vaults, glazed domes and ceramic tiles, among others. The classes correspond to specific categories of original architectural components present in the interior and exterior architecture of the library buildings, which are relevant to the focus of this ongoing research project. This initial annotation step was done using ©Adobe Bridge, which makes it easy to assign labels to pictures and then to filter them accordingly. Next, around 2000 images have been sub-sampled from the initial dataset, by filtering out the images

that did not contain the labels required for the analysis of the libraries' ventilation systems. The relevant ones correspond to the following 4 classes:

a) ventilation turret (prefabricated ventilation components typically found on top of pitched roofs);
b) ventilation tower (built-in elements with the same function as the corresponding prefabricated ones);
c) ventilator grille external (customised or standardised);
d) ventilator grille internal (mainly standardised).

Starting from this reduced sample, the components were isolated within the images, using VGG Image Annotator (VIA). This is an online open source software developed by the Visual Geometry Group at the University of Oxford, that enables defining regions in images (e.g. bounding boxes) with associated textual descriptions. As a result, 3815 bounding boxes were retrieved with the following distribution: a) 701; b) 382; c) 1100; d) 1631. Ventilation towers were included as part of this study to check the accuracy of the algorithm in recognising subtle differences in similar components. For similar reasons, the study includes small elements, which are challenging to identify, such as internal and external grilles.

2.2 Data processing and augmentation

In order to create the image classification dataset, each sub-image delimited by a bounding box is first cropped away from the initial images. As these samples are extracted from larger images, randomness in the crop size and centering can be introduced. This added variability results in a more realistic dataset where objects of interest are not necessarily perfectly centered nor normalized. Aside from training a model that is able to discriminate the diverse architectural components from one another, it seems also important to distinguish these objects from the background. To that end, random extracts containing none of the objects of interest are sampled randomly from the initial images. The inclusion of background images as a concrete classification class is even more important in the event that the final classifier is used as part of an object detection pipeline. Indeed, detection requires the correct discrimination of objects from the background.

Data augmentation has been proven to be an important driver of performance in modern computer vision pipelines (Perez et al. 2017). Therefore, in this work, a wide range of data augmentation techniques were performed on the initial dataset. More precisely, the final complete dataset is comprised of several copies of each initial image after application of various random transformations such as variations in colour balance, contrast, brightness, sharpness or rotation angle. As its name suggests, this process artificially increases the dataset size and its richness without any requirement for additional data collection. Furthermore, the use of such a dataset for training produces models that are inherently more robust to variabilities in all transformations applied during the augmentation. Overall, there is generally no major drawback to implement data augmentation, while its advantages can have a significant impact on the quality of the learning process. The challenge is to incorporate sufficient transformations that capture the expected variability in the real world.

A principle similar to that of data augmentation can also be applied for prediction. Thus, just as the application of many different transformations on the training dataset for training yields a model that is more robust to such transformations, predictions can be made on images with different characteristics in colour, contrast, sharpness etc. Hence, as for training augmentation, one can first make several copies of any test sample with different transformations and then feed these to the network. This process produces an entire array of predictions for each test sample. The final prediction is then obtained by aggregation of these individual predictions. The benefits of such a technique are twofold: the final predictions display less variance and most importantly they are generally more accurate.

2.3 Model architecture and training

In recent years, deep neural networks have displayed state-of-the-art results in numerous fields ranging from natural language processing to action recognition in videos (LeCun et al., 2015; Schmidhuber, 2015). As image classification is not an exception, this work will focus on deep learning models to solve the architectural object classification problem described above. This section presents the two deep architectures selected for this task as well as a classic machine learning model. The latter is used in section 2.5 to measure the potential performance gain in using end-to-end trained models with learnt features rather than multi-steps ones with hand-crafted features.

2.3.1 Traditional machine learning benchmark model

In classic machine learning methods feature design and classification are performed separately, which means that classic machine learning models are typically multi-phased processes. In other words, different algorithms are applied one after another. This paper presents a Traditional Machine Learning (TML) pipeline composed of a sequence of three steps: SIFT, K-means clustering, and standard RF. As will be discussed below, in all the three phases a set of alternative algorithms can be used to substitute or to coordinate with those suggested here. However, the proposed benchmark model is chosen as it is representative enough of common machine learning workflows and because its performance is expected to be at the higher end of the spectrum. A good performance is in fact required to get a meaningful result out of the final comparative test.

In this model, firstly, the images' keypoints are detected and described using SIFT. At this point each image is represented by an inhomogeneous bag/collection of features, whose number varies from case to case. Since machine learning algorithms work better with well-defined inputs, in a second step, a bag of visual word is generated by performing a classical K-means clustering (MacQueen, 1967; Steinhaus, 1956) with k=50 on all SIFT features of the training images dataset. Each image is then described by mapping each SIFT feature to its visual word (i.e. cluster) and by computing the distribution of occurrences of each word in the image. Here, the number of visual words is kept low to prevent overfitting and its subsequent negative influences on the remaining part of the object recognition pipeline. However, a significant discrepancy between in-sample and out-of-sample performance suggests that in our case using a small value of k is not enough to fully prevent the benchmark model to overfit the training data. A possible alternative to obtain uniform image representations is to use other algorithms such as, for instance, a PCA (Pearson, 1901). Finally, a standard RF (Breiman, 2001) of 100 trees is used to perform the classification using as input the visual word occurrence distribution of the training samples. Again, other classification models such as Support-Vector Machines (SVM), (Cortes and Vapnik, 1995) could have been used in the third step instead of RF, but the latter is inherently more suitable for multi-class problems. The final performance score (see table 1) is computed by applying the trained random forest model on the test samples.

https://doi.org/10.5194/isprs-annals-IV-2-W6-123-2019

2.3.2 A deep learning benchmark model

As a deep learning benchmark, we propose the use of a classical convolutional neural network (CNN). The model differs from well-known architectures (Krizhevsky et al., 2012; Simonyan et al., 2014) only in its number of convolutional layers, filters and nodes. Indeed, since the architectural dataset at hand is smaller than the standard datasets used for deep learning, which typically reach the size of millions of images (Deng et al., 2009), the model size is kept moderately small to prevent overfitting. More precisely, the representation learning part of the network is comprised of seven (3x3) convolutional layers with 16 to 32 filters, intertwined with ReLu activations and max-pooling layers. This first step transforms the (512x512) input images into 32 meaningful convolutional feature maps. After that, two fully-connected layers followed by a softmax transformation are used to map the representation space to class probabilities.

2.3.3 Feature Pyramid Network

Standard convolutional neural networks have demonstrated outstanding performance on a wide variety of tasks. Their structure is however not developed to cope easily with the detection or classification of objects of different scales. A scale invariance property is important in the context of architectural objects. To that end, an alternative architecture is tested, namely the feature pyramid network (FPN) (Lin et al., 2017). Throughout the years, feature pyramids have been used in numerous computer vision tasks to cope with scale variability. The FPN model simply applies this classical concept to deep convolutional neural networks. On a more technical level, once again, the number of filters is adjusted to suit the quite limited size of available data; so, each convolutional layer is comprised of 16 to 32 filters. In addition, as our input image size differs from that of the original research (we use larger images), additional convolutional and max-pooling layers are used to produce the final convolution representations that are then fed to the fully connected-layers.

2.3.4 Training

Regardless of their architecture, both deep learning models are trained using the gradient-based Adam optimizer (Kingma et al., 2014). A total of 30000 training steps with 32 images per batch are used to complete the training. A link to the full model architecture, TensorFlow implementations and additional details will soon be available on the official website of the AHRC funded "Shelf-Life; reimagining the future of Carnegie public libraries" project.

2.4 Model Evaluation

In order to assess the performance of the three presented models (TML, CNN and FPN), the dataset is first split into train and test sets. The same training and testing splits have been used for the classical machine learning benchmark and for the two deep learning models in order to obtain an unbiased comparison of relative performance in image classification. In contrast to classical cross validation methods, no validation set is used, as no hyper-parameter optimization is performed. Furthermore, in our pipeline the test set is not simply sampled uniform randomly from the dataset. Indeed, images taken at similar dates and times are regrouped and automatically put into the training set. This criterion ensures that similar images cannot end up in both training and test sets, which would certainly bias the final results. In addition, the test set is further sampled in such a way that each class is represented by a sufficient number of images. This manual balancing of the test sets ensures that the various per-class performance measures are not the product of single (or a few) predictions, thus both reducing the variance and increasing the relevance of the result. Once the separation of the dataset is done, the model is trained using solely the training set. After completion of the learning phase, the test images are fed to the network and the outputted predictions are then compared to their true class. More precisely, the per-class f1-score (which is the harmonic mean between precision and recall) and the overall mean f1-score are chosen as measures to assess the model performance:

$$F1 = 2* (precision * recall) / (precision + recall) \quad (1)$$

This entire process is then repeated 5 times, and the final cross-validated results are defined as the mean result over all independent evaluations. Performing the splitting of the sets and the evaluation several times assures a greater variety of test sets and allows most images to be part of the test set at least once. The final f1 averaged measure is known to be a good performance indicator for the two proposed models on new unseen input data.

2.5 Performance and Results

The performance of the traditional machine learning benchmark (TML), classical CNN and Feature Pyramid Network (FPN) on the architectural object classification task is summarized in Table 1 and Figure 2. Overall, with a F1-score of around 80%, the results of the deep learning models are promising, especially if compared to the 56% achieved by the TML model. The TML overall results are quite disappointing with each class scoring significantly below its deep learning counterparts (it is however noted that a problem with overfitting as described in paragraph 2.3.1 might have affected the results of the TLM). If we retrain the comparison to those, it is possible to see that the more advanced FPN architecture performed, as expected, slightly better than the classical CNN architecture.

Perf.	Turret a	Tower b	Gr.ext c	Gr.int d	Othe r	Mean
TML	0.60	0.15	0.62	0.61	0.82	**0.56**
CNN	0.85	0.60	0.77	0.78	0.91	**0.78**
FPN	0.88	0.67	0.81	0.80	0.92	**0.82**

Table 1. Comparative analysis of classification performance (F1 measures) for the machine learning benchmark and the two proposed deep learning architectures

A few additional class-specific observations can be made for all models:

- First, the confusion matrices in figure 2 reveal that distinguishing between exterior and interior ventilation grilles is challenging, which could be expected given the similarities that are often encountered between these two object classes. This phenomenon, although present also in the deep learning models, is much more significant in the TLM. This effect could be alleviated by feeding images (and bounding boxes) that are more loosely focused on the components, thus providing more information about context and background.
- Secondly, the recognition performance of ventilation towers, for which the number of observations was comparatively limited, is significantly lower than that of any other classes. This highlights the importance of gathering enough training samples for each object class to make the best use of the proposed methods. Here, the TML model's performance is particularly low, with only 15% of success rate for class b.

https://doi.org/10.5194/isprs-annals-IV-2-W6-123-2019 | © Authors 2019. CC BY 4.0 License

- Third, a non-negligible number of objects have been classified as ‘other’, which is an indicator of the dataset’s complexity.

One important consideration is unique to the TML benchmark:

- Class a, corresponding to the ventilation turrets presents a lot of predictions, but almost half of them are false.

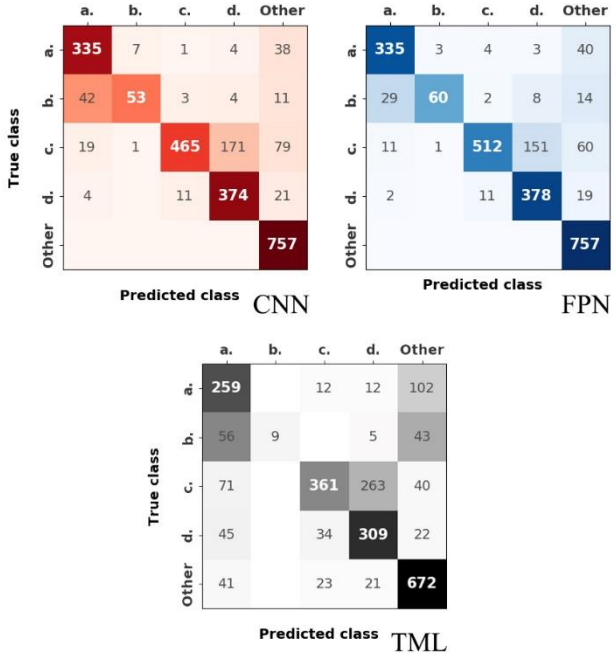

Figure *2*. confusion matrices of benchmark TML model, classic CNN model and Feature Pyramid Network (FPN) model

Noticeably, the TLM model is not as fast to train as one might expect as the K-means clustering of the SIFT features to generate the visual words results is computationally intensive. Overall, the model is still faster to train than both the CNN and the FPN, however the limited time gain does not compensate for the significant loss of performance for most applications, including those presented in this paper. On top of this, we can expect this computation burden to further increase if the TML model is applied to a larger dataset. The emergence of such undesirable properties in the TML model is rather hard to explain given that the two deep learning models, which have been trained on the exact same dataset, do not display the same behaviour. One tentative explanation for this performance discrepancy could possibly be found in the characteristics of the input data. In fact, there are numerous photos in the collected image dataset displaying only a few SIFT features (< 5), meaning that the dataset information content is not suitably described by means of these features alone.

The TML’s results might have benefitted from augmentation of local SIFT features with other descriptors, and/or from fine-tuning the representation mapping as well as the classification model. However, the outcome of our tests shows that in this and other similar cases, deep learning models (even with no hyper-parameter tuning) clearly outperform classic TML methods. Furthermore, the outcomes of the performance test demonstrate that classical hand-defined feature descriptors, including SIFT, SURF and HOG, might not be fully optimal for the dataset analysed here and, possibly, also for other similar ones. Conversely, by learning their own feature representation based solely on the training data, deep neural networks can tailor their predictions based on a specific dataset independently of the degree of its inherent complexity; which better suits the characteristics of the data at hand.

As shown by Figure 3, which depicts the 3 best and 3 worst predictions of a specific run for each object class, the classification problem is far from trivial. Indeed, the image reveals that some samples are very challenging. For instance, the three worst predicted towers include an image totally occluded by a tree, a rotated image and a low-resolution image with high levels of distortion. In addition, Figure 3 reveals how diverse each object class is in terms of general shape, structure and background. In this context, the final overall high classification accuracy of CNN and FPN confirms the potential benefits of adopting deep learning models for automating the recognition of complex architectural elements.

Figure 3. deep learning best and worst predictions sample

2.6 Beyond the Black Box

Viewed negatively, deep learning models have long been considered opaque black boxes that operate without any human interpretability. However, over the years, understanding the representation learning and decision process underlying these models has become an active area of research (Simonyan et al., 2013; Zeileret al., 2014; Zhang et al., 2018). In this section, one of the many currently available interpretability tools, namely Grad-CAM, is applied to the classification model presented above (Selvaraju et al., 2017).

https://doi.org/10.5194/isprs-annals-IV-2-W6-123-2019

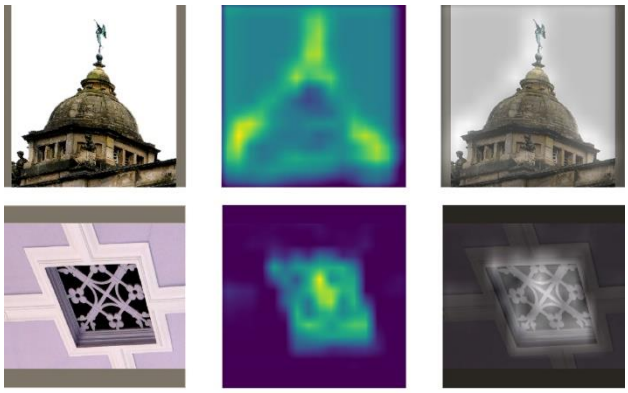

Figure 4. Grad-CAM heat maps

Informally, for each input, this method determines which regions of the image are the most relevant for classifying the object correctly. This provides some information about where the model is focusing on to make its predictions. In Figure 4, the coarse Grad-CAM activation map is presented for two different images. It can be observed that, in this specific case, the proposed deep learning model focuses on the general 3D shape of the tower (less on its ventilation components) and on the articulated structure of the internal ventilator grille, rather than the background, to make its decisions.

This example shows how such interpretability tools can provide us with a better understanding of the CNN internal decision-making process. This is particularly relevant when applied to architectural imagery as it may help in disclosing deep analogies useful for highlighting complex commonalities (e.g. morphological patterns) among heritage buildings as well as among their parts.

3. OBJECT DETECTION

The previous paragraphs have presented a deep learning classification pipeline that is able to successfully recognise images of a relevant set of architectural components such as ventilation turrets, towers and grilles. The main restriction of the proposed model resides in the assumption underlying the input images: the object of interest has to be the central element of the image and well-defined in order to be precisely recognized. However, these limitations can be alleviated by extending the current object recognition model to become a complete object detection model. A trivial extension of the proposed classifier to an object detection model would be to use a sliding window to scan through the image and attempt to recognize the presence of objects within each of these sub-images. This simple exhaustive search method presents a wide range of limitations such as the fixed detection scale and the slow processing time, which make this method unsuitable for most practical applications. As an alternative, region proposal algorithms could be used; these methods generate numerous bounding boxes representing areas where objects are more likely to appear. The presence and localization of objects can then be determined by classifying each of the sub-images defined by these bounding boxes.

In this work, class-independent region proposals are generated using selective search (Uijlings et al., 2013), similar to that of the classical R-CNN object detection model (Girshick et al., 2014). The algorithm groups pixel regions based on colour, texture, size and shape similarities. This method is known for its high recall, meaning that the set of bounding boxes is likely to contain our objects of interest.

Each sub-image delimited by a bounding box is then fed to our FPN object recognition model as defined in Section 2.3.2. This produces a probability estimation for the presence of an object of interest in the image. In the event that the probability of recognition reaches a value above an arbitrary threshold, the specific bounding box is considered to contain the corresponding object of interest. Finally, as the region proposal algorithm produces several thousand different bounding boxes, the process is bound to produce overlapping detections and unwanted false-positives. Therefore, as a last step, a non-maximum suppression has to be applied to clean the detections. The entire process is summarized in Figure 5.

As shown by the results on three test images in Figure 5 and 6, the proposed object detection pipeline appears suitable for the detection of architectural components. These examples underline however two limitations of the proposed pipeline. First, an exterior ventilation grille was not detected on the leftmost image. This omission is due to the fact that no bounding box was proposed around that object by the region proposal algorithm, thus hinting at the need for a task specific region proposal model. Second, the scale of one of the detections in Figure 5 is significantly too large. The scale invariance that the FPN attempts to achieve is likely the cause of it, indicating that the use of scale sensitive object recognition models might be more preferable as part of object detection pipelines.

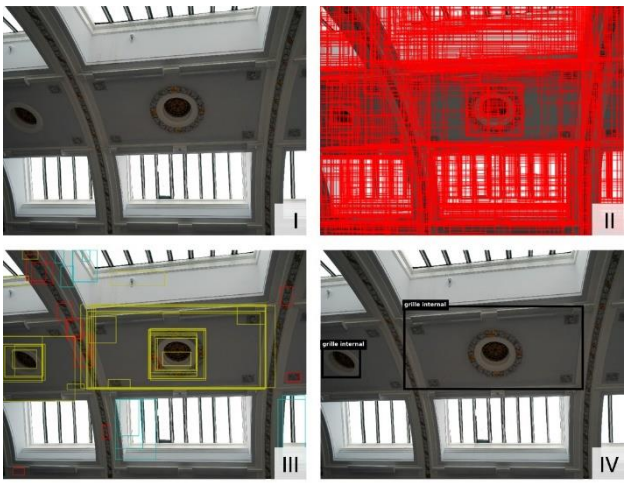

Figure 5. Detection process phases: I) original image; II) regional proposal; III) all-region prediction; IV) final prediction

Nevertheless, a larger dataset is required for a more in-depth analysis. In future work, other more advanced region-proposal based detection approaches could be investigated such as Fast R-CNN (Girshick, 2015) or Faster R-CNN (Ren et al., 2015). The regression-based YOLO detector (Redmon et al., 2016) also stands as a relevant option. In any of these cases, transfer learning (Weiss et al., 2016) will have to be leveraged for training, since the model complexity far outweighs the richness of the available dataset.

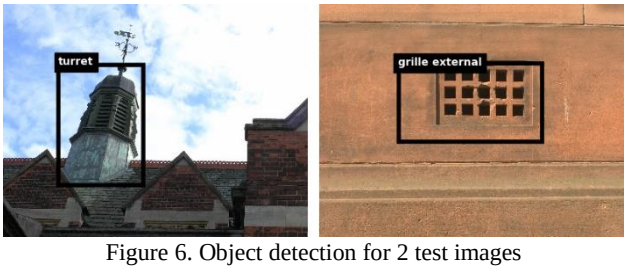

Figure 6. Object detection for 2 test images

https://doi.org/10.5194/isprs-annals-IV-2-W6-123-2019

4. CONCLUSIONS

In order to better inform decision-making strategies regarding the conservation or adaptation of these early 20th century buildings, it is crucial to take advantage of progress in other disciplines. This means to adopt an interdisciplinary approach and exploit the most advanced computational modelling and analysis techniques available to date. For architects, mage recognition does not in itself indicate the probable construction layers built beneath a building component. Nonetheless, by collating such data with technical literature and specifications of the period and focussing on manufactured elements as opposed to bespoke pieces, steps are made towards enabling educated guesses as to the likely make up of that 'invisible' information. In this preliminary study the automated recognition of elements, which are tiny, difficult to spot, blocked or simply elements that have been altered over time, such as the internal and external ventilator grilles, has proved useful to highlight some of the key principles underlying the design of Carnegie library buildings. It illustrates the consistency of their design, which is demonstrated by the careful engineering of their ventilation systems and so supports a better understanding of the consequences of alterations to the buildings' environmental functions. In principle this could assist in the swifter application of relevant environmental principles to guide practitioners in the analysis of similar historic buildings, thus speeding up the initial visual analysis of the building and supporting the decision-making process of experts.

The main difficulties in the Scan-to-BIM process to date in both academia and practice relate to issues of investigating buildings' geometry and identifying the shapes and structures to be captured. Professionals still rely heavily on orthogonal drawings to share design and production information. There is a challenge in the capacity to interpret 3D scenes produced by technologies such as laser scanners in the form of point clouds, to distinguish between ambiguities and then create coherent HBIM models. In order to speed up the survey process, practitioners may tend to avoid the recording of RGB values while scanning buildings (unless a coloured point cloud is required by the client). This makes the visual support offered by complementary high resolution 2D images extremely important to identify the nature of the architectural, structural and/or MEP elements surveyed (e.g. to distinguish pillars from cupboards, cable trays from beams, or even electrical from cardboard boxes). Furthermore, pictures are often used to check the 3D model for Quality Assurance prior to the delivery of the final models to the client.
This being the case, the proposed method not only potentially assists the path towards further advances in the field in time, but, if suitably adapted, could also offer useful support for use in contemporary practice. The key advantages of using the presented deep learning (FPN) classifier instead of classical machine learning methods, are its emphasis on building components and its robust architecture. This means that the proposed method can accept, as input data, photos collected for photogrammetry, which typically capture buildings in fragments and thus may include incomplete representations of components.

4.1 Future work

There is potential for this workflow to inform the creation of valuable datasets augmenting the process of conservation through the creation of richer parametric libraries in the service of HBIM. HBIM parametric families could be created in a two-step process, namely Photos-to-Specifications-to-BIM, by firstly automating the collection of relevant technical specifications and then transferring this knowledge into a set of 3D parametric models of standard components. A library of CAD elements crafted in such a way would be a precious resource to deal with the complexities involved in the refurbishment and renovation of early 20th century buildings. Among other things, it would enable a faster deployment of all kinds of simulations. Furthermore, the combination of HBIM and automated image classification systems would: foster quality control during the diagnosis, design and construction phases; enable rapid interventions in case of hazardous events; as well as simply foster awareness in the ongoing care of heritage buildings among all stakeholders.

Future research will hence address: (i) the classification of other building components with highly complex shapes, such as (glazed) barrel vaults and (glazed) domes, which bring specific challenges to the object recognition and detection tasks such as dealing with reflective surfaces and sharp gradients of lighting conditions in the images; (ii) the matching of representations found in the trade literature (e.g. pictures, drawings, architectural representations) with the photos of corresponding building components; (iii) the creation of semantically rich parametric families of objects; (iv) creation of a shareable HBIM library of standardised components for early 20th century buildings. Future work could also tackle the issue of 3D point cloud semantic segmentation using deep learning (deep segmentation) and that of the automation of positioning of HBIM objects within 3D point clouds of suitably surveyed heritage buildings.

5. ACKNOWLEDGMENTS

The Arts and Humanities Research Council, (UK) have funded this project: *Shelf-Life; Re-imagining the future of Carnegie Public Libraries*. The project is led by Dr Oriel Prizeman at the Welsh School of Architecture, with co-investigators Professor of Geographical Information Systems, Chris Jones at the School of Computer Science and Informatics, Cardiff University and Professor Alistair Black from the School of Library and Information Science, University of Illinois at Urbana-Champaign.

The Arts and Humanities Research Council (AHRC) funds world-class, independent researchers in a wide range of subjects: ancient history, modern dance, archaeology, digital content, philosophy, English literature, design, the creative and performing arts, and much more. The quality and range of research supported by this investment of public funds not only provides social and cultural benefits but also contributes to the economic success of the UK. For further information on the AHRC, please go to: www.ahrc.ac.uk

REFERENCES

Amato, G., Falchi, F., Gennaro, C., 2015. Fast Image Classification for Monument Recognition. *Journal on Computing and Cultural Heritage*, 8(4), 18:1-18:25.

Bay, H., Ess, A., Tuytelaars, T. and Van Gool, L., 2008. Speeded-up robust features (SURF). *Computer vision and image understanding*, 110(3), pp. 346-359

Bertram, J., 1911. *'Notes on the Erection of Library Bildings'*. New York, Columbia University, Rare Book and Manuscript

Library, Carnegie Collections, CCNY Records, series VIII, Printed Material A.3, 48(1), Miscellaneous Pamphlets.

Breiman, L., 2001. Random forests. Machine learning, 45(1), pp. 5-32.

Bruno, S., De Fino, M., Fatiguso, F., 2018. Historic Building Information Modelling: performance assessment for diagnosis-aided information modelling and management. *Automation in Construction*, 86, pp. 256-276.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, 20(3), pp. 273-297.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009 Imagenet: A large-scale hierarchical image database. In CVPR Conference on Computer Vision and Pattern Recognition.

Fathalla, R., Vogiatzis, G., 2017. A Deep Learning Pipeline for Semantic Facade Segmentation. In BMVC British Machine Vision Conference.

Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR Conference on Computer Vision and Pattern Recognition, pp. 580-587.

Girshick, R., 2015. Fast R-CNN. In IEEE international conference on computer vision, pp. 1440-1448.

Grilli, E., Menna F., Remondino, F. 2017. A review of point clouds segmentation and classification algorithms. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, pp. 339-344.

Grilli, E., Dininno, D., Petrucci, G., Remondino, F., 2018. From 2D to 3D supervised segmentation and classification for cultural heritage applications. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 62, pp. 399-406.

Kelly, T., Femiani, J., Wonka, P., Mitra, N.J., 2017. BigSUR. *ACM Transactions on Graphics*, 36, pp. 1–16.

Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv*:1412.6980.

Korc, F. & Forstner, W., 2009. eTRIMS Image Database for interpreting images of man-made scenes. Available at http://www.ipb.uni-bonn.de/html/projects/etrims/.

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pp. 1097-1105.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *Nature*, 521, pp. 436-444.

Li, W., & Yang, M. Y., 2016. Efficient Semantic Segmentation of Man-Made Scenes Using Fully-Connected Conditional Random Field. *ISPRS International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 41, pp. 633-640.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S., 2017, July. Feature pyramid networks for object detection. In CVPR Conference on Computer Vision and Pattern Recognition, 1(2), pp. 2117-2125.

Liu, H., Zhang, J., Zhu, J., Hoi, Steven C. H., 2017. Deepfacade: A deep learning approach to facade parsing. Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI 2017, pp. 2301-2307.

Lotte, R. G., Haala, N., Karpina, M., Aragão, L. E.O.C., Shimabukuro, Y. E. 2018. 3D Façade Labeling over Complex Scenarios: A Case Study Using Convolutional Neural Network and Structure-From-Motion. *Remote Sensing*, 10: 1435.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), pp. 91-110.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1 (14), pp. 281-297.

Martinovic, A.,Van Gool, L., 2013. Bayesian grammar learning for inverse procedural modeling. In CVPR Conference on Computer Vision and Pattern Recognition.

Murphy, M., McGovern, E., Pavia, S., 2013. Historic Building Information Modelling – Adding intelligence to laser and image based surveys of European classical architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 76, pp. 89-102.

Ochmann, S., Vock, R., Wessel, R., Klein, R., 2016. Automatic reconstruction of parametric building models from indoor point clouds. *Computers & Graphics*, 54, pp. 94-103.

Oses, N., Dornaika, F., & Moujahid, A., 2014: Image-based delineation and classification of built heritage masonry. *Remote Sensing*, 6(3), pp. 1863-1889.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), pp.559-572.

Perez, L. and Wang, J., 2017. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv*:1712.04621.

Prizeman, O., 2012. *Philanthropy and light: Carnegie libraries and the advent of transatlantic standards for public space*. Farnham, Surrey, England, Ashgate.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., 2016. You only look once: Unified, real-time object detection. In CVPR Conference on Computer Vision and Pattern Recognition, pp. 779-788.

Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, pp .91-99.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, pp. 618-626.

Simonyan, K., Vedaldi, A. and Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv*:1312.6034.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. *Neural Networks*, 61, pp. 85-117.

Schmitz, M., Mayer, H., 2016. A Convolutional Network for Semantic Facade Segmentation and Interpretation. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 709–715.

Steinhaus, 1956. Sur la division des corps materiels en parties. Bulletin de l’Académie Polonaise des Sciences, Classe III, 4(12) pp. 801-804.

Teboul, O., Kokkinos, I., Simon, L., Koutsourakis, P., Paragios, N., 2013. Parsing facades with shape grammars and reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7), pp. 1744-1756.

Teboul, O., Simon, L., Koutsourakis, P., Paragios, N., 2010. Segmentation of building facades using procedural shape priors. In CVPR Conference on Computer Vision and Pattern Recognition, pp. 3105-3112.

Tyleček R., Šára R., 2013. Spatial Pattern Templates for Recognition of Objects with Regular Structure. In: Weickert J., Hein M., Schiele B. (eds) Pattern Recognition. GCPR 2013. *Lecture Notes in Computer Science*, 8142. Springer, Berlin, Heidelberg.

Uijlings, J.R., Van De Sande, K.E., Gevers, T. and Smeulders, A.W., 2013. Selective search for object recognition. *International Journal of Computer Vision*, 104, 2, pp. 154-171.

Van Slyck, A. A., 1998. *Free to all: Carnegie libraries & American culture*, 1890-1920. Chicago, University of Chicago Press.

Wang, C., Cho, Y. K., Kim, C., 2015. Automatic BIM component extraction from point clouds of existing buildings for sustainability applications. *Automation in Construction*, 56, pp. 1-13.

Weiss, K., Khoshgoftaar, T.M. and Wang, D., 2016. A survey of transfer learning. *Journal of Big Data*, 3:9.

Zeiler, M.D. and Fergus, R., 2014, September. Visualizing and understanding convolutional networks. In European conference on computer vision, pp. 818-833. Springer, Cham.

Zhang, Q.S. and Zhu, S.C., 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1), pp. 27-39.

Zhou, Z.-H., & Feng, J. 2017. Deep Forest: towards an alternative to deep neural networks. In IJCAI 26th International Joint Conference on Artificial Intelligence. pp. 3553-3559.