Computer Networks 168 (2020) 107034

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

A modular CNN-based building detector for remote sensing images

Dimitrios Konstantinidis, Vasileios Argyriou*, Tania Stathaki, Nikolaos Grammalidis

Kingston University London, SEC, Kingston, London KT12EE, United Kingdom

ARTICLE INFO

Article history:
Received 14 April 2019
Revised 4 November 2019
Accepted 26 November 2019
Available online 28 November 2019

Keywords:
Remote sensing
Modular-CNN
Building detection

ABSTRACT

Convolutional neural networks (CNNs) have resurged lately due to their state-of-the-art performance in various disciplines, such as computer vision, audio and text processing. However, CNNs have not been widely employed for remote sensing applications. In this paper, we propose a CNN architecture, named Modular-CNN, to improve the performance of building detectors that employ Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP) in a remote sensing dataset. Additionally, we propose two improvements to increase the classification accuracy of Modular-CNN. The first improvement combines the power of raw and normalised features, while the second one concerns the Euler transformation of feature vectors. We demonstrate the effectiveness of our proposed Modular-CNN and the novel improvements in remote sensing and other datasets in a comparative study with other state-of-the-art methods.

1. Introduction

In the last few decades, the image sensors attached to satellites have evolved in a way that nowadays allows the capture of high-resolution multi-spectral satellite images, [1–4]. As a result, land cover classification became a widely-studied field providing also solutions on the detection and classification of buildings and other structures. A few important application areas, where the development of a system capable of monitoring and modelling urban changes can nd usefulness are [5,6] a) sociology, for the monitoring the dynamic processes that occur in a complex urban environment, b) citizen welfare, for city planning, c) city protection, for the analysis and assessment of the impact of fire, flood and natural disasters in an urban environment, d) illegal construction for detecting illegal building activity and e) navigation, for the development and constant update of accurate urban maps that can be employed for navigation purposes. Several remote-sensing applications, such as city planning, urban mapping and urban change detection can be improved using building detection systems that employ satellite images and reconstructed 3D representations. Additionally, urban expansion or decline can be studied and correlated to climatic changes and social, economic or natural factors in order to provide solutions and ensure human prosperity. Lately, 2D and 3D building detection from remote sensing images is tackled by means of machine learning and, more specifically, convolutional neural networks [7–9]. Convolutional neural networks were heavily employed in the 1990s [10] but were later abandoned, when the SVMs were introduced [11]. The interest in CNNs was rekindled when Krizhevsky et.al [12] showed the superior performance of CNNs on the ImageNet Large Scale Visual Recognition Challenge [13].

In this work, we implement a CNN architecture for building detection able to accommodate models when the amount of training data is low as in the case of remote sensing datasets (e.g. WorldView-2, Quickbird and Benedek in [14]). Furthermore, the proposed CNN architecture tends to perform better due to its modular structure and the ability to optimise easily. The optimisation simplicity of the suggested CNN allows us to analyse in depth the effect of this improvements on the accuracy of the overall detection. Our first contribution is the combined use of both normalised and raw features inside the CNN. Although normalisation makes a classifier more robust to intensity variations, the use of raw features can increase the discrimination ability of a CNN. Moreover, we propose the Euler transformation of the feature vectors before their classification based on the use of the cosine-based distance function that was proposed by Fitch as a metric for the separation between classes [15]. Overall the proposed method offers advantages due to its modular structure and the optimisation simplicity of the CNN architecture. Also supports both raw and processed data and it can be extended including deeper modules. Furthermore, the introduced layers provide improved robustness to noise and low-quality data. The main disadvantages are related to the complexity of the proposed system that it is higher in terms on computational time and required operations. Regarding the training stage it may be more time consuming, but it doesn't affect

* Corresponding author.
E-mail addresses: d.konstantinidis12@imperial.ac.uk (D. Konstantinidis), Vasileios.Argyriou@kingston.ac.uk (V. Argyriou), t.stathaki@imperial.ac.uk (T. Stathaki), ngramm@iti.gr (N. Grammalidis).

https://doi.org/10.1016/j.comnet.2019.107034
1389-1286/

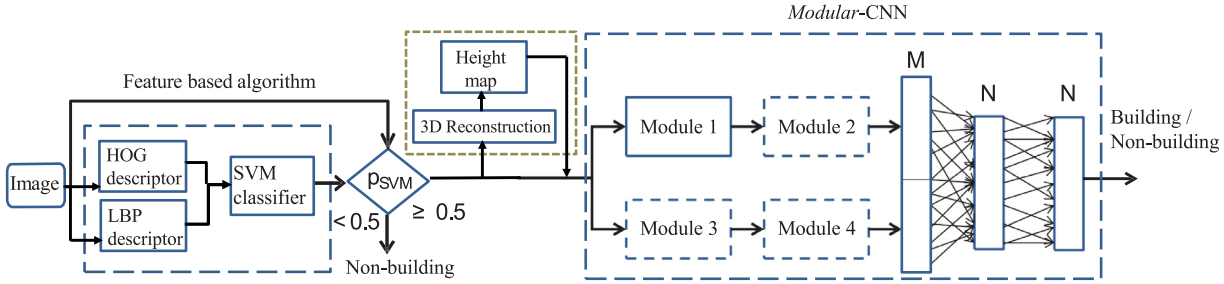

Fig. 1. Flowchart of the proposed building detection methodology.

further the performance during the deployment. We demonstrate using different datasets that a cosine-based distance function can make a classifier more robust to noise and outliers and increase the performance of a CNN.

2. Related work

Building detection is a significant, yet challenging task for remote sensing applications, since buildings present significant size, 3D shape, colour and texture variations. Several building detection methodologies have been proposed with varying degree of success. Energy functions based on building properties were constructed and employed in a level-set segmentation framework to achieve accurate building segmentation results [16]. Lines were utilised for building detection, since building shapes favour line detection [5]. Shadow detection has also been incorporated in several methodologies, as a way to denote the existence of tall structures, which can be candidate buildings [17]. Corner and texture features, whose distribution maxima can be considered as observations of building presence were also considered [18]. On the other hand, Ilsever et.al in [19] employed HOG [20] features for the identification of building regions. Konstantinidis et.al in [21] proposed an accurate building detector based on the features suggested in [22], along with a new distance function that can be employed to improve the robustness of an SVM classifier to noise. Last but not least, Markov Random Fields were employed for building segmentation in [23,24].

Regarding the work on CNNs, several modifications have been proposed to increase their classification performance. A detailed overview of recent improvements to CNNs can be found in [25–27]. Next, we present and focus on the improvements that are relevant to our work. Nair and Hinton introduced the Rectified Linear Unit (ReLU) as an alternative to the sigmoid and hyperbolic tangent activation functions [28]. It has been shown that ReLU outperforms other activation functions and allows a CNN to be trained faster and obtain easier sparse representations [12,29]. Dropout is a regularisation technique proposed by Hinton et.al in order to prevent overfitting during the training of deep neural networks [30]. Several modifications to the dropout method, such as maxout and adaptive dropout were later proposed [31,32]. To enhance model discriminability and avoid overfitting Lin et.al proposed the Network in Network (NIN), which concerns the use of multi-layer perceptrons inside the deep neural network [33]. Their work led Szegedy et.al to propose the Inception module [34], which uses variable filter sizes to capture patterns of different size. Finally, He et.al proposed residual learning to address the problem of degradation in deep neural networks, achieving state-of-the-art performance on several benchmark datasets [35].

Different from the CNN improvements discussed above, we propose the NL and ETL layers that consist alternative ways of increasing the accuracy and robustness of CNNs. The novel NL and ETL layers perform simple transformations of CNN feature representations without adding additional training parameters to the problem, although the next layers have their inputs doubled due to the use of the proposed NL and ETL layers. As a result, the proposed NL and ETL layers can be considered efficient due to the lack of training weights, especially if combined with an operation that reduces their output features.

Furthermore, in order to demonstrate the link between the accuracy and the appropriate method selection for satellite building detection the work presented at the survey papers and frameworks [36,37] demonstrates the difference performance expectations in relation to the selected methods and the corresponding datasets and applications.

3. Proposed *modular-CNN architecture*

Our method takes advantage of the accurate feature based building detectors such as HOG and LBP. In this work we extend and improve the building detection methodology by employing our *Modular*-CNN. A flowchart of the proposed methodology can be seen in Fig. 1. A tested image is split in overlapping windows and multiple scales and is fed to proposed building detection methodology. Initially feature based algorithms are employed as the first processing step in order to acquire an as accurate as possible initial set of image blocks that represent candidate buildings. The detected buildings at this stage are provided as input to the proposed *Modular*-CNN architecture and there is an option to apply 3D reconstruction methods [38,39] aiming to obtain an estimate of the buildings' height map. Our *Modular*-CNN is then employed to further improve and refine the building detection results by discarding false detections. In this way, we take advantage of both the power of the discriminative HOG and LBP features and the ability of a CNN to automatically generate descriptive features. Furthermore, the use of the *Modular*-CNN on the positive output of the feature based algorithm allows a speed up of the detection procedure as the *Modular*-CNN is not applied to the entire image and the introduction of new false alarms from the *Modular*-CNN detector is suppressed. The disadvantage of this approach is that buildings lost by the feature based classifier cannot be recovered at a later stage.

In this work, we implement a CNN that consists of maximum two units or *modules* placed in a sequential and/or parallel configuration, as shown in Fig. 1. We call this architecture *Modular*-CNN for sake of the *modules* that it consists of. A *module* is a basic CNN that has a combination of layers such as convolutional layers, activation functions and pooling layers (e.g. similar to VGG-S or VGG-16). The top of the *Modular*-CNN architecture consists of two fully connected linear layers, the first of which reduces the number of features and, consequently, parameters that need to be optimised, while the second one performs a linear mapping without modifying the feature vector dimensionality. The last layer performs the classification of the feature vectors to classes. Each *module* has its own set of hyper-parameters that needs to be optimised. Our strat-

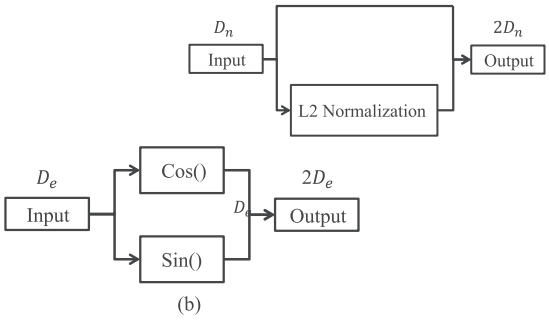

Fig. 2. Proposed Normalisation (top) and Euler transform (bottom) layers.

egy is to optimise the *Modular*-CNN as follows: *modules* are added to the CNN architecture one by one, their hyper-parameters are optimised independently of the hyper-parameters of other *modules* and then these hyper-parameters are kept fixed, while subsequent *modules* are introduced. The optimisation procedure lasts as long as the classification performance of the CNN increases or until the required depth or width is reached. Next, we present and analyse two novel improvements that come in the form of additional layers added to the *Modular*-CNN architecture. These new layers are the Normalisation and Euler transform layers, and as we demonstrate, they improve the performance and robustness of the tested CNNs.

3.1. Normalisation layer

Normalising the input data is a common data pre-processing method that increases the performance of a classifier, especially one that relies on stochastic gradient optimisation methods. This is needed due to the equal weighing of scaled features. Otherwise, too large input values can saturate some of the hidden neurons of a neural network, rendering the neurons of the next layers inactive and the neural network to get stuck in local optima. However, since the output of a CNN is a non-linear mapping of the normalised input, the effect of normalisation is in most cases lost in the feature space [40]. As a result, it is useful to normalise the computed feature vectors prior to their classification. On the other hand, since the raw features come from already normalised data, the neural network mapping may have led some important features to become prominent in the output and this discrimination can be lost after a new normalisation in the feature space. Based on these ideas and observations, we suggest the use of a shortcut connection before the feature vector classification. The proposed Normalisation Layer (NL), shown at the top of Fig. 2, takes as input a feature vector $\mathbf{x}$ and creates a new feature vector that has twice the size of the initial vector. The first half of the new feature vector is a copy of the initial vector (i.e. $\mathbf{x}$), while the second half is a normalised by l_2-norm copy of the initial feature vector (i.e. $\frac{\mathbf{x}}{\|\mathbf{x}\|_2}$).

3.2. Euler transform layer

Fitch et.al was the first to introduce a new distance function as a replacement to the l_2-norm in the computation of displacement between video frames [15]. The l_2-norm is known to be significantly affected by large values that can be attributed to noise. The new distance function was proposed to counter this sensitivity of the l_2-norm, as it is considered to be robust to noise and outliers. Given two feature vectors $\mathbf{x_i}$ and $\mathbf{x_j}$ that have values in the range [0,1] and are of length L, an ideal distance function can be approximated by a limited number P of sinusoidal terms, giving rise to the cosine-based dissimilarity measure

$$d(\mathbf{x_i}, \mathbf{x_j}) \approx \sum_{p=1}^{P} \sum_{l=1}^{L} b_p (1 - \cos(a_p \pi (\mathbf{x_i}(l) - \mathbf{x_j}(l)))) \quad (1)$$

In the special case, where only one sinusoidal term is considered (i.e. $P = 1$), the cosine-based distance function of Eq. (1) boils down to the measure

$$d(\mathbf{x_i}, \mathbf{x_j}) = \sum_{l=1}^{L} (1 - \cos(\alpha\pi (\mathbf{x_i}(l) - \mathbf{x_j}(l)))) \quad (2)$$

The cosine-based dissimilarity measure of Eq. (2) is controlled by a single variable α that affects the response of the dissimilarity measure to large differences. Small values of α make the cosine-based function to behave similarly to the l_2-norm, meaning that the distance between two feature vectors increases as their difference becomes larger. On the other hand, large values of α make the cosine-based dissimilarity measure to suppress its response to large differences. Since large differences between feature vectors can be attributed to outliers, the cosine-based distance function attempts by regularising its control variable α to suppress the effect of noise and outliers. The optimal value of the parameter α can be determined by an exhaustive search on a validation set. The cosine-based distance function has the ability to suppress noise because its derivative is equivalent to Andrew's M-Estimate [15,41], defined in Eq. (3), for difference values in the range $[-1, 1]$. The Andrew's M-Estimate is a redescending m-estimator, which is considered as an outlier rejection technique. This holds because the cosine-based distance function is not a monotonically increasing function as the difference between two vectors increases, like the l_2-norm, but it redescends smoothly towards zero for large difference values. This allows the cosine-based distance function to smoothly suppress large differences, which can be attributed to noise or outliers.

$$\psi(r) = \begin{cases} sin(\pi r) & \text{if } -1 \le r \le 1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The cosine-based distance function can either be directly employed as an alternative to the l_2-norm [15] or the feature vectors can be transformed to their Euler representation before they are mapped to classes. Applying the cosine-based dissimilarity measure to a pair of vectors $\mathbf{x_i}$ and $\mathbf{x_j}$ is equivalent to transforming the feature vectors to their Euler representation $\mathbf{z_i}$ and $\mathbf{z_j}$, where $\mathbf{z_i} = \frac{1}{\sqrt{2}} e^{i\alpha\pi\mathbf{x_i}}$ and subsequently employ the l_2-norm function. With the use of a few trigonometric identities and the substitution $\theta_\mathbf{i} = \alpha\pi\mathbf{x_i}$, the proof is presented in Eq. (4). In this work, we choose to employ the Euler transformation method since it is easily integrated in a neural network framework. Therefore, we propose the addition of a new layer, called Euler transform layer (ETL), in the *Modular*-CNN architecture. The ETL layer, which is introduced just before the layer that classifies feature vectors and after the NL layer, transforms features to their Euler representation, meaning that each feature vector $\mathbf{x}$ is now described by a cosine (i.e. $\cos(\alpha\pi\mathbf{x})$) and a sine (i.e. $\sin(\alpha\pi\mathbf{x})$) part that are concatenated together. The ETL layer is depicted at the bottom of Fig. 2.

$$\begin{aligned} \|\mathbf{z_i} - \mathbf{z_j}\|^2 &= \frac{1}{2} \sum_{l=1}^{L} \|(\cos(\theta_i(l)) + i\sin(\theta_i(l)) \\ &\quad - (\cos(\theta_j(l)) - i\sin(\theta_j(l))\|^2 \\ &= \frac{1}{2} \sum_{l=1}^{L} \left\| 2\sin\left(\frac{\theta_i(l) - \theta_j(l)}{2}\right) \left(ie^{i\frac{\theta_i(l)+\theta_j(l)}{2}}\right) \right\|^2 \\ &= \sum_{l=1}^{L} 1 - cos(\theta_i(l) - \theta_j(l)) \end{aligned} \quad (4)$$

4. Experiments

In this section, we present the implementation details and optimisation procedure of our *Modular*-CNN. Moreover, we analyse

and compare the effect of our proposed layers on the performance of the *Modular*-CNN and PlainNet, a deep CNN employed in [35], on three different datasets. Finally, we evaluate the overall building detection performance of our proposed method and compare it with other state-of-the-art methods.

4.1. Implementation details

The construction of an optimal *Modular*-CNN follows a *module*-based optimisation scheme. This means that each *module* is initially introduced to the current CNN and its hyper-parameters are optimised using stochastic gradient descent. In the case that the addition of the *module* is beneficial to the performance of the CNN, the *module* is added permanently in the CNN and its hyper-parameters are kept fixed, while new *modules* are introduced. Otherwise, the optimisation procedure terminates and the current CNN without the latest *module* is returned. The optimisation of the *Modular*-CNN is performed without the proposed novel NL and ETL layers. These layers are introduced afterwards and the CNN is retrained with its hyper-parameters kept fixed to the optimal values.

Each *module* is optimised with respect to the hyper-parameters of the basic layers that it consists of. These hyper-parameters are the convolution type, the convolutional filter size, the number of convolutional filters, the activation function, the pooling type and the pooling size and stride. Only the activation function is fixed to the ReLU unit, while the other hyper-parameters are optimised as described below. Two types of convolution are examined, the local convolution and the full convolution, proposed in [42]. The full convolution is employed for dense predictions since it has the ability to output features of various sizes, not necessarily smaller than the input size. In our implementation of the full convolution, we choose the output size to be equal to the input size, by adding appropriate zero padding. The number of convolutional filters is selected by the value pool {25, 50, 75, 100}. We also experiment with symmetrical ($m \times m$) and asymmetrical ($1 \times m$, $m \times 1$) convolutional filters, where $m = \{3, 5, 7\}$, in an attempt to extract useful features from the images. Furthermore, we test both max and average pooling using either a kernel of size 2×2 with a stride of 2 (i.e. non-overlapping pooling) or a kernel of size 3×3 with a stride of 1 (i.e. overlapping pooling). It has been shown that overlapping pooling can prevent overfitting [12].

Other hyper-parameters that affect the *Modular*-CNN architecture and training are optimised with respect to the remote sensing dataset and kept fixed for the other datasets. The dimensionality of the feature vectors introduced to the second fully connected layer of the *Modular*-CNN is optimised to the value of 2000. Linear and Euclidean layers for the classification of feature vectors are tested and we conclude that the Euclidean layer that performs clustering of the feature vectors by employing the l_2-norm distance function outperforms the linear layer. Dropout is examined but leads to sub-optimal results. Finally, the size of the mini-batch and the learning rate are optimised to the value of 32 and 0.05 respectively.

The CNN training with each hyper-parameter configuration is performed for a maximum number of 100 iterations. In each iteration, the training set is introduced to the CNN in mini-batches and the CNN is then evaluated on a validation set. During the evaluation, the loss on the validation set, which is equal to the average negative log-likelihood of each sample to belong to the correct class, is computed. The training phase is terminated when the loss on the validation set does not decrease for 5 consecutive iterations. This strategy is employed to prevent overfitting of the CNN to the training data, since after a few epochs the loss on the training set keeps decreasing, while the loss on the test set starts increasing. This typical behaviour of a CNN is reported during our experiments with the *Modular*-CNN. The CNN training is repeated for 3 rounds, where a new initialisation/reset of the weights is performed in the beginning of each new round. The average performance of the CNN on the validation set is computed and used for the selection of the optimal hyper-parameter configuration. The *Modular*-CNN and the novel layers were developed using the Torch software and a NVIDIA Tesla K40 GPU was used for a boost in the computational speed.

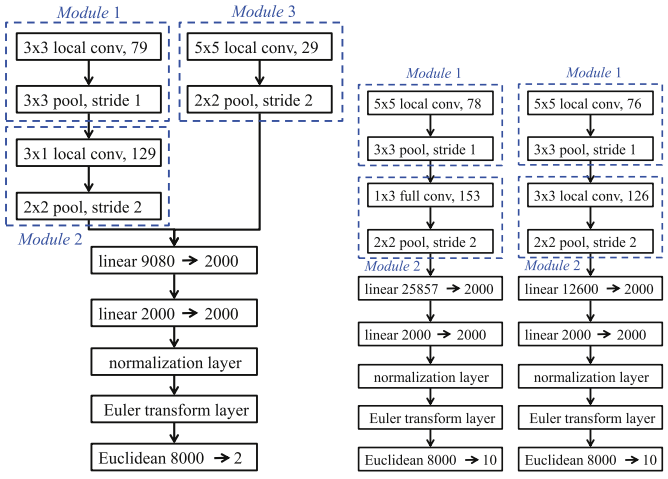

Fig. 3. Optimal *Modular*-CNN architectures for the (a) QuickBird/WorldView-2, (b) CIFAR-10 and (c) MNIST datasets.

4.2. Evaluation of proposed NL and ETL layers

In order to evaluate the performance of the proposed NL and ETL layers, we consider two different CNN architectures (i.e. the proposed *Modular*-CNN and PlainNet), and we perform experiments by deploying the proposed layers in both of them. PlainNet is a version of the plain-net for $n = 1$ defined in [35]. More specifically, PlainNet consists of stacks of two (3×3) convolution layers for each feature map size. The feature map size is progressively halved from D to $D/4$, while the number of filters is doubled from 16 to 64. For the evaluation we used remote sensing images from QuickBird/WorldView-2, the CIFAR-10 and MNIST datasets. Experiments are performed to optimise the *modules* that the CNN consists of. The optimal performance on the validation set is achieved by the *Modular*-CNN architecture depicted in Fig. 3(a). The average loss on the validation set of our *Modular*-CNN with and without the addition of the NL and ETL layers for values of α in the range $[0 - 1.9]$ is depicted in Fig. 4(a). Table 1 summarises the performance of the *Modular*-CNN and PlainNet with and without the addition of the proposed layers on the three tested datasets.

From Table 1 and Fig. 4(a), one can conclude that the addition of the proposed layers to the *Modular*-CNN and PlainNet leads to a decrease in the error on the remote sensing validation set. Furthermore, the addition of the ETL layer reduces significantly the average loss on the validation set, thus increasing the generalisation power of the *Modular*-CNN. The optimal value of the parameter α for the ETL layer is included next to the corresponding error on Table 1. A comparison between PlainNet and *Modular*-CNN reveals that our proposed novel *Modular*-CNN slightly outperforms PlainNet with or without the addition of the novel layers, however at the expense of increased CNN parameters. To further validate the advantages of the proposed NL and ETL layers, experiments are performed on CIFAR-10 [45] and MNIST [46] benchmark datasets. The images of CIFAR-10 and MNIST datasets are small in size (i.e. 32×32 and 28×28 pixels respectively) and thus appropriate for the testing of the performance of the proposed CNN

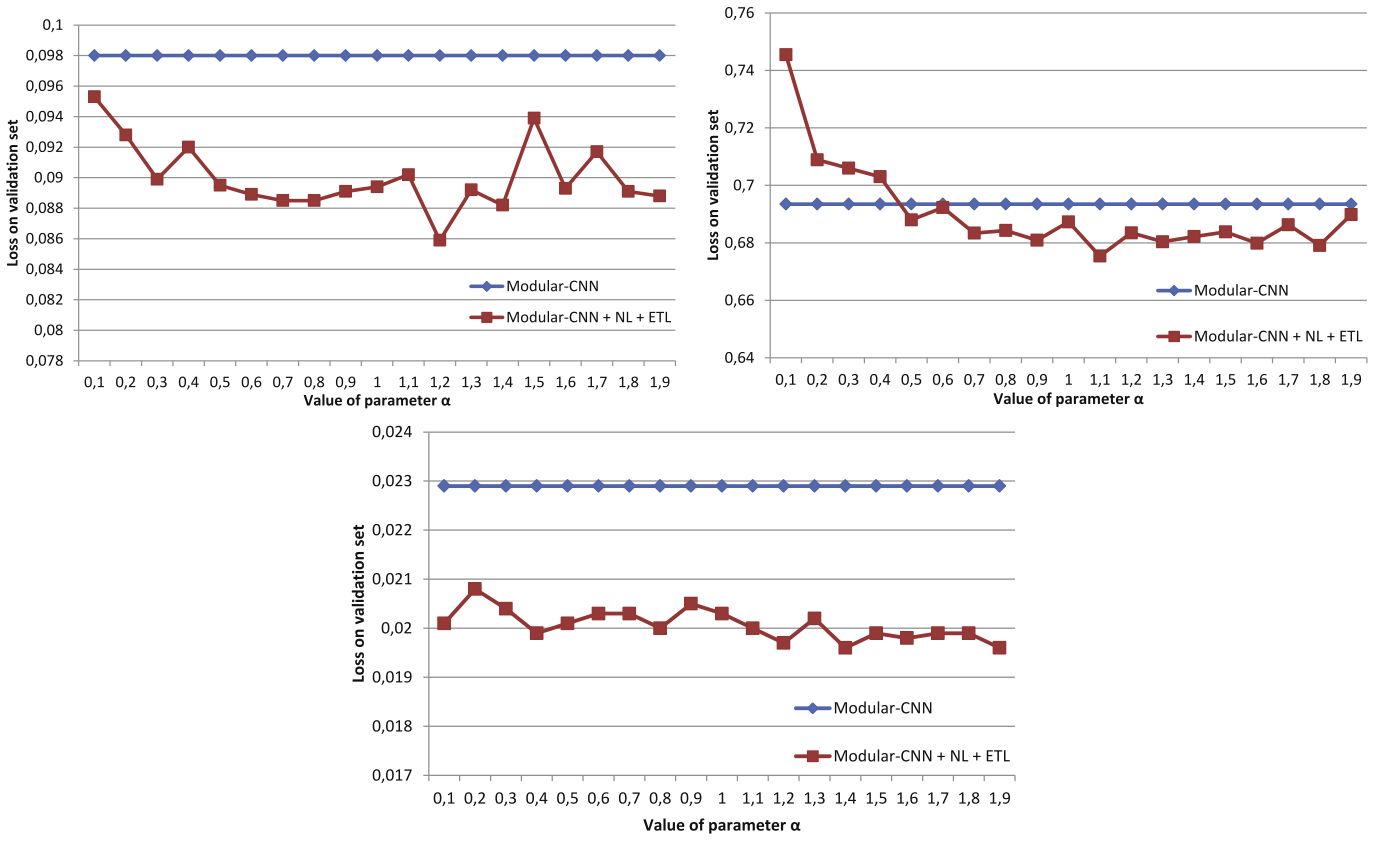

Fig. 4. Performance of *Modular*-CNN w/o NL and ETL layers with respect to average loss on validation set for (a) QuickBird/WorldView-2, (b) CIFAR-10 and (c) MNIST datasets.

Table 1
Classification performance of CNN architectures on the 3 tested datasets.

Method	Error (%)			No. of params
	QuickBird/WorldView-2	CIFAR-10	MNIST	
NIN [33]	–	10.41	0.47	0.97M
DSN [43]	–	9.69	0.39	0.97M
RCNN-96 [44]	–	9.31	0.31	0.67M
PlainNet [35]	3.404	25.94	0.71	2.69M
PlainNet+NL	3.118	25.35	0.69	2.69M
PlainNet+NL+ETL	3.049 (α=1.3)	25.35 (α=1.3)	0.63 (α=1.4)	2.70M
Modular-CNN	3.123	21.75	0.54	22.20–25.43M
Modular-CNN+NL	2.871	21.29	0.53	22.21–25.44M
Modular-CNN+ NL+ETL	2.866 (α=1.2)	21.29 (α=1.1)	0.51 (α=1.4)	22.22–25.45M

architecture. The CIFAR-10 training set consists of 50000 labelled images equally distributed between 10 different classes, while the test set consists of 10000 images. The validation set is formed by randomly selecting 10000 images out of the CIFAR-10 training set. The MNIST training set consists of 60000 grayscale images depicting digits $0-9$, while the test set consists of 10000. The validation set is formed by randomly selecting 10000 images out of the MNIST training set. The optimisation procedure leads to the CNN architectures described in Fig. 3, where only two sequential *modules* are employed. From Table 1 and Fig. 4 one can conclude that the addition of the NL layer improves the accuracy of both our *Modular*-CNN and PlainNet with respect to the cases without the NL layer. Moreover, the introduction of the ETL layer leads to a smaller loss on the validation set and thus, better generalisation ability of the *Modular*-CNN. Finally, our *Modular*-CNN with the novel layers outperforms by about 5.4% PlainNet with the novel layers on CIFAR-10. Although PlainNet demonstrates a larger depth, capable of learning complex features, the proposed *Modular*-CNN with a higher number of parameters can more effectively describe the dataset. Also for the MNIST dataset, our *Modular*-CNN outperforms PlainNet and achieves comparable performance with other state-of-the-art methods.

The proposed NL and ETL layers are not limited to the specific deep learning architectures or classification problem presented in this paper and as a result, the improved performance they achieve can be utilised in deeper deep learning architectures and in other image or video classification tasks, potentially leading to breakthroughs as far as accuracy and robust of deep networks is concerned. Finally, the notion of Euler Transform can be employed inside convolutional layers in order to provide more enriched and robust feature representations that can lead to better classification performance on several image and video classification tasks.

4.3. Experimentation on QuickBird/WorldView-2 Dataset

We employ QuickBird and WorldView-2 remote sensing images for the comparative evaluation of the proposed *Modular*-CNN in the task of building detection from satellite images. The training set consists of 900 positive and 1400 negative manually segmented and annotated images of size 20 $\times$ 20 pixels, depicting buildings and other structures (i.e. roads, trees etc) respectively. Since the training set is too small for accurate training of a CNN, it is augmented by taking horizontal and vertical flips of the images. The validation set consists of 20000 images randomly cropped from 5 labelled satellite images, while the test set consists of 24 labelled satellite images. The images consist of 4 spectral channels, namely red, green, blue, and near-infrared plus the height maps. Examples of the obtained height maps are shown in Fig 5. In our experiments, we employ the YUV colour space and the near-infrared channel, since this spectral configuration leads to the best performance on the validation set. Generally, we found that the CNN performance is strongly affected by the selected colour channels, which makes the selection of an optimal colour representation for the images critical to the performance of a CNN. The dataset is normalised to have zero mean and unity variance before it is fed to the CNNs for training and testing.

By employing the *Modular*-CNN with the proposed layers for the detection of buildings in the satellite images of the test set, we notice that our *Modular*-CNN can work complementary with a feature based algorithm as the locations of the false alarms differ between the two algorithms. Due to the small training set, applying the *Modular*-CNN directly to the test images leads to suboptimal results, as it is reported on Table 2. So, we propose a build-

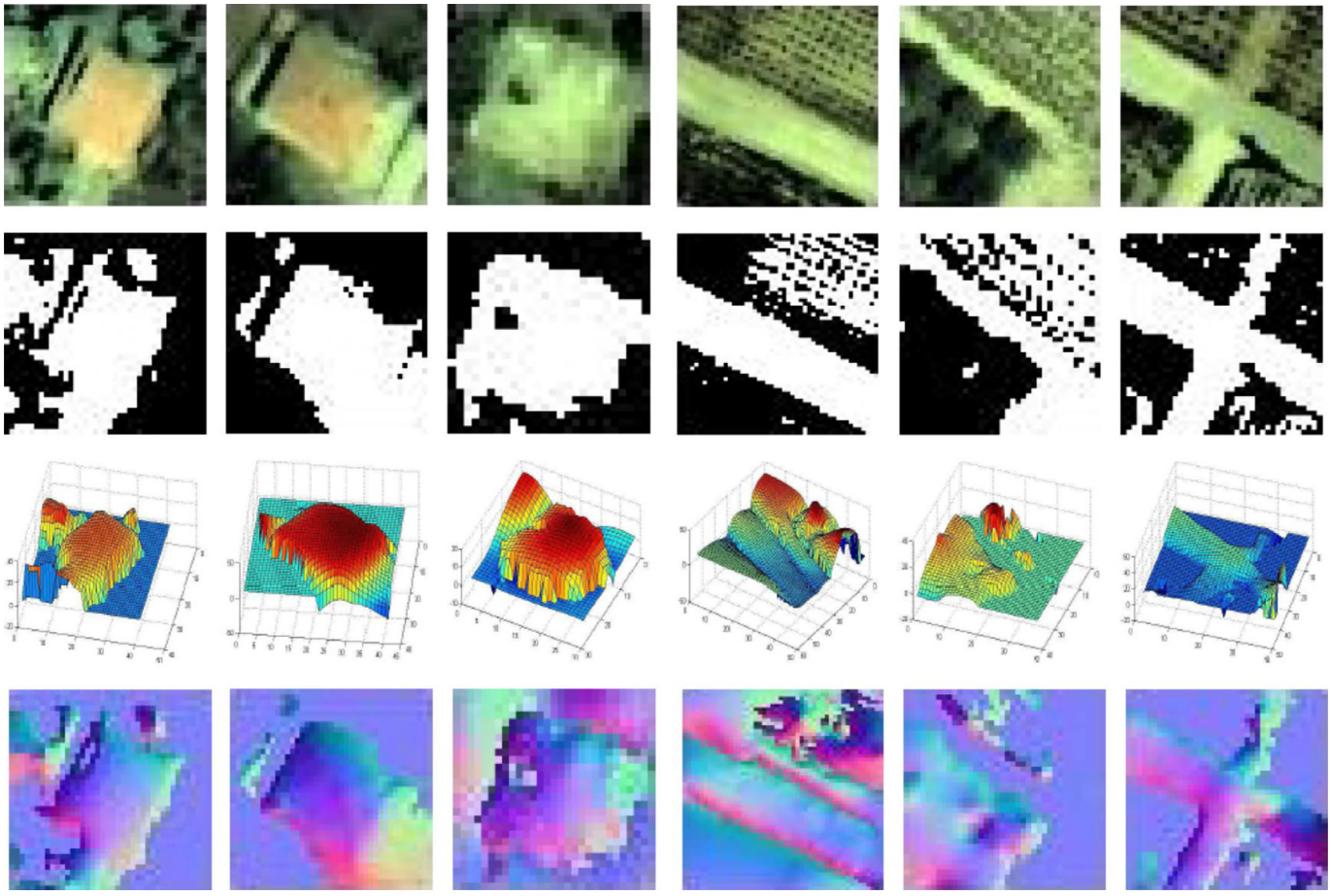

Fig. 5. Examples of the 3D reconstructed buildings and the corresponding height and normal maps.

Table 2
Performance (average and standard deviation) of building detectors on QuickBird/WorldView-2 test set. DAB: Discrete Adaboost.

Method	Recall	Precision	F_1-score	Av. eval. time
DAB with HAAR	0.767 $\pm$ 0.07	0.641 $\pm$ 0.105	0.691 $\pm$ 0.067	22.4 secs
LogitBoost with LBP	0.901 $\pm$ 0.041	0.706 $\pm$ 0.105	0.786 $\pm$ 0.068	24.1 secs
Fisherfaces	**0.998** $\pm$ 0.003	0.466 $\pm$ 0.136	0.624 $\pm$ 0.123	18.5 secs
Sirmacek [18]	0.552 $\pm$ 0.046	0.489 $\pm$ 0.133	0.509 $\pm$ 0.082	39.1 secs
Ilsever [19]	0.962 $\pm$ 0.008	0.209 $\pm$ 0.098	0.323 $\pm$ 0.121	51.2 secs
Konstantinidis [21]	0.953 $\pm$ 0.07	0.814 $\pm$ 0.106	0.871 $\pm$ 0.058	55.9 secs
Modular-CNN	0.968 $\pm$ 0.019	0.596 $\pm$ 0.1	0.733 $\pm$ 0.081	59.7 secs
Proposed method	0.937 $\pm$ 0.082	**0.859** $\pm$ 0.083	**0.891** $\pm$ 0.055	62.4 secs

ing detection method that combines the abilities of both feature based and *Modular*-CNN classifiers. In the new approach, the feature based classifier is initially applied to a test image and positive detections are extracted. Afterwards, the *Modular*-CNN is applied only in the positive detections (i.e. image regions), resulting in a set of final positive detections. This approach speeds up the detection procedure as the CNN is not applied to the entire image and avoids the introduction of new false alarms from the *Modular*-CNN detector. Table 2 compares the object-based performance of our proposed building detector with other state-of-the-art methods. The conclusion that can be drawn is that our proposed building detector discards several false alarms that the feature based algorithm produces, thus achieving an increase in the metric of F_1-score by 2.3%. We demonstrate the ability of our proposed building detector to suppress false alarms in the test set in Fig 6.

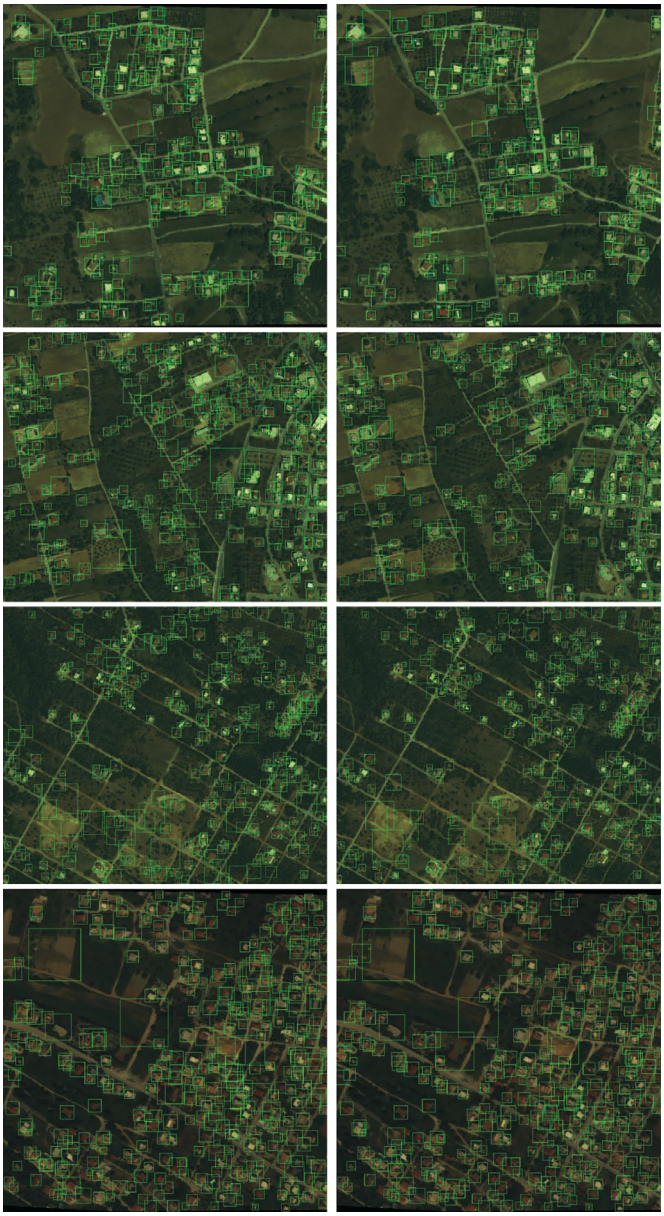

Fig. 6. Detections shown as green bounding boxes in QuickBird/WorldView-2 from feature based (first column) and our proposed (second column) building detectors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Conclusions and future work

In this paper, we propose a novel CNN architecture, called *Modular*-CNN that can be combined with a feature based classifier to improve the building detection performance on 2D and 3D remote sensing data. Furthermore, we propose two novel layers that can be added to CNN architectures in order to increase their discrimination ability and robustness. We analyse the effect of the novel layers on both our *Modular*-CNN and other deep CNN architecture, named PlainNet and demonstrate their beneficial effect in a comparative study with other state-of-the-art methods for building detection on remote sensing images.

As a future work, the proposed novel NL and ETL layers can be adopted by other deeper deep networks and applied to other image or video classification tasks boosting accuracy and robustness of deep networks. Additionally, of great research interest is a study on the use of Euler Transform inside convolutional layers providing more enriched feature representations, as well as a study on the performance of deep networks when multiple ETL layers are employed.

Declaration of Competing Interest

None.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.comnet.2019.107034.

References

[1] A. Sobotkova, S. Ross, High-resolution, multi-spectral satellite imagery and extensive archaeological prospection: case studies from Apulia, Italy, and Kazanlak, Bulgaria, 2010, pp. 25–28.
[2] Y.O. Ouma, R. Tateishi, J. t. Sri-Sumantyo, Urban features recognition and extraction from very-high resolution multi-spectral satellite imagery: a micro–macro texture determination and integration framework, IET Image Process. 4 (4) (2010) 235–254.
[3] A.M. Samad, N.S. Iliyas, N. Sahriman, F.A. Ruslan, M.Z. Zainal, N. Ghazali, N.A.M. Zaki, K. Zainuddin, Mangrove area detection by using high resolution satellite imagery, in: 2017 IEEE 13th International Colloquium on Signal Processing its Applications (CSPA), 2017, pp. 293–298.
[4] S.V. Koneru, M. Arul Raj, M. Padmaja, P.K. Kollu, L. Bokinala, A. Ravi Raja, A. Jitendra, Detection and enumeration of trees using cartosat2 high resolution satellite imagery, in: 2018 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES), 2018, pp. 1–6.
[5] Q.N. D. Woo and Q. Nguyen and, D. Park, Y. Jung, Building detection and reconstruction from aerial images, in ISPRS Congress (2008).
[6] A. Manno-Kovacs, T. Sziranyi, Orientation based building outline extraction in aerial images, in ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3, XXII ISPRS Congress (2012) 141–146.
[7] M. Volpi, D. Tuia, Dense semantic labeling of subdecimeter resolution images with convolutional neural networks, IEEE Trans. Geosci. Remote Sens. 55 (2) (2017) 881–893.
[8] K. He, J. Sun, Convolutional neural networks at constrained time cost, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[9] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, N.E. O'Connor, Shallow and deep convolutional networks for saliency prediction, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[10] G. Hinton, P. Dayan, B. Frey, R. Neal, The "wake-sleep" algorithm for unsupervised neural networks, Science 268 (5214) (1995) 1158–1161.
[11] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory (COLT), 1992, pp. 144–152.
[12] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems 25, 2012, pp. 1106–1114.
[13] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), 2009, pp. 248–255.
[14] C. Benedek, X. Descombes, J. Zerubia, Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics, IEEE Trans. Pattern Anal. Mach.Intell. 34 (1) (2012) 33–50.
[15] A. Fitch, A. Kadyrov, W. Christmas, J. Kittler, Fast robust correlation, IEEE Trans. Image Process. 14 (8) (2005) 1063–1073.

[16] J. Wegner, J. Montoya, K. Schindler, Road networks as collections of minimum cost paths, ISPRS J. Photogram. Remote Sens. 108 (2015) 128–137.
[17] J. Wegner, S. Branson, D. Hall, K. Schindler, P. Perona, Cataloging public objects using aerial and street-level images – urban trees, 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016).
[18] B. Sirmacek, C. Unsalan, A probabilistic framework to detect buildings in aerial and satellite images, IEEE Trans. Geosci. Remote Sens. 49 (1) (2011) 211–221.
[19] M. Ilsever, C. Unsalan, Building detection using HOG descriptors, in: 6th International Conference on Recent Advances in Space Technologies (RAST), 2013, pp. 115–119.
[20] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1, 2005, pp. 886–893.
[21] D. Konstantinidis, T. Stathaki, V. Argyriou, N. Grammalidis, Building detection using enhanced hog–lbp features and region refinement processes, IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. PP (99) (2016) 1–18.
[22] T. Ojala, M. Pietikainen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recognit. 29 (1) (1996) 51–59.
[23] M. Vakalopoulou, K. Karantzalos, N. Komodakis, N. Paragios, Building detection in very high resolution multispectral data with deep learning features, in: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2015, pp. 1873–1876.
[24] I. Hedhli, G. Moser, J. Zerubia, S.B. Serpico, A new cascade model for the hierarchical joint classification of multitemporal and multiresolution remote sensing data, IEEE Trans. Geosci. Remote Sens. 54 (11) (2016) 6333–6348.
[25] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, Recent advances in convolutional neural networks (2015). arXiv: 1512.07108.
[26] N. Audebert, A. Boulch, H. Randrianarivo, B. Le Saux, M. Ferecatu, S. Lefèvre, R. Marlet, Deep learning for urban remote sensing, in: 2017 Joint Urban Remote Sensing Event (JURSE), Dubai, 2017, pp. 1–4.
[27] A. Ben Hamida, A. Benoit, P. Lambert, L. Klein, C. Ben Amar, N. Audebert, S. Lefèvre, Deep learning for semantic segmentation of remote sensing images with rich spectral content, in: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, 2017, pp. 2569–2572.
[28] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 807–814.
[29] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS-11), 15, 2011, pp. 315–323.
[30] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors (2012). arXiv: 1207.0580.
[31] L.J. Ba, B. Frey, Adaptive dropout for training deep neural networks, in: Advances in Neural Information Processing Systems 26 (NIPS), 2013, pp. 3084–3092.
[32] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler, Efficient object localization using convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 648–656.
[33] M. Lin, Q. Chen, S. Yan, Network in network, in: International Conference on Learning Representations, 2013.
[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S.E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.
[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[36] G. Cheng, J. Han, A survey on object detection in optical remote sensing images (2016). arXiv: 1603.06201.
[37] M. Aamir, Y.-F. Pu, Z. Rahman, M. Tahir, H. Naeem, Q. Dai, A framework for automatic building detection from low-contrast satellite images, Symmetry 11 (2019) 3.
[38] J.T. Barron, J. Malik, Shape, illumination, and reflectance from shading, TPAMI (2015).
[39] K. Wang, J.-M. Frahm, Single image parametric building model estimation from satellite imagery, Int. Conf. 3D Vision (3DV) (2017).
[40] A.B. Graf, A.J. Smola, S. Borer, Classification in a normalized feature space using support vector machines, IEEE Trans. Neural Netw. 14 (3) (2003) 597–605.
[41] P. Huber, Robust Statistics, Wiley Series in Probability and Statistics, New York: Wiley, 1981.
[42] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Conf. Comput. Vision Pattern Recognit. (CVPR) (2015) 3431–3440.
[43] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2015.
[44] M. Liang, X. Hu, Recurrent convolutional neural network for object recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3367–3375.
[45] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images, Technical Report, University of Toronto, 2009.
[46] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

Dimitrios Konstantinidis Department of Electrical Engineering, Imperial College London, London, U.K. Dimitrios Konstantinidis received the Bachelor's degree in electrical engineering from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2009. In 2012, he received the Advanced Master's degree in artificial intelligence from the Katholieke Universiteit Leuven, Leuven, Belgium. He is currently working toward the Ph.D. degree in the Department of Electrical Engineering, Imperial College, London, U.K. on the topic of monitoring of urban changes from satellite and aerial images. He is simultaneously working as a Research Assistant in the Information and Technology Institute, Centre for Research and Technology Hellas, Thermi, Greece, on Greek and European research projects. His research interests include the field of image processing, computer vision, and artificial intelligence.

Vasileios Argyriou Department of Computer Sciences and Mathematics, Kingston University, Surrey, U.K. Vasileios Argyriou (M'15) received the B.Sc. degree in computer science from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2001, and the M.Sc. and Ph.D. degrees from the University of Surrey, Surrey, U.K., in 2003 and 2006, respectively, both in electrical engineering working on registration. In 2007, he joined the Communications and Signal Processing Department, Imperial College, London, U.K., where he was a Research Fellow working on 3-D object reconstruction. He is currently an Associate Professor at Kingston University, London, U.K., working on computer vision and AI for computer games, entertainment, security, and medical applications. His research interests include educational games and human computer interaction for developing systems.

Tania Stathaki Department of Electrical Engineering, Imperial College London, London, U.K. Tania Stathaki (M'08) was born in Athens, Greece. She received the Master's degree in electronics and computer engineering from the Department of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece, and the Ph.D. degree in signal processing from the Imperial College, London, U.K. She was a Lecturer with the Department of Information Systems and Computing, Brunel University, London, U.K., and an Assistant Professor with the Department of Technology Education and Digital Systems, University of Piraeus, Piraeus, Greece. She is currently a Reader in the Department of Electrical and Electronic Engineering, Imperial College. She has intensive research experience in image processing and computer vision and, more specifically, image fusion, image registration, change detection, object detection and recognition, and object tracking. She has mainly been involved in defence and security applications and has collaborated with various U.K. companies such as Dstl, General Dynamics, Selex Galileo, and BAE Systems. She is actively involved in various European Union and U.K. research projects. She has authored or co-authored many papers on signal and image processing and computer vision.

Nikolaos Grammalidis Information and Technology Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece Nikos Grammalidis (M'15) received the B.S. and Ph.D. degrees in electrical and computer engineering from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1992 and 2000, respectively. He is a Senior Researcher (Researcher Grade B) in the Information Technologies Institute, Centre of Research and Technology Hellas, Thermi, Greece. Prior to his current position, he was a Researcher in 3-D Imaging Laboratory, Aristotle University of Thessaloniki. His main research interests include computer vision, signal, image and video processing, and stereoscopy and multiview image sequence analysis and coding. His involvement with these research areas has led to the co-authoring of more than 25 articles in refereed journals and more than 75 papers in international conferences. Since 1992, he has been actively involved in more than 25 European commission and National projects. He served as a regular reviewer for a number of international journals and conferences.