Medical Image Analysis 000 (2016) 1–14

Contents lists available at ScienceDirect

Medical Image Analysis

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

Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks

Jelmer M. Wolterink[a,*], Tim Leiner[b], Bob D. de Vos[a], Robbert W. van Hamersvelt[b], Max A. Viergever[a], Ivana Išgum[a]

[a] Image Sciences Institute, University Medical Center Utrecht, Q.02.4.45, P.O. Box 85500, 3508 GA Utrecht, The Netherlands
[b] Department of Radiology, University Medical Center Utrecht, E.01.132, P.O. Box 85500, 3508 GA Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 10 January 2016
Revised 7 April 2016
Accepted 19 April 2016
Available online xxx

Keywords:
Coronary artery calcifications
Automatic calcium scoring
Convolutional neural network
Cardiac CT angiography

ABSTRACT

The amount of coronary artery calcification (CAC) is a strong and independent predictor of cardiovascular events. CAC is clinically quantified in cardiac calcium scoring CT (CSCT), but it has been shown that cardiac CT angiography (CCTA) may also be used for this purpose. We present a method for automatic CAC quantification in CCTA. This method uses supervised learning to directly identify and quantify CAC without a need for coronary artery extraction commonly used in existing methods.

The study included cardiac CT exams of 250 patients for whom both a CCTA and a CSCT scan were available. To restrict the volume-of-interest for analysis, a bounding box around the heart is automatically determined. The bounding box detection algorithm employs a combination of three ConvNets, where each detects the heart in a different orthogonal plane (axial, sagittal, coronal). These ConvNets were trained using 50 cardiac CT exams. In the remaining 200 exams, a reference standard for CAC was defined in CSCT and CCTA. Out of these, 100 CCTA scans were used for training, and the remaining 100 for evaluation of a voxel classification method for CAC identification. The method uses ConvPairs, pairs of convolutional neural networks (ConvNets). The first ConvNet in a pair identifies voxels likely to be CAC, thereby discarding the majority of non-CAC-like voxels such as lung and fatty tissue. The identified CAC-like voxels are further classified by the second ConvNet in the pair, which distinguishes between CAC and CAC-like negatives. Given the different task of each ConvNet, they share their architecture, but not their weights. Input patches are either 2.5D or 3D. The ConvNets are purely convolutional, i.e. no pooling layers are present and fully connected layers are implemented as convolutions, thereby allowing efficient voxel classification.

The performance of individual 2.5D and 3D ConvPairs with input sizes of 15 and 25 voxels, as well as the performance of ensembles of these ConvPairs, were evaluated by a comparison with reference annotations in CCTA and CSCT. In all cases, ensembles of ConvPairs outperformed their individual members. The best performing individual ConvPair detected 72% of lesions in the test set, with on average 0.85 false positive (FP) errors per scan. The best performing ensemble combined all ConvPairs and obtained a sensitivity of 71% at 0.48 FP errors per scan. For this ensemble, agreement with the reference mass score in CSCT was excellent (ICC 0.944 [0.918–0.962]). Aditionally, based on the Agatston score in CCTA, this ensemble assigned 83% of patients to the same cardiovascular risk category as reference CSCT.

In conclusion, CAC can be accurately automatically identified and quantified in CCTA using the proposed pattern recognition method. This might obviate the need to acquire a dedicated CSCT scan for CAC

* Corresponding author. Tel.: +31887569695.
E-mail addresses: j.m.wolterink@umcutrecht.nl (J.M. Wolterink), t.leiner@umcutrecht.nl (T. Leiner), b.d.devos-2@umcutrecht.nl (B.D. de Vos), r.w.vanhamersvelt-3@umcutrecht.nl (R.W. van Hamersvelt), m.viergever@umcutrecht.nl (M.A. Viergever), i.isgum@umcutrecht.nl (I. Išgum).

http://dx.doi.org/10.1016/j.media.2016.04.004

scoring, which is regularly acquired prior to a CCTA, and thus reduce the CT radiation dose received by patients.

1. Introduction

Cardiovascular disease (CVD) is the global leading cause of death. The amount of coronary artery calcification (CAC) as quantified in cardiac CT – the calcium score – is a strong and independent predictor of CVD events (Yeboah et al., 2012).

In a clinical cardiac CT exam, a calcium scoring CT (CSCT) scan and a coronary CT angiography (CCTA) scan are typically both acquired. The CCTA scan is used for stenosis detection or identification of non-calcified plaque, and the CSCT scan is used to determine the calcium score (Hecht, 2015). However, it has been shown that CAC may also be quantified in CCTA. In a study by Pavitt et al. (2014), 85% of patients with a high calcium score in CSCT also had a high calcium score in CCTA (specificity 99%). Moreover, Mylonas et al. (2014) showed excellent agreement between CVD risk categories based on calcium scoring in CCTA and categories based on calcium scoring in CSCT (Cohen's linearly weighted $\kappa = 0.93$). A recent survey reported typical radiation doses of 1 mSv for CAC scoring in CSCT (Messenger et al., 2015), while modern techniques allow CCTA acquisitions with 1.5 mSv radiation dose (Al-Mallah et al., 2014). Hence, performing calcium scoring in CCTA and omitting acquisition of the CSCT scan could reduce the radiation dose of a cardiac CT examination by 40–50% (Voros and Qian, 2012).

In clinical practice, CAC is standardly quantified in CSCT by manual identification of groups of connected voxels in the coronary artery that are above a 130 HU threshold and subsequent automatic 3D region growing (Agatston et al., 1990). This procedure is not applicable to CCTA, due to intravascular contrast material that typically enhances the arterial lumen well beyond 130 HU (Figs. 1(b) and (f)). Hence, higher global detection thresholds, ranging from 320 HU (Otton et al., 2012) to 600 HU (Glodny et al., 2009) have been proposed to emulate CAC scoring in CCTA. However, these fixed thresholds do not consider variations in lumen attenuation in CCTA, which might occur depending on protocols, scanners or contrast agents (Figs. 1(c) and (g)). This variation can be taken into account by using patient-specific or scan-specific attenuation thresholds, based on HU values taken from a ROI in the ascending aorta (Mylonas et al., 2014) or the proximal coronary arteries (Pavitt et al., 2014) (Figs. 1(d) and (h)).

Manual identification of CAC in cardiac CT requires substantial expert interaction, which makes it time-consuming and infeasible for large-scale or epidemiological studies. To overcome these limitations, (semi)-automatic calcium scoring methods have been proposed for CSCT (see e.g. Išgum et al. (2007); Kurkure et al. (2010); Shahzad et al. (2013); Wolterink et al. (2015a) and Ding et al. (2015)). Wolterink et al. (In press) provide a comparison of (semi-)automatic methods for calcium scoring in cardiac CT exams. Similarly, methods have been developed for auto-

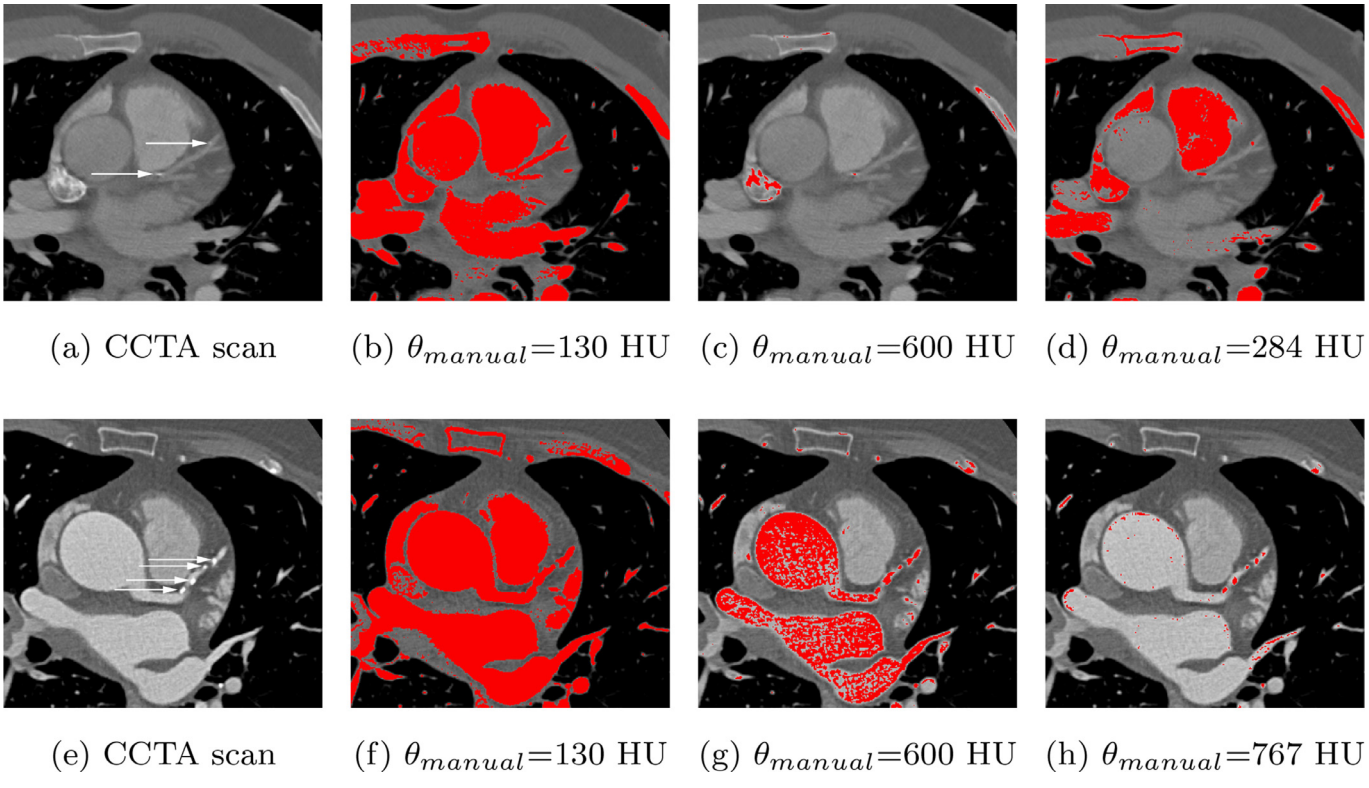


Fig. 1. Manual CAC identification in CCTA using diverse thresholds (θ_{manual}). (a), (e) Two example CCTA images with CAC indicated by white arrows. Voxels with attenuation $> \theta_{manual}$ shown in red. (b), (f) $\theta_{manual} = 130$ HU (Agatston et al., 1990) oversegments CAC in both images. (c),(g) $\theta_{manual} = 600$ HU (Glodny et al., 2009) misses one CAC lesion in (c) and oversegments CAC in (g). (d), (h) a patient-specific threshold based on attenuation in the ascending aorta (Mylonas et al., 2014) identifies the individual CAC lesions. Window and level are the same for all images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article as: J.M. Wolterink et al., Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks, Medical Image Analysis (2016), http://dx.doi.org/10.1016/j.media.2016.04.004

matic calcium scoring in CCTA. These methods typically require a (semi)-automatically extracted segmentation of the coronary arteries. Based on this segmentation, CAC has been identified as deviation from a trend line through the lumen intensity (Wesarg et al., 2006; Ahmed et al., 2014), as voxels in the extracted arteries with intensities above a patient-specific HU threshold (Teßmann et al., 2011), or as deviations from a model of non-calcified artery segments (Eilot and Goldenberg, 2014). Mittal et al. (2010) did not use a model or threshold to identify CAC, but trained classifiers to identify CAC lesions along an extracted coronary artery centerline. Coronary artery tree extraction methods generally show good performance, but they have been reported to fail in patients with complex anatomy, in the distal segments of the coronary arteries, in scans with motion or noise artifacts and in scans with occlusions in the coronary arteries. In addition, severe CAC deposits affect the performance of artery extraction algorithms, restricting their applicability in CAC identification (Schaap et al., 2009). Manual correction of incorrectly segmented coronary arteries is often time-consuming and tedious.

We propose identification of CAC without initial coronary artery tree extraction. In contrast to previously proposed methods, our algorithm uses supervised learning to directly identify CAC in CCTA. Supervised learning using nearest-neighbor, SVM and randomized decision tree classifiers has been previously applied to CAC identification in CSCT (e.g. (Isgum et al., 2012; Wolterink et al., 2015a; Shahzad et al., 2013)). However, these methods cannot be applied in CCTA, as they classify potential CAC lesions, extracted using a clinical 130 HU threshold. In CCTA, it is non-trivial to distinguish between CAC and attenuated lumen, and the application of a predefined single detection threshold to extract potential CAC lesions is not feasible. Instead, the proposed method identifies CAC voxels to segment lesions.

CAC voxel identification in CCTA is a challenging and extremely unbalanced classification problem. The proposed algorithm therefore first limits the volume-of-interest (VOI) to a bounding box around the heart, extracted using our previously proposed algorithm (de Vos et al., 2016). Thereafter, voxels in this VOI are classified using convolutional neural networks (ConvNets). Recently, ConvNets have been successfully used in natural image classification, image segmentation and object detection. In addition, they have been used in several medical image analysis tasks, for example knee cartilage segmentation (Prasoon et al., 2013) lymph node detection (Roth et al., 2014), brain tissue segmentation (Stollenga et al., 2015), and pulmonary nodule classification (Ciompi et al., 2015). In the proposed algorithm, ConvNets automatically extract texture features from triplanar 2.5D or volumetric 3D input samples, which are combined with spatial features derived from a normalized coordinate system defined in the VOI. To classify voxels as CAC or non-CAC, a pair of ConvNets is used. These ConvNets are linked by training and together are called a ConvPair. The first ConvNet identifies voxels likely to be CAC. Such voxels are further classified by the second ConvNet, which distinguishes between CAC and CAC-like negatives. We propose a purely convolutional ConvNet architecture, which allows for fast evaluation times and can be directly applied to arbitrarily sized CCTA images. In addition, we present experiments showing that combinations of different architectures can achieve higher CAC identification performance than individual architectures.

We have previously proposed a method for CAC scoring in CCTA using a combination of a ConvNet and a Random Forest classifier (Wolterink et al., 2015b). This work extends our previous work in several ways. First, the classification procedure has been modified. Our previously proposed method used a ConvNet for voxel classification and a Random Forest classifier for lesion classification. The current method uses two sequential ConvNets for voxel classification. Second, in our previous work, candidate voxels for classification were selected based on the image intensity histogram. In the current work, we classify all voxels within the VOI, regardless of intensity, hence no assumptions are made about CAC HU values. Third, location features were previously extracted using a time-consuming elastic registration preprocessing step. In the current method, this registration step is omitted in favor of our very fast ConvNet-based bounding box detection technique (de Vos et al., 2016). Fourth, in our previous work we only evaluated triplanar 2.5D input with one input size. In the current work, we provide a comparison between 2.5D and volumetric 3D input, between input with different sizes, as well as experiments with ensembles combining these input representations. Fifth, the ConvNet architecture in our previous work required a time-consuming scan algorithm with many redundant operations for neighboring candidates. Here, we use a purely convolutional network for efficient voxel classification. Finally, in this work an evaluation on a substantially larger set of scans has been performed, and a thorough comparison with clinically used CSCT CAC scores, as well as interobserver variability, are provided.

2. Data

In this study, clinically obtained cardiac CT exams of 250 consecutively scanned patients were included. Each exam consists of a CSCT and a CCTA scan, made on a 256-detector row scanner (Philips Brilliance iCT, Philips Medical, Best, The Netherlands). The CSCT scans were acquired using a standard calcium scoring protocol with 120 kVp tube voltage and 55 mAs tube current, with ECG-triggering and without contrast enhancement. Reconstructed sections had 3.0 mm spacing and thickness. The CCTA scans were acquired with 120 kVp tube voltage and 210–300 mAs tube current, with ECG-triggering and contrast enhancement. Reconstructed sections had 0.45 mm spacing and 0.90 mm thickness. In both CSCT and CCTA, in-plane resolution was 0.4–0.5 $\times$ 0.4–0.5 mm.

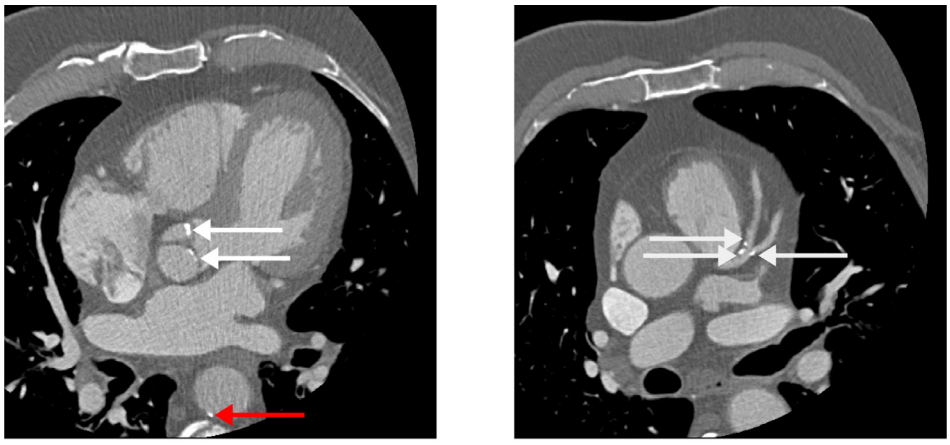
The set of 250 cardiac CT exams was divided into two sets. The first 50 exams were used to train an algorithm that detects bounding boxes around the heart. The remaining 200 exams were used to train and evaluate a voxel classification algorithm that identifies CAC in these bounding boxes. Two expert observers provided annotations in all (observer O_1) or in a subset (observer O_2) of the exams.

In each of the 50 cardiac CT exams used to train the bounding box detection algorithm, observer O_1 manually determined a 3D rectangular bounding box around the heart in the CCTA scan. This bounding box included the pericardial sac, from below the pulmonary artery to the apex in the craniocaudal direction.

In each of the 200 cardiac CT exams used to train and evaluate the voxel classification algorithm, manual reference annotations for CAC were obtained in the CCTA and the CSCT scan. Manual annotations in the CCTA scans were obtained similarly to the methods proposed in (Mylonas et al., 2014; Pavitt et al., 2014). The expert observer first manually identified a point in the center of the ascending aorta at the level of the origin of the left coronary artery. This point was automatically grown to a 200 mm^3 volume of interest (VOI). The mean ($mean_{aorta}$) and standard deviation (SD_{aorta}) of HU values in this ROI were used to compute a patient-specific threshold $mean_{aorta} + 3SD_{aorta}$. The expert observer then marked calcification in the coronary artery by a mouse click on a single voxel, which was followed by automatic 3D region growing of voxels with density above the defined threshold. In addition, to compare obtained CAC scores in CCTA to the clinically used standard, CAC in CSCT scans was also manually identified with a clinically used threshold of 130 HU (Agatston et al., 1990).

Observer O_1 annotated CAC in all 200 cardiac CT exams. These annotations were considered the reference standard, used for training and evaluation of the voxel classification algorithm. Observer

Please cite this article as: J.M. Wolterink et al., Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks, Medical Image Analysis (2016), http://dx.doi.org/10.1016/j.media.2016.04.004



(a) (b)

Fig. 2. CAC-like voxels in CCTA: (a) aortic valve calcification (white arrows), calcification in the descending aorta (red arrow), and (b) CAC in the left anterior descending artery (white arrows). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

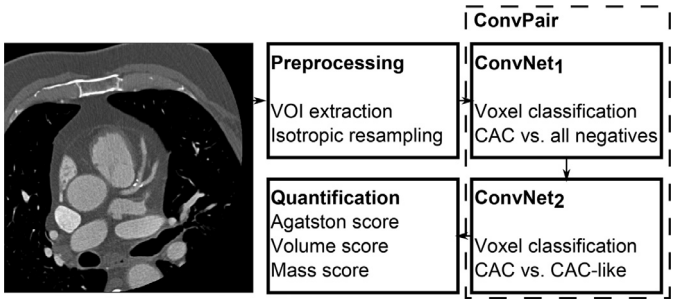


Fig. 3. Overview of the proposed system. The image is preprocessed by determination of a bounding box around the heart and isotropic resampling. Voxels resembling CAC are extracted and subsequently classified by a pair of ConvNets (ConvPair). ConvNet_1 identifies CAC-like voxels, among which ConvNet_2 distinguishes between CAC and other CAC-like candidates. The resulting segmentation is quantified using the CAC Agatston, volume and mass score.

O_2 annotated CAC in a subset of 100 cardiac exams. These annotations were used to determine interobserver variability and for comparison with the automatic method.

3. Method

CAC was identified by voxel classification. Besides CAC, a typical CCTA scan contains many other voxels of appearance similar to CAC. These include extracardiac lesions like bones such as ribs, calcifications in the descending aorta and calcified lymph nodes, as well as intracardiac calcifications such as those in the mitral and aortic valve (Fig. 2). In addition, coronary artery lumen is often highly attenuated, hence resembling CAC.

The proposed algorithm is illustrated in Fig. 3. First, a bounding box around the heart is determined. This excludes most extracardiac calcifications and allows further analysis within this VOI only. Next, voxels in the VOI are classified with a pair of ConvNets (ConvPair), which share the same structure but have differently trained parameters. The first ConvNet (ConvNet_1) detects voxels likely to be CAC among all candidate voxels. The second ConvNet (ConvNet_2) separates CAC from CAC-like voxels, such as attenuated coronary artery lumen and aortic calcifications. Finally, identified CAC is quantified.

3.1. Preprocessing

CCTA scans are generally acquired with a standardized scan length in the craniocaudal direction ranging from mid-pulmonary artery to diaphragm (Raff et al., 2014). However, their field of view in the transverse plane is less standardized. Some scans might contain the ribs and spine, while others may be closely cropped around the heart. To reduce this variation, and to allow analysis only within the VOI, the field of view is standardized by finding a 3D rectangular bounding box around the heart. This bounding box is automatically determined using our previously developed algorithm described by de Vos et al. (2016). The algorithm uses three independent ConvNets, each determining the presence of the heart in the axial, sagittal or coronal view. A rectangular 3D bounding box around the heart is obtained by combining posterior probabilities obtained for axial, sagittal and coronal image slices.

ConvNets label a target voxel based on a square or cubic input patch centered at that voxel. For this, it is beneficial to have identical receptive fields along all image axes. Therefore, images cropped to the determined bounding box are resampled with B-spline interpolation to 0.45 mm isotropic voxels – the standard slice spacing in our data set. Finally, to allow robust ConvNet training, all data is rescaled to the mean and standard deviation of HU values in the training images.

3.2. Voxel classification

All voxels in the VOI are considered candidates for CAC. The proposed ConvNets take one or multiple patches of size w voxels centered at the candidate voxel as input and extract features based on that input. Using these features, the voxel is assigned a probability of being CAC p_{CAC}. Input patches are either 2.5D, i.e. three 2D patches from orthogonal image planes centered at the voxel, or 3D, i.e. volumetric patches centered at the voxel.

Features in a ConvNet are typically extracted by a stack of convolution layers, while classification is done by a stack of fully-connected hidden layers. The proposed method uses a purely convolutional ConvNet architecture for both the feature extraction stack and the classification stack, suitable for both 2.5D and 3D input. Fig. 4 shows an overview of the proposed architecture.

Depending on the size and the dimensionality of the input patches, the architecture of the feature extraction stack is generated as follows. Each convolutional layer with the exception of the last one consists of 16 small convolution kernels of 3×3 voxels in 2.5D or $3 \times 3 \times 3$ voxels in 3D. Choosing multiple stacked small convolution kernels over one larger convolution kernel has been shown to have two advantages (Simonyan and Zisserman, 2015). First, more stacked layers contain more non-linear activation layers, hence allowing the network to become more discriminative.

Please cite this article as: J.M. Wolterink et al., Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks, Medical Image Analysis (2016), http://dx.doi.org/10.1016/j.media.2016.04.004

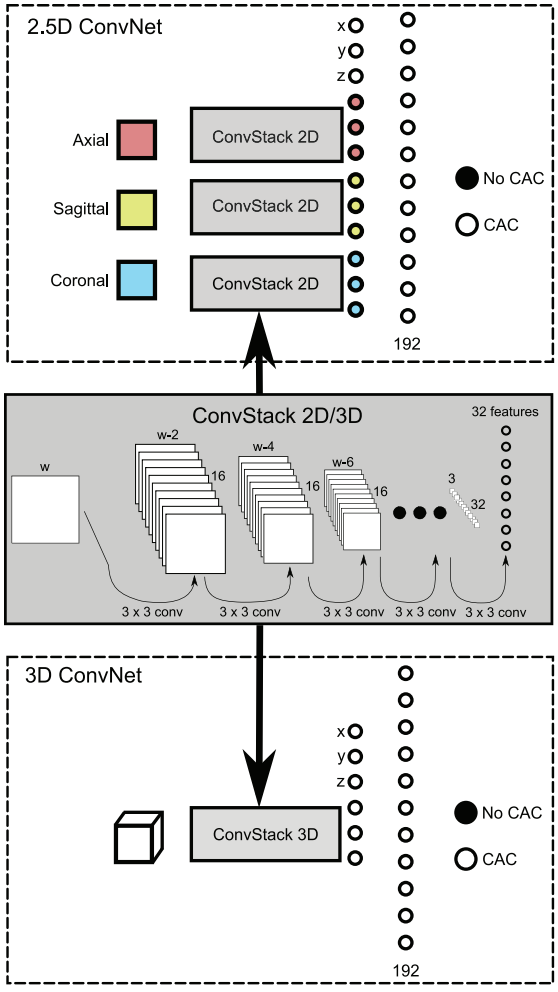


Fig. 4. Proposed convolutional neural network (ConvNet) architecture for 2.5D and 3D input. The network consists of a stack of feature extraction layers and a stack of classification layers. The feature extraction stack (ConvStack) has the same hyperparameters in both architectures. It consists of a sequence of layers with 3×3 voxel (in 2D) or $3 \times 3 \times 3$ voxel (in 3D) convolution kernels. The 2.5D ConvNet combines features from three identical 2D ConvStacks with shared weights, each processing an input patch from a different orthogonal viewing direction, i.e. axial, sagittal and coronal. These input patches are centered at the target voxel. The 3D ConvNet uses volumetric features extracted from a 3D input patch centered at the target voxel. In both the 2.5D and the 3D ConvNet, features are concatenated with *x*, *y* and *z* location features and connected to an output layer through one hidden layer.

Second, stacking small kernels reduces the number of trainable parameters, and hence the chance of over-fitting.

Convolutions are valid, i.e. no zero-padding is applied after convolution, so that each convolution reduces the input size by 2 voxels along each axis. In the final convolution layer, 32 convolution kernels reduce the input size to 1 voxel along each axis. The 32 obtained features are used for classification. Each convolution layer was followed by a rectified linear unit (ReLU) activation function (Glorot et al., 2011).

The convolutional stack does not include any max-pooling downsampling layers. These layers are typically used in image classification and object detection to rapidly decrease the input size, to reduce the number of weights in the network to prevent overfitting and to introduce spatial invariance. However, the spatial invariance introduced by pooling could mean that neighboring voxels are assigned the class label that is most expressed in that location. This in turn could lead to over- or undersegmentation of CAC lesions, which are generally small. In addition, the absence of pooling layers means that the convolutional stack is purely convolutional. Hence, the convolutions may be applied to full images, thereby avoiding redundant convolution operations.

A 2.5D ConvNet contains three convolutional stacks, which independently process axial, sagittal and coronal input. The three networks share weights, i.e. one 2D network was used for feature extraction in the three orthogonal planes, similar to the shared weight multi-scale approach in (Farabet et al., 2012). Tying the weights in these networks reduces the number of features and allows robust generic texture feature extraction. A 3D ConvNet contains one volumetric convolutional stack.

The features extracted by the convolutional stack are used to classify the target voxel as either CAC or non-CAC. The extracted features might only provide limited spatial information. However, studies on automatic CAC scoring in non-contrast-enhanced cardiac CT have shown that location information is essential for CAC identification (Isgum et al., 2012; Shahzad et al., 2013; Wolterink et al., 2015a). Therefore, a normalized heart coordinate system is used to describe the location of each voxel within the VOI. In this coordinate system, the origin is located at the center of the VOI and -1 and 1 are positioned at the boundaries of the VOI along each axis. For each candidate, the *x*-, *y*- and *z*-coordinate are determined as location features. These location features are concatenated to the texture features derived by the network to provide a feature vector.

The feature vector serves as input to one fully-connected hidden layer. This hidden layer is connected to an output layer with a softmax function to predict probabilities p_{CAC} and $1 - p_{CAC}$. To regularize the network, Dropout is applied before and after the hidden layer.

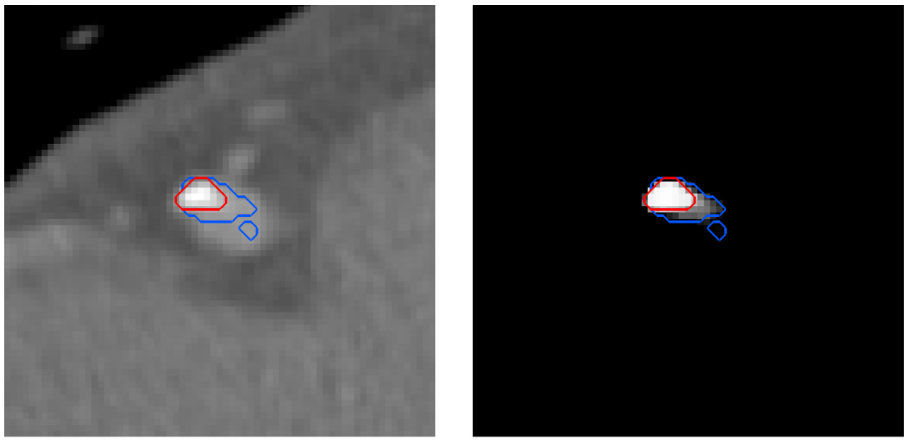
3.3. Training strategy

CCTA scans contain many more negative (background) than positive (CAC) samples, representing a heavily unbalanced classification problem. Identifying CAC among all voxels in a cardiac VOI in CCTA poses two challenges. The vast majority of negatives such as those representing lung or fatty tissue, share very few similarities with CAC. Hence, given sufficiently descriptive features, they might easily be discarded. Other negatives, such as bone (e.g. sternum), calcifications in the ascending aorta and coronary artery lumen enhanced by contrast material, are more challenging to distinguish from CAC. Our method uses a ConvPair, a pair of ConvNets, each of which have a specific task. ConvNet_1 focuses on detection of CAC-like voxels, ConvNet_2 identifies CAC voxels among these candidates.

The two ConvNets are trained in sequence. ConvNet_1 is trained first, using all voxels in the VOIs of the calcium scoring training images. This ConvNet learns to discard the vast majority of negative voxels. For each calcium scoring training image, a CAC candidate mask is obtained using ConvNet_1. This mask contains CAC-like voxels, but no negatives such as lung tissue or fatty tissue. Subsequently, ConvNet_2 is trained using only the samples in the CAC candidate mask. To leverage already learned knowledge, the ConvNet_2 is initialized as the final version of ConvNet_1 and training is resumed using the samples from the CAC candidate mask. Hence, ConvNet_1 and ConvNet_2 share their architecture but not their trainable parameters.

During testing, ConvNet_1 and ConvNet_2 in a ConvPair can be evaluated sequentially, i.e. by first obtaining a candidate mask from ConvNet_1 and classifying only those candidate voxels in the mask with ConvNet_2. Alternatively, both ConvNets may be merged into one network with two parallel stacks of layers, where the first stack contains ConvNet_1 and the second stack contains ConvNet_2. To obtain a probabilistic CAC map, p_{CAC} values generated by ConvNet_1 are thresholded and the resulting binary image is multiplied with the p_{CAC} values generated by ConvNet_2.

Please cite this article as: J.M. Wolterink et al., Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks, Medical Image Analysis (2016), http://dx.doi.org/10.1016/j.media.2016.04.004



(a) CCTA scan (b) p_{CAC}

Fig. 5. (a) CCTA scan showing a right coronary artery (RCA) CAC lesion, and (b) aligned posterior CAC probability p_{CAC} map. The blue contour shows manual lesion segmentation based on a patient-specific intensity threshold in the CCTA scan, and the red contour shows automatic lesion segmentation based on a threshold θ_{CAC} in the posterior probability map. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. Implementation

All ConvNets were implemented using Theano (Bastien et al., 2012). The purely convolutional nature of the networks was exploited to rapidly process an image of size $nx \times ny \times nz$ voxels. In 2.5D, full image slices along the three principal axis were independently processed. This resulted in three 4D texture feature tensors of size $nx \times ny \times nz \times nf$, where $nf = 32$ in the proposed architecture (Fig. 4). These tensors were concatenated along the last dimension. Similarly, in 3D one subimage of size $nx \times ny \times w$ voxels was processed for every axial slice, to obtain a texture feature tensor. In both 2.5D and 3D, the texture feature tensor was concatenated with a $nx \times ny \times nz \times 3$ location feature tensor. Fully connected hidden layers were implemented as convolutions with kernel size 1. Hence, two consecutive convolutions for the hidden and output layer resulted in a probability distribution for each voxel.

3.5. Evaluation

Reference lesions were segmented using 26-connected 3D region growing of voxels above a patient-specific intensity threshold (Mylonas et al., 2014). Although this patient-specific threshold identifies CAC better than a global threshold (Fig. 1), it is based on aortic attenuation, which can differ from coronary attenuation due to for example luminal narrowing, imaging artifacts or partial volume effects. Hence, intensity-based region growing may under- or oversegment CAC lesions (Fig. 5). Therefore, automatically obtained lesions were not segmented based on attenuation, but based on posterior probabilities, using 26-connected 3D region growing of voxels with a CAC probability $p_{CAC} \geq \theta_{CAC}$, where p_{CAC} was predicted by the ConvPair and θ_{CAC} was determined using ROC analysis. Lesions smaller than 1.0 mm^3 were discarded as these likely represented noise.

Given that CAC lesions in CCTA were created differently in the reference and automatic results, they might not contain the same voxels. Therefore, true positive lesions were automatically found lesions having overlap with the reference lesions. False positive lesions were those in the automatic result having no overlap with the reference lesions. False negative lesions were lesions identified in the reference that had no overlap with any automatically found lesion.

In addition to an evaluation of the ability of the method to detect individual lesions, its ability to determine per patient Agatston, volume and mass CAC scores was established. The Agatston score weighs calcified plaque area by peak intensity, and was computed as $\sum_{s \in S} a_s \cdot d_s \cdot \frac{\Delta_z}{3.0}$, where S is the set of slices containing the lesion, a_s is the lesion area (in mm^2) in s, d_s is a density factor based on the lesion's peak intensity in s (130–199 HU: 1, 200–299 HU: 2, 300–399 HU: 3, $\geq$ 400 HU: 4), and Δ_z is the image slice increment. A linear scaling factor $\frac{\Delta_z}{3.0}$ is standardly used to correct for image slice increments different than the 3.0 mm increment for which the method was originally developed (Agatston et al., 1990). The CAC volume score (in mm^3) quantifies CAC volume and was computed as the number of identified voxels multiplied by voxel volume (in mm^3). The CAC equivalent mass score weighs lesion attenuation linearly and was computed as the product of the volume of a lesion and its mean intensity (McCollough et al., 2007). Per lesion CAC scores were summed to obtain per patient scores.

Automatically obtained volume and mass scores in CCTA were compared with reference volume and mass scores in CCTA by observer O_1. In current clinical practice, CAC burden is determined in CSCT. Hence, automatically obtained volume and mass scores in CCTA were compared with reference volume and mass scores in CSCT. In addition, manual volume and mass scores defined by observer O_1 and observer O_2 in CCTA scans were compared with the reference scores in CSCT. Finally, to establish interobserver agreement, volume and mass scores in CSCT and CCTA scans were compared between the observers. Agreement between two measurements was determined using the intra-class correlation coefficient (ICC) for absolute agreement, with 95% confidence interval. In addition, Bland–Altman plots were generated and the bias and limits of agreement ($\pm$1.96 SD) were reported.

In clinical practice, patients are assigned to a CVD risk category based on their Agatston score. Because CVD risk categorization is not defined for Agatston scores in CCTA, agreement between standard Agatston score based risk categorization in CSCT and the Agatston score in CCTA was determined as follows: patients were first categorized based on their CSCT Agatston score according to standard risk categories (Very low: 0, Low: 1-100, Intermediate: 101-400, High: >400). Thereafter, patients were ranked based on their CCTA Agatston score. Based on this ranking, to each of four CCTA risk categories the same number of patients was assigned as in the corresponding CSCT category. Patients who ended up in the same category based on both CSCT and CCTA Agatston score were correctly categorized, patients who were assigned to a different category were incorrectly categorized. Categorization accuracy and Cohen's linearly weighted κ were computed.

Please cite this article as: J.M. Wolterink et al., Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks, Medical Image Analysis (2016), http://dx.doi.org/10.1016/j.media.2016.04.004

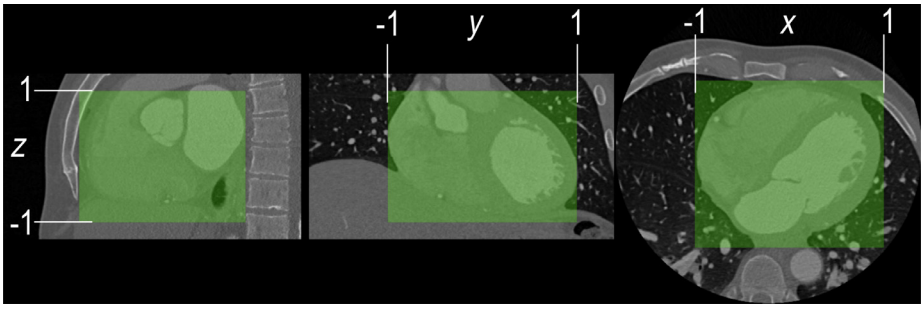


Fig. 6. Automatically obtained heart bounding box in CCTA in sagittal, coronal and axial view, respectively. The bounding box reduces the volume of interest (VOI) by up to 80%. In addition, the boundaries provide a normalized coordinate space with the origin in the center of the box.

Table 1
Details of the evaluated ConvNet pairs. Number of layers, without input and output layers (#Layers), number of trainable weights (#Weights), average and SD processing time per image for $ConvNet_1$ in s ($Time_1$) and for $ConvNet_2$ in s ($Time_2$).

		#Layers	#Weights	$Time_1$	$Time_2$
$w = 15$	2.5D	8	35,986	42 ± 5	26 ± 4
	3D	8	56,242	52 ± 6	107 ± 44
$w = 25$	2.5D	13	47,586	46 ± 7	28 ± 4
	3D	13	90,882	147 ± 15	201 ± 113

Table 2
Lesion identification with respect to the reference standard in CCTA. Four ConvPairs with dimensionality 2.5D or 3D, and input patch size $w = 15$ or $w = 25$ were trained. Ensembles of ConvPairs were formed by averaging of probabilities. Bullet points indicates membership of an ensemble. Lesion identification sensitivity is reported, as well as the average number of false positive (FP) lesions per scan.

2.5D		3D		Sens.	FP/scan
$w = 15$	$w = 25$	$w = 15$	$w = 25$		
•				68%	0.90
	•			72%	0.85
		•		67%	1.69
			•	72%	1.21
•	•			71%	0.64
		•	•	69%	0.77
•		•		71%	0.68
	•		•	71%	0.57
•	•	•	•	71%	0.48

4. Experiments and results

The set of 250 exams was divided into four sets. First, a set of 50 exams was used to train the bounding box extraction algorithm. Second, a training set of 90 exams was used to train the ConvNet pairs for CAC classification. Third, a validation set of 10 exams was used to optimize hyperparameters of the ConvNets. Finally, a test set of 100 exams was only used to evaluate the performance of the method. For the test set, annotations by both observer O_1 and O_2 were available.

4.1. Bounding box extraction

Bounding box extraction took 6.9 ± 0.5 s, discarding up to 80% of the original CCTA image. Fig. 6 shows an example of an automatically determined bounding box. Each VOI contained around 20,000,000 negative voxels and on average 800 positive voxels. Visual inspection of the results showed that in all cases the bounding box contained the whole heart.

4.2. Experimental settings

Network weights were initialized according to the procedure specified by Glorot and Bengio (2010). Dropout probability in fully connected layers was set to $p = 0.5$ (Srivastava et al., 2014). The categorical cross-entropy between reference and predicted labels was minimized using stochastic gradient descent with Nesterov momentum and learning rate $\alpha = 0.001$. $ConvNet_1$ was trained with 200,000 mini-batches, $ConvNet_2$ with 100,000 mini-batches. Mini-batches were balanced, containing 64 negative and 64 positive samples. Probability maps generated by $ConvNet_1$ were thresholded at $p_{CAC} \geq 0.5$ to provide a mask for CAC detection. All models were trained and tested on single Titan X GPUs.

Four ConvPairs were trained, namely for 2.5D and 3D input patches with size $w = 15$ and $w = 25$. Table 1 lists the number of layers and the number of trainable parameters of each ConvPair architecture, as well as the average time required for processing by its components $ConvNet_1$ and $ConvNet_2$. For 2.5D input, the networks with input size $w = 25$ took slightly more time than the network with size $w = 15$. In 3D, the difference in required processing time between input sized $w = 15$ and $w = 25$ was larger.

4.3. Lesion identification

ConvPairs are specified by input dimensionality and input size, and results are presented for merged output of their members $ConvNet_1$ and $ConvNet_2$. Automatically extracted lesions were compared to the reference standard in CCTA test scans, which in total contained 260 CAC lesions.

Table 2 lists sensitivity for lesion identification, as well as the average number of false positive (FP) errors per scan. Results are shown for individual ConvPairs, as well as ensembles combining ConvPairs with different input sizes and dimensionality. Recent large-scale 2D natural image classification challenges have been dominated by ensembles of ConvNets. It has been shown that combinations of ConvNets improve classification results, particularly when ConvNets have been trained with different hyperparameters.

The results indicate that among architectures with the same dimensionality, those with an input size $w = 25$ perform better than those with an input size $w = 15$. Among architectures with the same input size, 2.5D networks obtain better results than 3D network. The strongest individual architecture combines the largest input size $w = 25$ with a 2.5D input representation.

In all cases, ensembles make fewer FP errors per scan at a similar sensitivity level. The ensemble with $w = 25$ outperforms the ensemble with $w = 15$, and the 2.5D ensemble outperforms the 3D ensemble. An ensemble of all ConvPairs provides the best result, at 71% lesion sensitivity and 0.48 FP error per scan.

Fig. 7 shows typical examples of false negative (FN) and FP errors. A CAC lesion of low density compared to the surrounding

Please cite this article as: J.M. Wolterink et al., Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks, Medical Image Analysis (2016), http://dx.doi.org/10.1016/j.media.2016.04.004

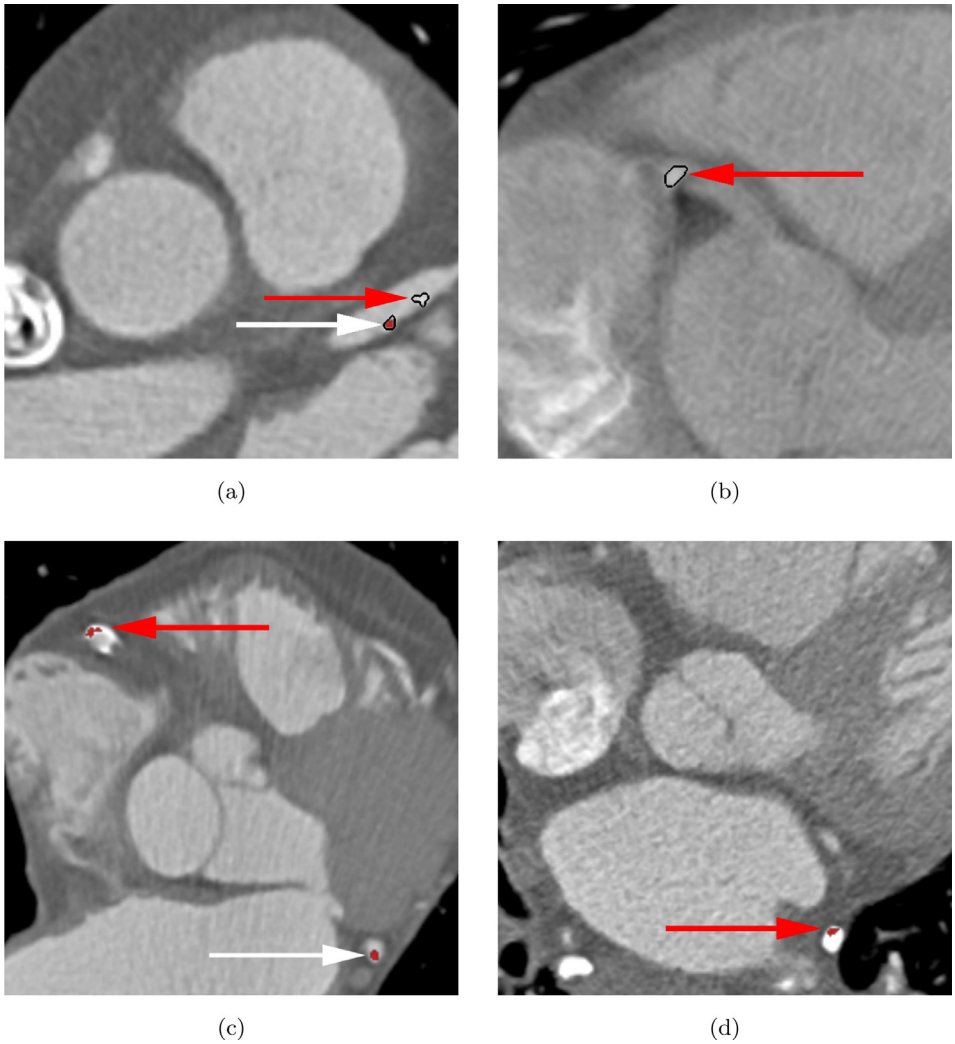


Fig. 7. Examples of false negative (FN) and false positive (FP) errors. (a) Two CAC lesions in the left anterior descending artery. One was missed by the algorithm (red arrow), one was found (white arrow). (b) CAC lesion affected by motion artifact in right coronary artery (RCA) which was not identified by the algorithm. (c) CAC lesion in the left circumflex artery which was correctly identified (white arrow), and part of a stent which was incorrectly identified as CAC (red arrow). (d) Calcified lymph node, part of which was incorrectly identified as CAC (red arrow). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

lumen in the left anterior descending (LAD) artery was missed (Fig. 7(a)), as well as a CAC lesion that was deformed by a motion artifact in the RCA (Fig. 7(b)). In a small number of scans, all ConvPairs and ensembles of ConvPairs made the same errors. One scan contained a stent in the RCA, parts of which were incorrectly identified as CAC (Fig. 7(c)). A second scan contained large calcified lymph nodes, parts of which were incorrectly identified as CAC (Fig. 7(d)). Several other false positive errors were due to high HU values in the coronary artery lumen.

The effect of ConvNet_1 and ConvNet_2 on voxel classification was investigated. Fig. 8 shows typical probabilistic maps obtained by ConvNet_1 and ConvNet_2, obtained with 2.5D input with size $w = 25$. While ConvNet_1 detected CAC, it also detected lumen, bone in the rib and aortic calcification (Fig. 8(b)). ConvNet_2 correctly identified CAC, but also gave a weak response in voxels dissimilar to those which were used to train ConvNet_2, i.e. high-density voxels similar to CAC (Fig. 8(c)). Because the network was fine-tuned only with samples similar to CAC, it had lost its ability to classify other candidates. Finally, in the merged result of ConvNet_1 and ConvNet_2 (Fig. 8(d)), only CAC received a high probability.

To investigate the effect of the normalized $x, y, z-$coordinates on CAC identification, an additional ConvPair was trained with the best performing architecture, i.e. 2.5D input with $w = 25$, adapted to omit the feature coordinates. ROC analysis showed that at all sensitivity levels, this ConvPair made slightly more FP errors than the ConvPair with location features.

4.4. Per patient CAC quantification

For each patient, the volume and mass scores in CCTA were determined by observer O_1, observer O_2 and the automatic method. Also, for the automatic method, Agatston scores in CCTA were determined. Automatic results were based on the complete ensemble of trained architectures.

Fig. 9 shows a Bland–Altman plot for the agreement between reference CAC scores in CCTA and automatically obtained CAC scores in CCTA. Automatically obtained CAC scores were typically lower than those in the reference standard. Bland–Altman bias and limits of agreement were −19.6 (−144.7–105.4) mm^3 for CAC volume and −10.8 (−81.8–60.2) for CAC mass score. The ICC was 0.768 (0.660–0.842) for CAC volume and 0.872 (0.808–0.915) for CAC mass score.

Fig. 10 shows Bland–Altman plots for the agreement between reference CAC scores in CSCT and CAC scores in CCTA determined

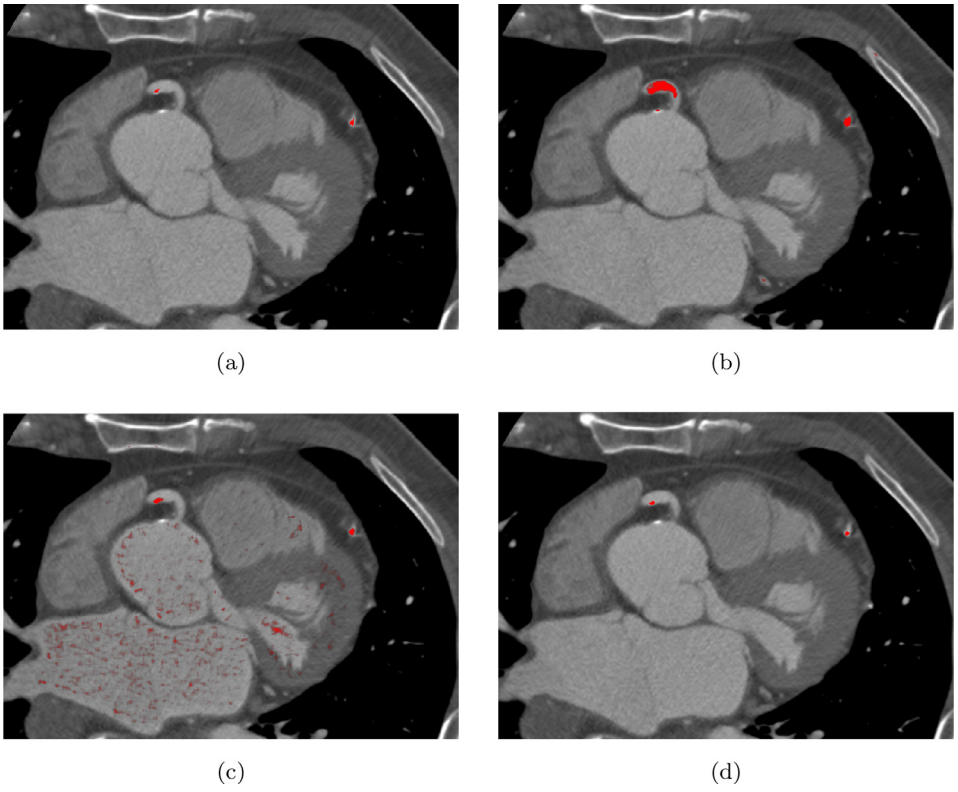


Fig. 8. (a) Overlay showing reference annotation with CAC lesions in the left anterior descending (LAD) and right coronary artery (RCA) in bright red. (b) Probabilistic map generated by ConvNet_1. The map contains high probabilities (bright red) for CAC, but also for coronary lumen and for calcification in the ascending aorta. (c) Probabilistic map generated by ConvNet_2. The map shows high probabilities for CAC (bright red), while the probabilities for coronary lumen and aortic calcification are zero. However, ConvNet_2 also assigns CAC probabilities to blood in the left atrium, left ventricle and ascending aorta (dark red), as it was specifically trained on CAC-like voxels. (d) thresholded merged probabilistic maps of ConvNet_1 and ConvNet_2, showing identified CAC voxels (bright red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

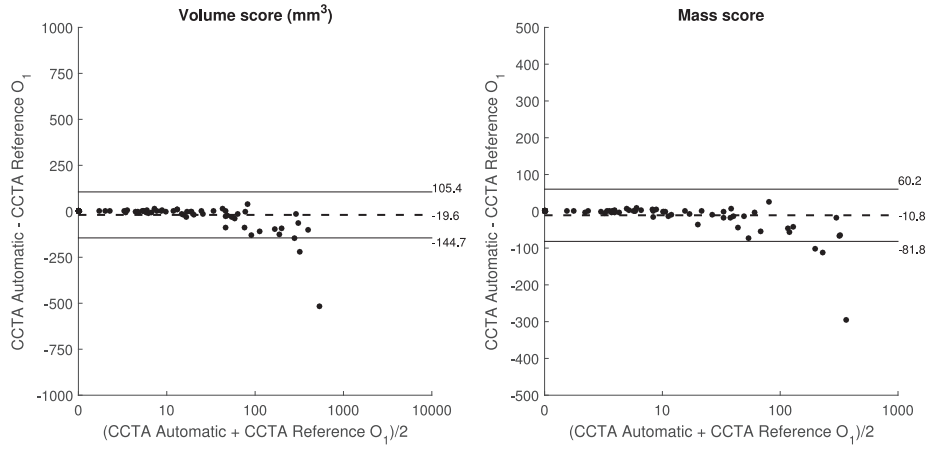


Fig. 9. Bland–Altman plots comparing automatically obtained CAC volume and mass scores in CCTA with reference annotations by observer O_1 in CCTA. Bland-Altman bias and limits of agreement are indicated.

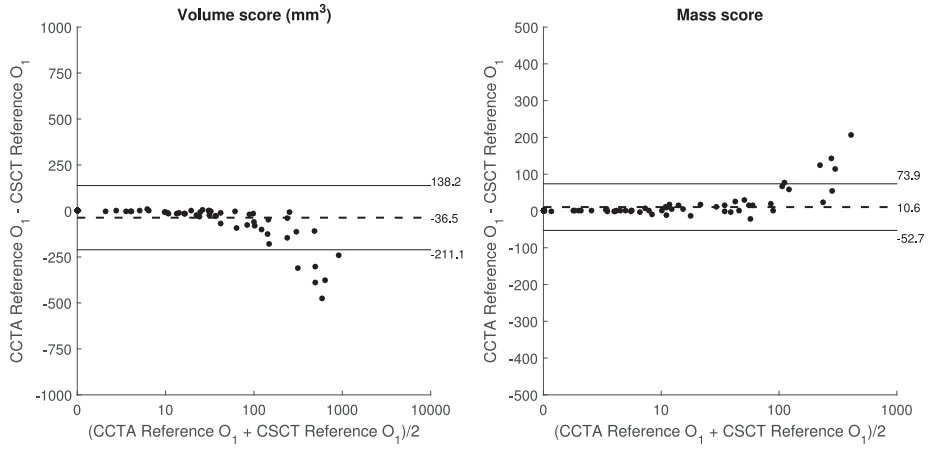
by observer O_1, observer O_2, and the automatic method. One patient was left out of the statistical analysis due to a large motion artifact in CSCT.

CAC volume in CCTA was lower than the reference CAC volume in CSCT. This effect was stronger for the automatic method than for the observers. Bland–Altman bias and limits of agreement were −36.5 (−211.1–138.2) mm^3 and −22.8 (−175.4–129.8) mm^3 for observers O_1 and O_2, and −56.1 (−319.4–207.2) mm^3 for the automatic method. The ICC was 0.828 (0.719–0.891) and 0.900 (0.848–0.934) for observers O_1 and O_2, and 0.538 (0.347–0.679) for the automatic method.

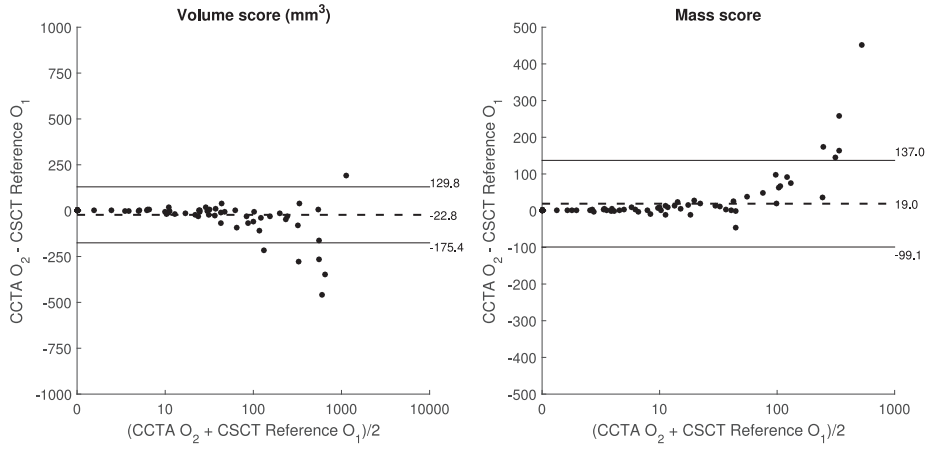
For CAC mass score, values in CCTA were lower than the reference in CSCT for the two observers, but not for the automatic method. Bland–Altman bias and limits of agreement were −10.6 (−52.7–73.9) for observer O_1, 19.0 (−99.1–137.0) for observer O_2, and −0.2 (−38.7–38.3) for the automatic method. The ICC was 0.895 (0.837–0.932) for observer O_1, 0.761 (0.650–0.838) for observer O_2, and 0.944 (0.918–0.962) for the automatic method.

The confusion matrix in Table 3 compares Agatston-score based risk categorization in the reference CSCT annotations and CVD risk categorization based on automatically determined Agatston scores in CCTA. CVD risk categorization accuracy was 83%, with

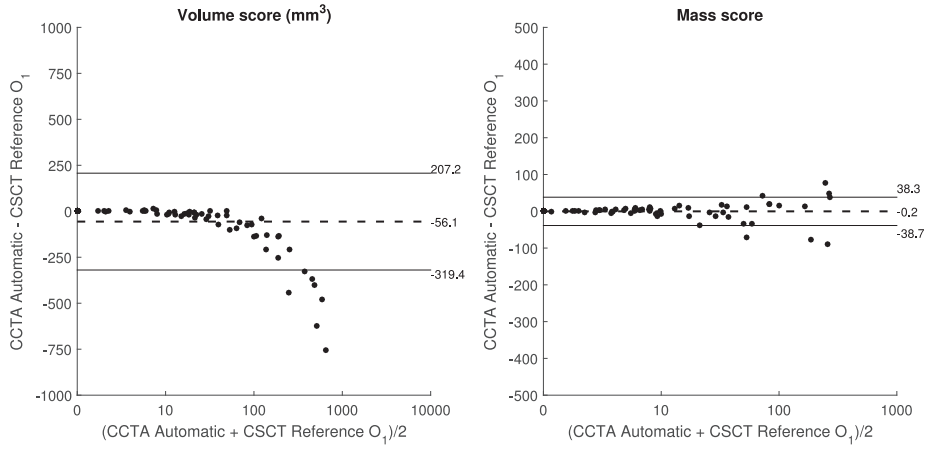
Please cite this article as: J.M. Wolterink et al., Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks, Medical Image Analysis (2016), http://dx.doi.org/10.1016/j.media.2016.04.004



(a) Observer O_1



(b) Observer O_2



(c) Automatic method

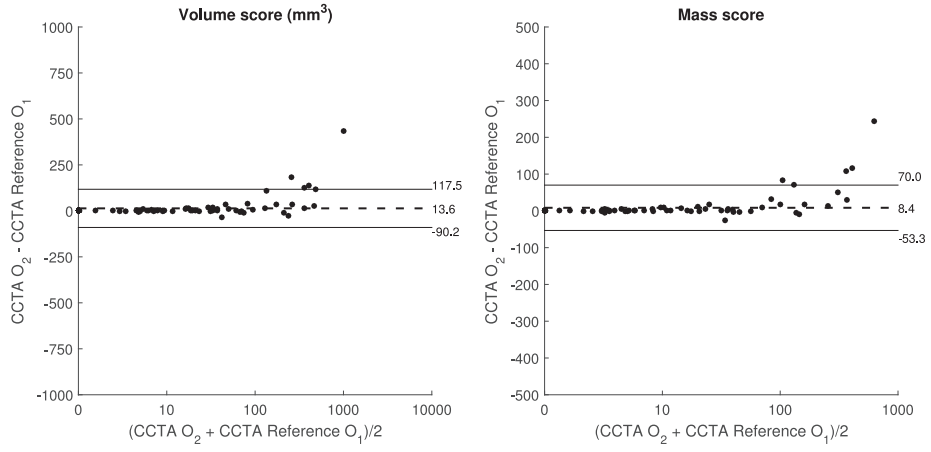
Fig. 10. Bland–Altman plots comparing reference CAC volume and mass scores in CSCT with annotations in CCTA by (a) observer O_1, (b) observer O_2 and (c) the automatic method. Bland–Altman bias and limits of agreement are indicated.

linearly weighted $\kappa = 0.83$. No patient was more than one category off. The test set contained 48 patients with zero CAC and 52 patients with a positive CAC score. The automatic method identified 43/48 patients with zero CAC. Conversely, the automatic method correctly identified 48/52 patients with a positive CAC score. Missed CAC lesions in the remaining patients were small and low-density lesions, with the exception of one patient whose scan contained a CAC lesion in the RCA which was deformed due to a motion artifact and therefore missed by the algorithm (Fig. 7b).

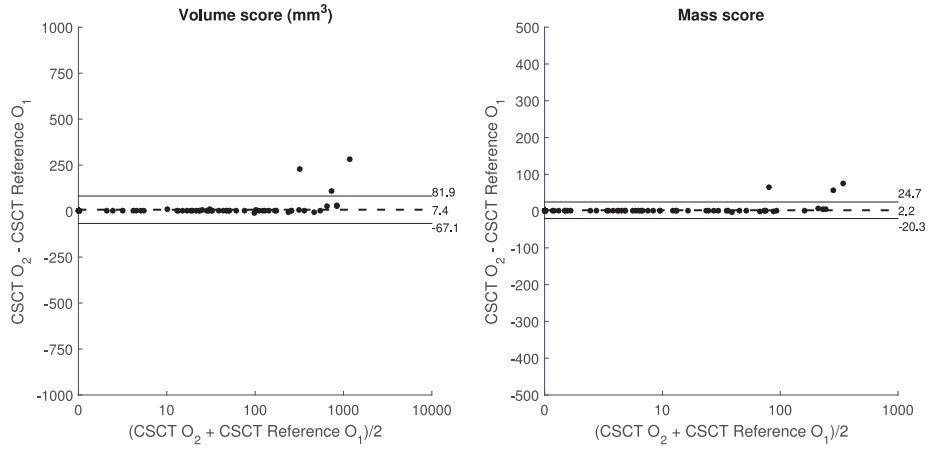
4.5. Interobserver agreement

Both observers scored CAC in the full test set. To score CAC in CCTA, a patient-specific CAC threshold was determined using a manually placed ROI in the ascending aorta. This threshold showed

Please cite this article as: J.M. Wolterink et al., Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks, Medical Image Analysis (2016), http://dx.doi.org/10.1016/j.media.2016.04.004



(a) Interobserver agreement in CCTA



(b) Interobserver agreement in CSCT

Fig. 11. Bland–Altman plots comparing (a) CAC volume and mass scores in CCTA and (b) CAC volume and mass scores in CSCT between observer O_1 and observer O_2. Bland–Altman bias and limits of agreement are indicated.

Table 3
Agreement in cardiovascular risk categorization based on Agatston score categories in the CSCT reference standard (I: 0, II: 1 – 100, III: 101 – 400, IV: >400) and derived categories in the automatically obtained CCTA Agatston score.

Automatic	Reference				Total
	I	II	III	IV	
I	**43**	4	0	0	47
II	5	**23**	3	0	31
III	0	3	**10**	1	14
IV	0	0	1	**7**	8
Total	48	30	14	8	100

excellent agreement between the two observers (ICC 0.994 [0.990–0.997]). The thresholds ranged from 332 to 898 HU for observer O_1 and from 333 to 930 HU for observer O_2.

Agreement between CAC scores of the two observers in both CCTA and CSCT was excellent. For CAC volume and mass in CCTA, the ICC between the two observers was 0.928 (0.891–0.952) and 0.949 (0.922–0.966), respectively. Bland–Altman analysis (Fig. 11a) had bias and limits of agreement 13.6 (−90.2 – 117.5) mm^3 and 8.4 (−53.3–70.0) for CAC volume and CAC mass, respectively. For CAC volume and mass in CSCT, the ICC between the two observers was 0.983 (0.974–0.988) and 0.982 (0.974–0.988). Bland–Altman analysis (Fig. 11b) had bias and limits of agreement −7.4 (−67.1–81.9) mm^3 and 2.2 (−20.3–24.7) for CAC volume and CAC mass, respectively.

Observer O_2 identified 45/48 patients with zero CAC score in CCTA as annotated by observer O_1, the reference. Conversely, observer O_2 identified 51/52 patients with a positive CAC score as annotated by observer O_1.

4.6. Comparison with previous methods

The proposed method was compared with results reported by other previously published algorithms on CAC scoring in CCTA. Table 4 lists, where available, the number of scans used for evaluation in each study, the evaluation criterion, Pearson's ρ correlation, ICC and/or Bland–Altman statistics between manual and reference scores, CVD risk categorization accuracy and lesion detection sensitivity, as well as the average number of FP errors per image. The listed results cannot be directly compared, as different data sets, different CAC quantification metrics, different correlation metrics and different CVD risk categories were used. For example, some studies specifically excluded patients with moderate or poor image quality, patients without CAC or patients with stents. All methods except the proposed method require coronary artery extraction for CAC detection.

Please cite this article as: J.M. Wolterink et al., Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks, Medical Image Analysis (2016), http://dx.doi.org/10.1016/j.media.2016.04.004

Table 4
Previously published results on automatic CAC scoring in CCTA. For each study, the following are reported: the number of scans in the test set (#Scans), the evaluated CAC quantification score (Score, AS = Agatston score [modified for CCTA], MS = mass score), Pearson correlation (Pearson ρ), intra-class correlation coefficient (ICC), Bland–Altman mean and limits of agreement in units or percentages (Bland–Altman), CVD risk categorization accuracy and weighted Cohen's κ (CVD risk), lesion identification sensitivity (Sens.) and false positive errors per scan (FP/scan) are listed.

	#Scans	Score	Pearson ρ	ICC	Bland–Altman	CVD risk	Sens., FP/scan
Schuhbaeck et al. (2015)	44	AS_{CCTA} vs. AS_{CSCT}	0.94	–	−56 (−518–407)	88.6% $\kappa = 0.87$	–
Ahmed et al. (2014)	100	AS_{CCTA} vs. AS_{CSCT}	0.949	0.863	−3 (−174–168) %	76.0% $\kappa = 0.588$	–
Eilot and Goldenberg (2014)	263	AS_{CCTA} vs. AS_{CSCT}	0.95/0.91	–	−1 (−80–78) %	82.7%	0.94, 0.9
Teßmann et al. (2011)	53	AS_{CCTA} vs. AS_{CCTA}	0.95	–	–	–	0.94, 0.9
Mittal et al. (2010)	165	–	–	–	–	–	0.70, 0.1
Wesarg et al. (2006)	10	–	–	–	–	–	1.00, –
Proposed method	100	MS_{CCTA} vs. MS_{CSCT}	0.950	0.944	−0.2 (−38.7–38.3)	83% $\kappa = 0.83$	0.72, 0.48

5. Discussion

A method for automatic coronary artery calcium scoring in coronary CT angiography employing convolutional neural networks has been presented. In contrast to previously proposed methods for CAC scoring in CCTA, our method does not require coronary artery extraction. Instead, CAC voxels are directly identified using pairs of ConvNets.

Automatically obtained as well as reference CAC volume scores in CCTA were lower than in CSCT. This is in accordance with previous studies (van der Bijl et al., 2010; Mylonas et al., 2014). However, automatically obtained CAC mass scores showed a strong correlation (ICC 0.944 [0.918–0.962]) with reference CAC mass scores in CSCT. A comparison of CVD risk categories based on reference Agatston scores in CSCT and automatically obtained Agatston scores in CCTA showed excellent agreement, with 83% of patients assigned to their reference risk category ($\kappa = 0.83$). Hence, patients with a high reference Agatston score in CSCT also had a high automatically determined Agatston score in CCTA. In addition, discrimination between patients with zero CAC and patients with a positive CAC score was good. Large scale studies have shown that patients with zero CAC have an excellent prognosis (Joshi et al., 2012), underlining the clinical relevance of this distinction.

False positive errors were caused by high intensity voxels in the coronary artery lumen. Lumen attenuation may be very different among patients, thereby posing a challenge for accurate CAC segmentation. In our test set, CAC detection thresholds determined based on the attenuation of the ascending aorta ranged from 333 to 930 HU. The method made some false positive errors, for example calcified lymph nodes, that are less likely to occur in methods using coronary artery extraction. Nevertheless, although our method does not use coronary artery extraction, it is likely that the ConvNet implicitly learns a representation of tubular structures. Calcifications in the coronary arteries were identified, but calcifications in the aorta, with similar intensity and shape characteristics but a different context, were not a source of false positive errors.

A VOI containing the heart was determined using a ConvNet-based bounding box extraction algorithm. Although this VOI primarily contained cardiac structures, non-cardiac structures such as ribs (see Fig. 6) were occasionally partially included due to the rectangular nature of the identified VOIs. Alternatively, segmentations which more closely follow the boundaries of the heart may be obtained using for example graph cuts (Funka-Lea et al., 2006), morphological operations (Kurkure et al., 2010) or atlas-based methods (Zhuang et al., 2015). The results showed that for the current application a 3D rectangular bounding box was sufficient. Moreover, this bounding box was successfully acquired in all cases, with an average processing time of 6.8 ± 0.5 s, compared to 13.2 m reported by Zhuang et al. (2015). Results presented by de Vos et al. (2016) illustrate that the method tightly follows a predefined standard anatomical VOI. Finally, the method only required retraining with manually defined 3D bounding boxes in 50 CCTA images.

Similarly to our previous work, *x, y, z*-coordinates were used to describe the location of each candidate in the image (Wolterink et al., 2015b). While in our previous work these features were crucial for accurate scoring, we found that here they only moderately affected the performance of the method. The reason for this may be two-fold. First, the patches provided sufficient texture information for voxel classification. Second, the VOI was sufficiently limited in size that location features did not provide much additional information. Nevertheless, *x, y, z*-coordinates would likely be valuable to provide artery-specific CAC scores, potentially leading to better prediction of CVD events (Brown et al., 2008). It is likely that the proposed method could straightforwardly be extended to such a multi-class analysis. However, this would increase the complexity of the method and it would therefore require a larger training set size than available in the presented work.

The purely convolutional network architecture used in this study allows training with patches and testing with whole images. To improve efficiency during ConvNet training, Long et al. (2015) proposed end-to-end training with whole images, i.e. by minimizing the difference between a predicted and a reference label map for a whole image instead of a single sample. However, the extreme imbalance in our classification problem necessitates balanced sampling of negative and CAC candidates during training. In whole image training, a sampling mask should therefore be applied to the predicted and reference label maps, which would reduce the overlap between patches and the potential increase in efficiency. Furthermore, considering whole images and not patches as samples would reduce the number of possible training batches. Therefore, in this study ConvNets were trained with mini-batches containing balanced samples from random training images.

The proposed network architecture (Fig. 3) does not include any pooling layers, hence no spatial invariance is introduced. Therefore, the ConvNets were trained to predict a label only for the voxel at the center of the odd-sized input patch and not for other voxels in the patch that might have a different label. The experiments showed a benefit of larger input patches in terms of specificity, with input patch sizes of 15 and 25 voxels corresponding to receptive fields of 6.75 mm vs. 11.25 mm along each axis. Note that the typical diameter of a coronary artery is 4 mm (Dodge et al., 1992). Therefore, a larger patch size provides a wider margin around a coronary artery, which likely allows reduction of FP errors. In this work, only two input sizes were evaluated. As shown by the experiments, smaller inputs would not be likely to provide better results. Larger inputs might provide better results, but this effect might be mitigated by the increase in the number of trainable parameters and the limited number of training samples. Hence, a multi-scale approach using a combination of small high resolution input patches to provide detailed local analysis and larger low resolution

Please cite this article as: J.M. Wolterink et al., Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks, Medical Image Analysis (2016), http://dx.doi.org/10.1016/j.media.2016.04.004

input patches to provide spatial context might further improve the method, while keeping the number of trainable parameters low.

Our method used either 2.5D or 3D input. We did not use individual 2D planar inputs. Although these are highly efficient, they fail to capture the volumetric aspect of the data. Volumetric 3D patches can provide more information, but their size may pose computational challenges. In our experiments, 3D testing took substantially more time than processing with 2.5D, on average 147 s vs. 52 s for ConvNet_1 and 201 s vs. 107 s for ConvNet_2. Similarly, training with 3D input took much longer. In addition, with a limited number of training samples, as is often the case in medical image analysis, 3D ConvNets are more likely to overfit to the training data. Our experiments showed a performance drop between 2.5D and 3D architectures in terms of CAC lesion identification in CCTA. It is likely that the number of trainable parameters and input voxels for the 3D patch causes the network to overfit, considering that the number of parameters was high compared to the number of positive training samples. In other applications in medical image analysis, e.g. (Prasoon et al., 2013), 2.5D input outperformed 3D input as in our experiments. As a potential means to overcome the limitations of 3D input patches, yet capture volumetric information, Zheng et al. (2015) proposed separable 3D kernels with a reduced number of trainable parameters. Setio et al. (2016) extended 2.5D input by using 9 rotated 2D views for ConvNet-based lung nodule classification, and hypothesized that, to some extent, more 2D views may lead to better performance. However, this advantage may be problem-specific and not applicable to voxel classification. In future work, we will further investigate the trade-off between complexity and performance of different input representations for voxel classification. In addition, the current data set could be enlarged to provide a more diverse set of training samples.

It has previously been shown in 2D natural image classification that ensembles of ConvNet models can outperform individual models, e.g. by Krizhevsky et al. (2012). In our experiments, ensembles of ConvPairs with different dimensionality or input size improved lesion identification in all cases. It is likely that these models captured different aspects of the data, and hence were prone to make different errors. In future work, we will investigate to what extent the combination of models with identical architectures improves results, compared to models with different architectures that were combined in the present study.

A clinical standard for CAC scoring in CCTA is lacking, and intensity thresholds have been determined in various ways (Glodny et al., 2009; Otton et al., 2012; Mylonas et al., 2014; Pavitt et al., 2014). The proposed method was trained using manual annotations in CCTA, based on a patient-specific threshold of $mean_{aorta} + 3SD_{aorta}$ determined with a ROI in the ascending aorta. As a consequence, in several cases, high variability in lumen HU values in the CCTA image caused oversegmentation of the CAC lesion. In the current work, we did not correct for this oversegmentation in the reference. A comparison between automatically determined and manually determined CAC volume and mass scores in CCTA showed that the automatic method generally determined lower scores. In (van der Bijl et al., 2010), calcium was annotated fully manually, by contouring of calcified lesions, an extremely time-consuming process. The performance of our method is likely to increase by using such annotations, i.e. by removing noisy labels from the training data set.

In large studies, CAC scores in CSCT have been shown to be highly predictive of cardiovascular events (Yeboah et al., 2012). For CAC scores in CCTA, such studies have as of yet not been performed. Agatston and volume scores in CCTA are typically much lower in CCTA than in CSCT (van der Bijl et al., 2010). CSCT images are reconstructed with 3.0 mm slice spacing, while CCTA images typically have sub-millimeter slice spacing. Therefore, due to partial volume effects high-density CAC appears to have a larger volume in CSCT than in CCTA, but a higher peak intensity in CCTA than in CSCT. The Agatston score uses a stepwise function, which assigns the same weight to all voxels above 400 HU, thus not weighing higher peak intensities in CCTA accordingly. In previous studies, Agatston scores and CAC volume scores in CCTA were converted to modified Agatston scores using empirically determined linear conversion factors (Mylonas et al., 2014; Schuhbaeck et al., 2015). In this study, we primarily evaluated our method using the mass score, which weighs intensities linearly, and therefore better captures differences between CSCT and CCTA. The mass score has previously been shown to yield high correlations between CAC quantification in CSCT and CCTA (Hong et al., 2002) and to have lower inter-scan variability than the volume and Agatston score (Hoffmann et al., 2006). In addition, we did compute unmodified Agatston scores in CCTA to assign patients to CVD risk categories. Although these scores were lower than Agatston scores in CCTA, they ranked patients almost equally, indicating that CAC quantification in CCTA may be used to assign patients to CVD risk categories, without the application of a conversion factor.

ConvNets are pattern recognition networks, which means that they learn from examples and hence, are not likely to correctly classify unfamiliar samples. An example of this was found in our test set, which contained large calcified lymph nodes in the pericardium. It is likely that the number of scans in the training set was too small to capture all the variability expected in coronary CT angiography scans. Therefore, in future work a larger set of images covering a larger variety of examples will be included.

The presented method performed very fast voxel classification. Bounding box extraction took on average 7 s per patient. Average processing time for the best performing ConvPair was 46 s for ConvNet_1 and 28 s for ConvNet_2. Because our method is fully automatic, it shows potential for application in large-scale studies, as well as in a clinical settings where immediate processing would allow for a smooth workflow. To the best of our knowledge, the relation between CAC scores in CCTA and clinical CVD outcome and/or CVD risk categorization is not yet known. The presented automatic method allowing quick analysis might allow such studies.

In conclusion, CAC can be accurately automatically identified and quantified in CCTA using the proposed pattern recognition method. This might obviate the need to acquire a dedicated CSCT scan for CAC scoring, which is regularly acquired prior to a CCTA, and thus reduce the CT radiation dose received by patients.

Acknowledgments

This study was financially supported by the project FSCAD, funded by the Netherlands Organisation for Health Research and Development (ZonMw) in the framework of the research programme IMDI (Innovative Medical Devices Initiative); project 104003009.

Dr. Leiner is a recipient of a ZonMw Clinical Fellowship (2011-40-00703-98-11432).

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

Agatston, A.S., Janowitz, W.R., Hildner, F.J., Zusmer, N.R., Viamonte, M., Detrano, R., 1990. Quantification of coronary artery calcium using ultrafast computed tomography. J. Am. Coll. Cardiol. 15 (4), 827–832.

Ahmed, W., de Graaf, M.A., Broersen, A., Kitslaar, P.H., Oost, E., Dijkstra, J., Bax, J.J., Reiber, J.H., Scholte, A.J., 2014. Automatic detection and quantification of the Agatston coronary artery calcium score on contrast computed tomography angiography. Int. J. Cardiovasc. Imag. 31 (1), 151–161.

Al-Mallah, M.H., Aljizeeri, A., Alharthi, M., Alsaileek, A., 2014. Routine low-radiation–dose coronary computed tomography angiography. Eur. Heart J. Suppl. 16 (suppl B), B12–B16.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I. J., Bergeron, A., Bouchard, N., Bengio, Y., 2012. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

van der Bijl, N., Joemai, R.M., Geleijns, J., Bax, J.J., Schuijf, J.D., de Roos, A., Kroft, L.J., 2010. Assessment of Agatston coronary artery calcium score using contrast-enhanced CT coronary angiography. AJR Am. J. Roentgenol. 195 (6), 1299–1305.

Brown, E.R., Kronmal, R.A., Bluemke, D.A., Guerci, A.D., Carr, J.J., Goldin, J., Detrano, R., 2008. Coronary calcium coverage score: determination, correlates, and predictive accuracy in the multi-ethnic study of atherosclerosis. Radiology 247 (3), 669–675.

Ciompi, F., de Hoop, B., van Riel, S.J., Chung, K., Scholten, E.T., Oudkerk, M., de Jong, P.A., Prokop, M., van Ginneken, B., 2015. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. Med. Imag. Anal. 26 (1), 195–202.

de Vos, B.D., Wolterink, J.M., De Jong, P.A., Viergever, M.A., Isgum, I., 2016. 2D image classification for 3D anatomy localization; employing deep convolutional neural networks. In: SPIE Medical Imaging, 9784. International Society for Optics and Photonics, p. 97841Y.

Ding, X., Slomka, P.J., Diaz-Zamudio, M., Germano, G., Berman, D.S., Terzopoulos, D., Dey, D., 2015. Automated coronary artery calcium scoring from non-contrast ct using a patient-specific algorithm. In: SPIE Medical Imaging. International Society for Optics and Photonics, p. 94132U.

Dodge, J., Brown, B.G., Bolson, E.L., Dodge, H.T., 1992. Lumen diameter of normal human coronary arteries. influence of age, sex, anatomic variation, and left ventricular hypertrophy or dilation.. Circulation 86 (1), 232–246.

Eilot, D., Goldenberg, R., 2014. Fully automatic model-based calcium segmentation and scoring in coronary CT angiography. Int. J. Comput. Assist. Radiol. Surg. 9 (4), 595–608.

Farabet, C., Couprie, C., Najman, L., LeCun, Y., 2012. Scene parsing with multiscale feature learning, purity trees, and optimal covers. In: Proceedings of the International Conference on Machine Learning (ICML'12).

Funka-Lea, G., Boykov, Y., Florin, C., Jolly, M.-P., Moreau-Gobard, R., Ramaraj, R., Rinck, D., 2006. Automatic heart isolation for CT coronary visualization using graph-cuts. In: IEEE International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 614–617.

Glodny, B., Helmel, B., Trieb, T., Schenk, C., Taferner, B., Unterholzner, V., Strasak, A., Petersen, J., 2009. A method for calcium quantification by means of CT coronary angiography using 64-multidetector CT: very high correlation with Agatston and volume scores. Eur. Radiol. 19 (7), 1661–1668.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, pp. 249–256.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 15, pp. 315–323.

Hecht, H.S., 2015. Coronary artery calcium scanning: Past, present, and future. JACC: cardiovasc. Imag. 8 (5), 579–596.

Hoffmann, U., Siebert, U., Bull-Stewart, A., Achenbach, S., Ferencik, M., Moselewski, F., Brady, T.J., Massaro, J.M., ODonnell, C.J., 2006. Evidence for lower variability of coronary artery calcium mineral mass measurements by multi-detector computed tomography in a community-based cohort - consequences for progression studies. Eur. J. Radiol. 57 (3), 396–402.

Hong, C., Becker, C.R., Schoepf, U.J., Ohnesorge, B., Bruening, R., Reiser, M.F., 2002. Coronary artery calcium: absolute quantification in nonenhanced and contrast-enhanced multi–detector row CT studies. Radiology 223 (2), 474–480.

Isgum, I., Prokop, M., Niemeijer, M., Viergever, M.A., van Ginneken, B., 2012. Automatic coronary calcium scoring in low-dose chest computed tomography. IEEE Trans. Med. Imaging 31 (12), 2322–2334.

Išgum, I., Rutten, A., Prokop, M., van Ginneken, B., 2007. Detection of coronary calcifications from computed tomography scans for automated risk assessment of coronary artery disease. Medical physics 34 (4), 1450–1461.

Joshi, P.H., Blaha, M.J., Blumenthal, R.S., Blankstein, R., Nasir, K., 2012. What is the role of calcium scoring in the age of coronary computed tomographic angiography? J. Nucl. Cardiol. 19 (6), 1226–1235.

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

Kurkure, U., Chittajallu, D.R., Brunner, G., Le, Y.H., Kakadiaris, I.A., 2010. A supervised classification-based method for coronary calcium detection in non-contrast CT. Int. J. Cardiovasc. Imag. 26 (7), 817–828.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

McCollough, C.H., Ulzheimer, S., Halliburton, S.S., Shanneik, K., White, R.D., Kalender, W.A., 2007. Coronary artery calcium: A multi-institutional, multimanufacturer international standard for quantification at cardiac ct. Radiology 243 (2), 527–538.

Messenger, B., Li, D., Nasir, K., Carr, J.J., Blankstein, R., Budoff, M.J., 2015. Coronary calcium scans and radiation exposure in the multi-ethnic study of atherosclerosis. Int. J. Cardiovasc. Imag. 1–5.

Mittal, S., Zheng, Y., Georgescu, B., Vega-Higuera, F., Zhou, S.K., Meer, P., Comaniciu, D., 2010. Fast automatic detection of calcified coronary lesions in 3D cardiac CT images. In: Wang, F., Yan, P., Suzuki, K., Shen, D. (Eds.), Machine Learning in Medical Imaging. In: Lecture Notes in Computer Science, 6357. Springer Berlin Heidelberg, pp. 1–9.

Mylonas, I., Alam, M., Amily, N., Small, G., Chen, L., Yam, Y., Hibbert, B., Chow, B.J., 2014. Quantifying coronary artery calcification from a contrast-enhanced cardiac computed tomography angiography study. Eur. Heart J. Cardiovasc. Imag. 15 (2), 210–215.

Otton, J.M., Lønborg, J.T., Boshell, D., Feneley, M., Hayen, A., Sammel, N., Sesel, K., Bester, L., McCrohon, J., 2012. A method for coronary artery calcium scoring using contrast-enhanced computed tomography. J. Cardiovasc. Comput. Tomogr. 6 (1), 37–44.

Pavitt, C.W., Harron, K., Lindsay, A.C., Ray, R., Zielke, S., Gordon, D., Rubens, M.B., Padley, S.P., Nicol, E.D., 2014. Deriving coronary artery calcium scores from CT coronary angiography: a proposed algorithm for evaluating stable chest pain. Int. J. Cardiovasc. Imag. 30 (6), 1135–1143.

Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., Nielsen, M., 2013. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (Eds.), Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013. In: Lecture Notes in Computer Science, 8150. Springer Berlin Heidelberg, pp. 246–253.

Raff, G.L., Chinnaiyan, K.M., Cury, R.C., Garcia, M.T., Hecht, H.S., Hollander, J.E., O'Neil, B., Taylor, A.J., Hoffmann, U., 2014. SCCT guidelines on the use of coronary computed tomographic angiography for patients presenting with acute chest pain to the emergency department: a report of the Society of Cardiovascular Computed Tomography Guidelines Committee. J. Cardiovasc. Comput. Tomogr. 8 (4), 254–271.

Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M., 2014. A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014. In: Lecture Notes in Computer Science, 8673. Springer International Publishing, pp. 520–527.

Schaap, M., Metz, C.T., van Walsum, T., van der Giessen, A.G., Weustink, A.C., Mollet, N.R., Bauer, C., Bogunović, H., Castro, C., Deng, X., et al., 2009. Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms. Med. Imag. Anal. 13 (5), 701–714.

Schuhbaeck, A., Otaki, Y., Achenbach, S., Schneider, C., Slomka, P., Berman, D.S., Dey, D., 2015. Coronary calcium scoring from contrast coronary CT angiography using a semiautomated standardized method. J. Cardiovasc. Comput. Tomogr. 9 (5), 446–453.

Setio, A.A.A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., van Riel, S., Wille, M.W., Naqibullah, M., Sanchez, C., van Ginneken, B., 2016. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. IEEE Trans. Med. Imag. PP (99). doi:10.1109/TMI.2016.2536809.1–1

Shahzad, R., van Walsum, T., Schaap, M., Rossi, A., Klein, S., Weustink, A.C., de Feyter, P.J., van Vliet, L.J., Niessen, W.J., 2013. Vessel specific coronary artery calcium scoring: An automatic system. Acad. Radiol. 20 (1), 1–9.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Stollenga, M.F., Byeon, W., Liwicki, M., Schmidhuber, J., 2015. Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation. In: Advances in Neural Information Processing Systems, pp. 2980–2988.

Teßmann, M., Vega-Higuera, F., Bischoff, B., Hausleiter, J., Greiner, G., 2011. Automatic detection and quantification of coronary calcium on 3D CT angiography data. CSRDC 26 (1), 117–124.

Voros, S., Qian, Z., 2012. Agatston score tried and true: by contrast, can we quantify calcium on CTA? J. Cardiovasc. Comput. Tomogr. 6 (1), 45–47.

Wesarg, S., Khan, M.F., Firle, E.A., 2006. Localizing calcifications in cardiac CT data sets using a new vessel segmentation approach. J. Digit. Imaging. 19 (3), 249–257.

Wolterink, J.M., Leiner, T., Takx, R.A.P., Viergever, M.A., Isgum, I., 2015. Automatic coronary calcium scoring in non-contrast-enhanced ECG-triggered cardiac CT with ambiguity detection. IEEE Trans. Med. Imag. 34 (9), 1867–1878.

Wolterink, J.M., Leiner, T., Viergever, M.A., Isgum, I., 2015. Automatic coronary calcium scoring in cardiac CT angiography using convolutional neural networks. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. In: Lecture Notes in Computer Science, 9349. Springer International Publishing, pp. 589–596.

Wolterink, J. M., Leiner, T., de Vos, B. D., Coatrieux, J.-L., Kelm, B. M., Kondo, S., Salgado, R. A., Shahzad, R., Shu, H., Snoeren, M., Takx, R. A. P., van Vliet, L. J., van Walsum, T., Willems, T. P., Yang, G., Zheng, Y., Viergever, M. A., Išgum, I., An evaluation of automatic coronary artery calcium scoring methods with cardiac CT using the orCaScore framework. Medical Phys. doi:10.1118/1.4945696

Yeboah, J., McClelland, R.L., Polonsky, T.S., Burke, G.L., Sibley, C.T., OLeary, D., Carr, J.J., Goff, D.C., Greenland, P., Herrington, D.M., 2012. Comparison of novel risk markers for improvement in cardiovascular risk assessment in intermediate-risk individuals. JAMA 308 (8), 788–795.

Zheng, Y., Liu, D., Georgescu, B., Nguyen, H., Comaniciu, D., 2015. 3D deep learning for efficient and robust landmark detection in volumetric data. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. In: Lecture Notes in Computer Science, 9349. Springer International Publishing, pp. 565–572.

Zhuang, X., Bai, W., Song, J., Zhan, S., Qian, X., Shi, W., Lian, Y., Rueckert, D., 2015. Multiatlas whole heart segmentation of CT data using conditional entropy for atlas ranking and selection. Medical Phys. 42 (7), 3822–3833.

Please cite this article as: J.M. Wolterink et al., Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks, Medical Image Analysis (2016), http://dx.doi.org/10.1016/j.media.2016.04.004