Automated scale invariant interpretation of aerial imagery: from tennis courts to trampolines

Abstract—**For many years it has been thought that automated map generation from aerial imagery could be possible, and it has often been studied within the National Mapping Agency of Great Britain, Ordnance Survey (OS). The ImageLearn project aims to accomplish this using OS imagery and Convolutional Neural Networks (CNNs). However, it is well known that scale is an inherent problem within imagery and ignoring it can lead to poor classification accuracy. Combining aerial imagery, CNNs and scale has previously been done only in a pre-processing stage, thus the proposal for the MSc thesis is to implement a GoogLeNet style CNN with aerial imagery. Since scale will be inherent in the CNN, this will produce better results.**

A. *Introduction and Motivation*

Within Ordnance Survey there is a distinct impression that more benefit could be gained from capturing aerial imagery. It is used for verification, the Terrain products and as a product in itself. However, it has been thought for many years that automated map generation and image interpretation from aerial imagery should be possible for computers. Indeed as early as 1959, [22] foresaw the use of Geographic Information Systems and automated techniques. Yet this early promise was not realised due to issues with map generalisation (a technique to take one scale and create coarser scale products) [16].

With the advances of deep learning and new computer vision techniques this goal is being revisted in the ImageLearn collaboration between the University of Southampton and Ordnance Survey. Specifically, this project aims use state-of-the-art methods to extract the semantic meaning from OS aerial imagery that normally only a trained surveyor would be be able to distinguish. However, this project is not explicitly focussing on incorporating scale, the effects of such a decision have not yet been fully quantified. It is likely that it will not generalize well to the British landscape, due to the importance of size in the real world. Without incorporating scale, then picking objects of constant size, especially standardized items such as tennis courts and shipping containers, should be well executed. However, there are many items which a human interpreter would categorise similarly but a computer would not due to size differences . An example of this is trampolines, which come in a variety of sizes (8 - 16 feet). In the ImageLearn project it was found that searching using trampolines of a certain size would not find differently-sized trampolines. This concept will affect many other features in aerial photography and needs to be dealt with. Thus the problem can be stated as: *How to recognise similar objects in aerial images when objects can be a range of sizes*, which is known colloquially as 'The Trampoline Problem'.

In this review, section B introduces aerial imagery and its particular issues and, in section C the workings of Convolutional Neural Nets (CNNs) and a history of their usage is covered. Section D examines methods of introducing scale into models. Within section E the literature overlapping between the three previous subjects (Aerial Imagery, CNNs and Scale) is discussed and critiqued. Finally in section F, the gap in the literature is outlined and a topic for the MSc thesis is proposed.

B. *Aerial Imagery*

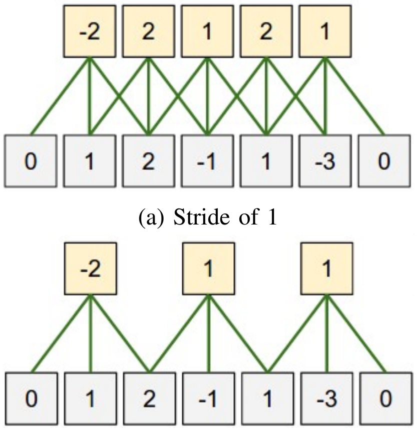
Manual extraction of data from aerial imagery has been performed for over a hundred years, but has continued to progress throughout this period. However, as Tranowski [23] points out, the explosion in photography available means that it is beyond human capability to complete, and inadvisable to continue doing so. Ordnance Survey wishes to make more use of its data, especially with around 1 GB of data for each 25km^2 of the UK, with a times series going back to 2001. Fewer articles have been written on machine learning using aerial imagery than on satellite imagery. [2], [14] and [17] are three papers comparing machine learning methods, but all use satellite imagery. This is due to the nature of satellite imaging, which usually provides a frequent time series of images of larger areas, whereas aerial imagery provides less of a time series, generally smaller areas, but higher resolution. Usually automation is less important for aerial imagery because of this, however Ordnance Survey is an outlier in this case, having a time series going back many years for the whole of the UK.

C. *Convolutional Neural Networks (CNNs)*

Convolutional Neural Networks operate in the same way as conventional Neural Networks with some additional processing stages. They introduces the idea of shared weights rather than having weights for every pixel and links to every node [5]. It also introduces a number of different types of layers in the hidden layers. The first type unique to CNNs is the convolution layer where learned filters (k filters of size $m \times m \times 3$) convolve with with the image (of size $n \times n \times 3$) creating a output map to the next layer of size $(m \times m \times k)$. With the use of striding (Figure 1) this significantly reduces the number of weights and resulting outputs compared to per-pixel weightings, whilst also avoiding overfitting.

Generally after each convolution layer there is a pooling layer which further reduces the size of the network. The most common type of pooling is taking the maximum of a particular slice which is specified using spatial extents and striding [5]. However, this is thought to reduce the spatial information content, according to [21]. The benefits of pooling is that it can be used in place to fully connected layers, which seem to offer better performance results [21] as well as creating invariance to position [11].

CNNs were often used to create deep networks before other types of neural nets, due to the error gradient training method developed by Rumelhart [18]. CNNs have had particular success within image classification and pattern recognition [1], having been used to win the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by several constestants [19].



(a) Stride of 1

(b) Stride of 2

Fig. 1: Example of striding. The image inputs are convolved with the kernel [1, 0, -1]. (a) shows all inputs have weights going to fully connected (FC) hidden layer. (b) shows the offset for the next set of weights. Recreated from [5].

However their prevalence within this domain did not materialise until 2012 where a "turning point" occured[19, p.17], when Krizhevsky trained a deep CNN and won ILSVRC. Prior to this, the Support Vector Machine was king with stochastic [8] and linear [13] types being used, together with feature engineering of the images. In 2012, Krizhevsky et al. [6] won the ILSVRC with a significant lead (11% less error) on all other applicants. In the successive years (2013 and 2014), the techniques used were mostly convolutional neural networks. Due to the competitive nature of the challenge, the methods have continued to progress, with the winning entry's error decreasing year on year [19]. The ILSVRC provides a good insight into the practices of the computer vision community with regard to algorithm changes. Every year since 2012 the winning teams have relied heavily on CNNs, which have significantly better performance than any other current method. This is not only for image classification but also single-object localization and object detection (the other two tasks within the ILSVRC). It is for this reason that CNNs are being used in the ImageLearn project.

So far the focus has been on CNNs with respect to image classification. Yet this is not the sole aim of the ImageLearn project, which is investigatinf an image retrieval task as a test of success. Techniques used in the ILSVRC up until the 'turning point' are still prevelant in image retrieval. These feature engineering techniques often use SIFT descriptors encoded using Bag of Visual Words (BoVW), or Fisher vectors [12], which are then classified. Recently it has been suggested that CNNs trained for image classification could be put to use in other visual recognition tasks such as image retrieval [12].

D. But what about scale?

The idea of scale is summed up by [9] as "*...[A] property of objects in the world is that they only exist as meaningful entities over certain ranges of scale*". It doesn't make sense to look at continents with a microscope; certain objects will only be viewable at certain scales. This is a particular problem in aerial imagery because objects that are considered the same by us can be very different sizes (such as an industrial warehouse and a shed). There have been many methods developed, in search of the 'solution' to scale (showing the importance of this concept [9]). The most popular are linear scale-space representations, pyramids, wavelets and multigrid methods. Simplest to understand of these is the pyramid of images (Figure 2) which takes an image and through using a low pass filter (often Gaussian) and subsampling creates a number of images that have fewer pixels and less detail. It is synonymous with viewing something at one distance and then moving back a distance and seeing less detail than you could close up. Both [25] and [7] use image pyramids, and they are popular in domains other than remote sensing. However they pose some problems; due to the coarse quantisation in the scale direction it is difficult to match objects between scales and they are not translationally invariant [9].

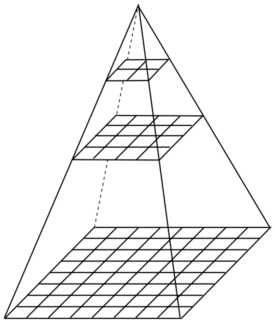


Fig. 2: Example showing how an 8x8 image is then reduced to a 4x4 and finally a 2x2. This illustrates the idea of a pyramid of images each with less detail than the previous using smoothing and subsampling Recreated from [9].

Scale-space representation is another method that produces a set of images which becomes coarser as the scale factor t increases (created through convolution with Gaussian kernels of increasing width) [9]. Compared to image pyramids they have some advantages: the size of the image is the same as the original with only the scale changing and they can be extended to be non-linear or spatio-temporal.

E. Discussion

With respect to Convolutional Neural Networks, there have been a number of lines of inquiry into how to include scale within them. As [25, page 1] states "Existing methods tend only consider low-level features and limited scales", which is a big problem for remote sensing due to the variety of sizes objects can take (such as manhole covers or trampolines). To combat this, [21] created the inception model which uses several different sizes of filter combined together, forming an 'inception unit'. This feature was inspired by [20], to handle multiple scales by making it scale-invariant, but differs in that these filters are learned through the CNN. By judiciously reducing dimensions and only implementing inception models at higher layers, they addressed the issue that with more filters, there is a higher chance of computational 'blow up'.

Similarly [12] uses the GoogLeNet architecture to evaluate its fitness for purpose in the field of image retrieval. However, where image classification often uses the final layers of a CNN, it was found that intermediate layers work best for image retrieval. They investigate the effect of scale on this problem domain by creating 'fake' high resolution images and showing that although the network isn't trained for them the accuracy for these images is greater than the original scale. They use this as argument to say that their extension to GoogLeNet, extracting Vector Locally Aggregated Descriptors (VLADs; using k-means clustering to create centres, where the distance to the centre is recorded and creates the VLAD) at each layer of the CNN, is resilient to scale. Their results are compared to other image retrieval benchmarks, in which, [15] has better accuracy, but [12] performed significantly better than the traditional methods of SIFT [10] with a BOVW.

Conversely [25] take a more traditional approach to scale, not incorporating it into the CNN but using a Laplacian spatial pyramid to subsample and input into the CNN. This is evaluated on hyperspectral data (103-band images), dimensionally reduced using Principal Componenets Analysis (PCA). Comparing this to a benchmark dataset [3], it performed well for certain classes, but other methods, such as clustering with a KNN together with a Support Vector Machine, had better overall accuracy. It is difficult to compare these results to other benchmark datasets because it uses very different parameters for its CNN (sigmoid activation rather than RELU, a voting scheme to combine them, only one filter size).

The aerial imagery literature has even more ways to deal with scale. Many analyses within remote sensing use object-based image analysis (OBIA), segmenting the image to create 'superpixels'. Much of the literature combining scale and aerial imagery pertains to extraction or detection of features from imagery and this is reflected in this section. The work in [4] combines OBIA and scale using a fractal net evolution approach (FNEA) to segment the image into multiple scales. This is particularly key for their purpose, extracting roads, where roads could be anything from small tracks to motorways, both of very different scales. They found that one scale parameter would not be optimal, either omitting the small or the large roads, and thus multiple parameters created several different results which were then combined. The FNEA was compared to a multiscale cognitive pyramid and showed good results, their respective completeness measures being 93% and 63% for the road map output. In fact, the problems facing [4] were less to do with the scale implementation than spectral similarities between landcovers. However the FNEA can offer a solution to scale issues in extraction and detection problems but would not provide a solution to large-scale image classification or retrieval tasks.

On the other hand, rather than trying to find different sizes of road, [7] exploits the fact that roads are often a certain width. They use scale-space theory to create coarser images, detecting roads in these scales and using these to prune the extraction results at finer scales. Using this technique is very limited compared to that of [4], as any differing road sizes such as dual carriageways or smaller tracks would be omitted. [26] uses a more traditional pyramid of images together with a convolutional autoencoder to classify landcover using the same techniques as Zhao's previous work ([25]) mentioned above.

F. The Gap/Proposal

It can be seen that most of the literature focuses on benchmark datasets and little can be found on domain-specific problems [24]. My project aims to focus on the domain of automated map generation from aerial imagery, by combining imagery with CNNs and scale. This combination has been attempted before by [25], but the scale component was not a part of the CNN and the aim was classification, which is where the project differs. My project aims to produce an output of "What area is most like this area?" (a mixture of classification and image retrieval), not pure classification. The scale incorporated into this method will not be a preprocessing step but will, like in [21], be put into the CNN using a family of filters of different sizes. This will be a novel application of CNNs to an important domain-specific project.

References

[1] Yoshua Bengio. Learning Deep Architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009. ISSN 1935-8237. doi: 10.1561/2200000006. URL http://dx.doi.org/10.1561/2200000006.

[2] Dennis C. Duro, Steven E. Franklin, and Monique G. Dub. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment*, 118:259–272, 2012. URL http://www.sciencedirect.com/science/article/pii/S0034425711004172.

[3] Paulo Gamba. Pavia Centre and University Hyperspectral Remote Sensing Scenes. URL http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes. Telecommunications and Remote Sensing Laboratory.

[4] Xin Huang and Liangpei Zhang. Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines. *International Journal of Remote Sensing*, 30(8):1977–1987, 2009. URL http://www.tandfonline.com/doi/abs/10.1080/01431160802546837.

[5] Andrej Karpathy and Li Fei-Fei. CS231n Convolutional Neural Networks for Visual Recognition. URL http://cs231n.github.io/convolutional-networks/.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-w.

[7] Ivan Laptev, Helmut Mayer, Tony Lindeberg, Wolfgang Eckstein, Carsten Steger, and Albert Baumgartner. Automatic extraction of roads from aerial images based on scale space and snakes. *Machine Vision and Applications*, 12(1):23–31, 2000. URL http://link.springer.com/article/10.1007/s001380050121.

[8] Yuanqing Lin, Fengjun Lv, Shenghuo Zhu, Ming Yang, Timothee Cour, Kai Yu, Liangliang Cao, and Thomas Huang. Large-scale image classification: fast feature extraction and svm training. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1689–1696. IEEE, 2011.

[9] Tony Lindeberg. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(1-2):225–270, 1994. URL http://www.tandfonline.com/doi/abs/10.1080/757582976.

[10] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[11] Jawad Nagi, Frederick Ducatelle, Gianni A Di Caro, Dan Cireşan, Ueli Meier, Alessandro Giusti, Farrukh Nagi, Jürgen Schmidhuber, and Luca Maria Gambardella. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *Signal and Image Processing Applications (ICSIPA), 2011 IEEE International Conference on*, pages 342–347. IEEE, 2011.

[12] Joe Ng, Fan Yang, and Larry Davis. Exploiting local features from deep networks for image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 53–61, 2015.

[13] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *Computer Vision–ECCV 2010*, pages 143–156. Springer, 2010.

[14] Yuguo Qian, Weiqi Zhou, Jingli Yan, Weifeng Li, and Lijian Han. Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. *Remote Sensing*, 7(1):153–168, 2014. URL http://www.mdpi.com/2072-4292/7/1/153/htm.

[15] Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. A baseline for visual instance retrieval with deep convolutional networks. *arXiv preprint arXiv:1412.6574*, 2014.

[16] David Rhind. Personality as a factor in the development of a discipline: The example of computer-assisted cartography. *The American Cartographer*, 15(3):277–289, 1988. URL http://www.tandfonline.com/doi/abs/10.1559/152304088783886928.

[17] John Rogan, Janet Franklin, Doug Stow, Jennifer Miller, Curtis Woodcock, and Dar Roberts. Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. *Remote Sensing of Environment*, 112 (5):2272–2283, May 2008. ISSN 0034-4257. doi: 10.1016/j.rse.2007.10.004. URL http://www.sciencedirect.com/science/article/pii/S003442570700449X.

[18] David E Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propogating errors. *Nature*, 323:533–536, 1986.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. URL http://link.springer.com/article/10.1007/s11263-015-0816-y.

[20] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(3):411–426, 2007.

[21] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015. doi: 10.1109/CVPR.2015.7298594.

[22] Waldo R. Tobler. Automation and cartography. *Geographical Review*, 49(4):526–534, 1959. URL http://www.jstor.org/stable/212211.

[23] Deborah Tranowski. A knowledge acquisition environment for scene analysis. *International journal of man-machine studies*, 29(2):197–213, 1988. URL http://www.sciencedirect.com/science/article/pii/S0020737388800464.

[24] Kiri Wagstaff. Machine learning that matters. *arXiv preprint arXiv:1206.4656*, 2012.

[25] Wenzhi Zhao and Shihong Du. Learning multiscale and deep representations for classifying remotely sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 113:155–165, 2016.

[26] Wenzhi Zhao, Zhou Guo, Jun Yue, Xiuyuan Zhang, and Liqun Luo. On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery. *International Journal of Remote Sensing*, 36(13):3368–3379, 2015.