COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 138 (2017) 49–56

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Classification of CT brain images based on deep learning networks

Xiaohong W. Gao [a,*], *Rui Hui* [b], *Zengmin Tian* [b]

[a] *Department of Computer Science, Middlesex University, London NW4 4BT, UK*
[b] *Neurosurgery Centre, Navy General Hospital, Beijing, China*

ARTICLE INFO

Article history:
Received 25 May 2016
Received in revised form
14 September 2016
Accepted 15 October 2016

Keywords:
Deep learning
Convolutional neural network
Classification
CT brain images
3D CNN
KAZE

ABSTRACT

While computerised tomography (CT) may have been the first imaging tool to study human brain, it has not yet been implemented into clinical decision making process for diagnosis of Alzheimer's disease (AD). On the other hand, with the nature of being prevalent, inexpensive and non-invasive, CT does present diagnostic features of AD to a great extent. This study explores the significance and impact on the application of the burgeoning deep learning techniques to the task of classification of CT brain images, in particular utilising convolutional neural network (CNN), aiming at providing supplementary information for the early diagnosis of Alzheimer's disease. Towards this end, three categories of CT images (N = 285) are clustered into three groups, which are AD, lesion (e.g. tumour) and normal ageing. In addition, considering the characteristics of this collection with larger thickness along the direction of depth (z) (~3–5 mm), an advanced CNN architecture is established integrating both 2D and 3D CNN networks. The fusion of the two CNN networks is subsequently coordinated based on the average of Softmax scores obtained from both networks consolidating 2D images along spatial axial directions and 3D segmented blocks respectively. As a result, the classification accuracy rates rendered by this elaborated CNN architecture are 85.2%, 80% and 95.3% for classes of AD, lesion and normal respectively with an average of 87.6%. Additionally, this improved CNN network appears to outperform the others when in comparison with 2D version only of CNN network as well as a number of state of the art hand-crafted approaches. As a result, these approaches deliver accuracy rates in percentage of 86.3, 85.6 ± 1.10, 86.3 ± 1.04, 85.2 ± 1.60, 83.1 ± 0.35 for 2D CNN, 2D SIFT, 2D KAZE, 3D SIFT and 3D KAZE respectively. The two major contributions of the paper constitute a new 3-D approach while applying deep learning technique to extract signature information rooted in both 2D slices and 3D blocks of CT images and an elaborated hand-crated approach of 3D KAZE.

1. Introduction

1.1. CT brain images

Today, in the UK, 800,000 people have been formally diagnosed with the condition of dementia [1]. In reality, it is estimated that 60% of people who are living with the condition go undiagnosed [2]. This is because the determination of dementia remains a convoluted process as symptoms come and go. As a result, more comprehensive data remain in need to provide complementary information. Due to its prevalent, non-invasive and inexpensive nature, computerised tomography (CT) is in service in nearly every hospital while presenting

* *Corresponding author.* Department of Computer Science, Middlesex University, London NW4 4BT, UK. Fax: +44 (0) 20 8411 2252.
E-mail address: x.gao@mdx.ac.uk (X.W. Gao).

http://dx.doi.org/10.1016/j.cmpb.2016.10.007
0169-2607/

good quality of visual information of human organs. In addition, CT remains probably the first imaging tool that was introduced into the study of human brain and has since been widely applied as the first choice to eliminate other possibilities when it comes to the diagnosis of Alzheimer's disease (AD).

Although most patients have undertaken this scanning as a prelude to imaging inspection, mainly for the purpose of ruling out the other alternatives (e.g., tumour, stroke, etc.), CT data have not been implemented into the clinical diagnosis of AD due to their relatively low resolution and variations among manual measurement of certain features, such as medial temporal lobe that is associated with AD. Moreover, with regard to CT brain images, specified brain atrophy is associated with not only AD but also normal ageing and cerebral vascular diseases. For example, the medial temporal lobe atrophy (MTA) together with cerebrospinal fluid (CSF) biomarkers has been demonstrated as the most important diagnostic markers for AD, which however may not be specific. In addition, the atrophy of hippocampus (in particular, left hippocampus) that has been found in AD also emerges in healthy ageing adults [3]. But in spite of these concerns, recently, it has been found that CT data can contribute significantly to the diagnosis of AD with 90.2% accuracy rate [3] through accurate measurement of atrophy factors between temporal horn ratio and suprasellar cistern ratio. Therefore, CT linear measurements can be of great value in the work-up processes of AD patients to a large extent.

To alleviate the considerable variations [4] that may incur during manual measurements, this study is to investigate a non-supervised automatic process employing the state of the art of convolutional neural network (CNN) to the classification of CT data between AD, healthy (normal) ageing and lesion brain data.

1.2. *Convolutional neural network (CNN)*

Deep learning models refer to a class of computing machines that can learn a hierarchy of features by building high-level attributes from low-level ones [5,6], thereby automating the process of feature construction. One of these models is the well-known convolutional neural network (CNN) [6]. Consisted of a set of algorithms in machine learning, CNN comprises several (deep) layers of processing involving learnable operators (both linear and non-linear), and hence has the ability to learn and build high-level information from low-level features in an automatic fashion [7]. Stemming from biological vision processes, a CNN applies a feed-forward artificial neural network to simulate variations of multilayer perceptrons whereby the individual neurons are tiled in such a way that they respond to overlapping regions in the visual field [8]. As a direct result, these networks are widely applied to image and video recognition. Specifically, CNNs have demonstrated as an effective class of models for understanding image content, proffering state of the art results on image recognition, segmentation, detection and retrieval. In other words, when trained with appropriate regularisation, CNNs can achieve superior performance on visual object recognition tasks without relying on any hand-crafted features, e.g. SIFT, SURF. In addition, CNNs have shown to be relatively insensitive to certain variations on the inputs [9]. Significantly, recent advances of computer hardware technology (e.g., graphics processing unit (GPU)) have propitiated the implementation of CNNs in representing images.

Table 1 – Notations of parameters in Eq. (2).

Parameter	Notation
tanh(.)	Hyperbolic tangent function
m	Index over the set of feature maps in the (i – 1)th layer
b_{ij}	Bias for the feature map f in Eq. (1)
w_{ijk}^{pq}	Value at the position (p, q) of the kernel connected to the kth feature map
(p,q)	2D position of a kernel
P_i, Q_j	Height and width of the kernel

Theoretically, CNN can be expressed in the following formulas. For example, for a set of training data $\left(x^{(i)}, y^{(i)}\right)$, where image $x^{(i)}$ is in three-dimension (inclusive of RGB channel as the 3rd dimension) and $y^{(i)}$ the indicator vector of affiliated class of $x^{(i)}$, the feature maps of an image, namely, $w_1, \ldots, w_L$ will be learnt based on CNN by solving Eq. (1).

$$argmin_{w_{1,\ldots,w_L}} \frac{1}{n}\sum_{i=1}^{n} \ell\left(f\left(x^i; w_1, \ldots, w_L\right), y^i\right) \tag{1}$$

where ℓ refers to a suitable loss function (e.g. the hinge or log loss) and f the selected classifier.

To obtain these feature maps computationally, in a 2D CNN, the operator of convolution is conducted at each convolutional layer to extract features from local neighbourhood on the feature maps acquired in the previous layer. Then an additive bias is applied and the result is passed through a sigmoid function as formulated in Eq. (2) mathematically in order to obtain a newly calculated feature value v_{ij}^{xy} at position (x,y) on the j_{th} feature map in the i_{th} layer.

$$v_{ij}^{xy} = tanh\left(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)}\right) \tag{2}$$

where the notations of those parameters in Eq. (2) are explained in Table 1.

As a result, CNN architecture can be constructed by stacking multiple layers of convolution and subsampling in an alternating fashion. The parameters of CNN, such as the bias b_{ij} and the kernel weight w_{ijk}^{pq} are trained using unsupervised approaches [10,11].

In the same way, the 3D convolution is achieved by convolving a 3D kernel to a block or box along both x–y (2D) and z directions where Eq. (2) is extended into Eq. (3) to calculate the value at position (x,y,z) on the j_{th} feature map in the i_{th} layer.

$$v_{ij}^{xyz} = tanh\left(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \tag{3}$$

where R_i indicates the size of the 3D kernel along the z dimension, w_{ijm}^{pqr} is the $(p, q, r)_{th}$ the value of the kernel connected to the m_{th} feature map in the previous layer.

While CNNs have lent themselves well to the computer vision field and achieved state-of-the-art results, they are built

mainly for 2D images. Although several papers report the work on 2D videos [12,13] considering time as the 3rd dimension, working on 3D still images is quite a different task to a certain extent [14]. In this study, both 2D and 3D forms of CNN are elaborated in details for the application of classification of CT brain images with detailed implementation presented in Section 2.

This paper is structured in the followings. Section 2 entails the methodologies that are employed in this study completed with detailed implementation of CNN. In addition, Section 3 puts forward the results as well as comparisons with a number of hand-crafted techniques whereas the conclusion and future recommendations are summarised in Section 4.

2. Methodology

2.1. Data pre-processing and averaging 3D CT images

Conformed to the patient informed consent, a total of 285 datasets of 3D are collected from Navy General Hospital, China, which compose 57, 115 and 113 data respectively in the category of Alzheimer's (AD), lesion and normal. In this collection, CT data vary in resolution in both depth between 16 and 33 slices and dimension with either 512×512 or 912×912 pixels. Fig. 1 depicts the montage of three datasets associated with Alzheimer's (top), lesion (middle) and normal (bottom) respectively.

In 2D form, all slices are cropped and normalised into 200×200 pixels to exclude information tags on each slice. For normal and AD data, the middle 20 slices are employed for the processing, whereas for lesion data, only slices that contain visual lesion features (e.g. tumour) are employed. As a result, although lesion datasets amount to the largest (N = 115) among three classes, the overall numbers of slices are similar to the other two.

In parallel, in 3D form, firstly each dataset undergoes processes of registration, segmentation and normalisation to arrive at a resolution of $200 \times 200 \times 20$ pixels. Then, because of their relatively thickness in-between CT slices (~3–5 mm) in comparison with their counterparts of MR images (0.5 mm), spatial normalisation is performed to align all 3D CT images into the same space and opts for the approach of rigid body geometric transformation [14]. After normalisation, each 3D dataset is divided into $40 \times 40 \times 10$ boxes. Similar to 2D form, for both AD and normal data, all the boxes/blocks are applied to train the CNN network whereas for lesion data, only blocks that contain lesion contents are employed.

2.2. Implementation of 2D and 3D CNNs

In this study, the implementation of 2D convolutional deep learning neural network (CNN) takes shape of MatConvNet [11] written in Matlab software, along the axial direction of the brain. Fig. 2 schematically illustrates the deep learning network architecture employed in this study integrating both 2D and 3D CNN networks. Each network consists of seven layers whereas each layer embodies a number operators (liner or nonlinear), mainly convolution and sub-sampling. The Conv operator usually computes the output of neurons to be re-connected to local regions in the input, by producing a dot product between their weights and a small region they are supposed to link to in the input volume. For example, the input size of $11 \times 11 \times 96$ in layer 1 (i.e. Conv-1) at Fig. 2 top graph indicates the neighbourhood filter (F) size, i.e. weight, being 11×11 whereas 96 bank of filters are chosen to apply. In the subsampling as illustrated in Fig. 2, the resolution of the feature maps is reduced by pooling over local neighbourhood on the feature maps in the previous layer, thereby increasing invariance to distortions on the inputs. The down size rate is controlled by pooling stride (P-S) and is set to be 2 at layer 1, i.e. the data size has been halved after pooling stage.

$$y = f(x) = f_L(\ldots, f_2(f_1(x, w_1), w_2), \ldots), w_L) \tag{4}$$

Specifically, in each layer, to learn jointly, both forward and backward processing are staged composed of several operators in an end-to-end manner. As such, a forward neural network tends to be the composition of a number of functions as formulated in Eq. (4) [11].

Each function f_l takes a datum x_l as input that has a size of M × N pixels × K channels (default K being 3 representing R, G, and B colour channels) and a parameter vector w_l, then produces an output datum x_{l+1}. The very first input of $x = x_1$ indicates a CT image that is to be processed whereas the rest of x_l (l>1) are intermediate feature maps. For each convolutional layer, the initial input filter bank of w_i is randomly generated but with pre-defined filter sizes. For example, in Fig. 2 top graph, for Conv-1, the filter size is set as $11 \times 11 \times 3$, generating 96 filter banks. The output of the convolution with this bank of filters, y, is assessed in Eq. (5).

$$y_{i'j'k'} = sum_{ijk}\left(w_{ijkk'} x_{i+i',j+j',k}\right) \tag{5}$$

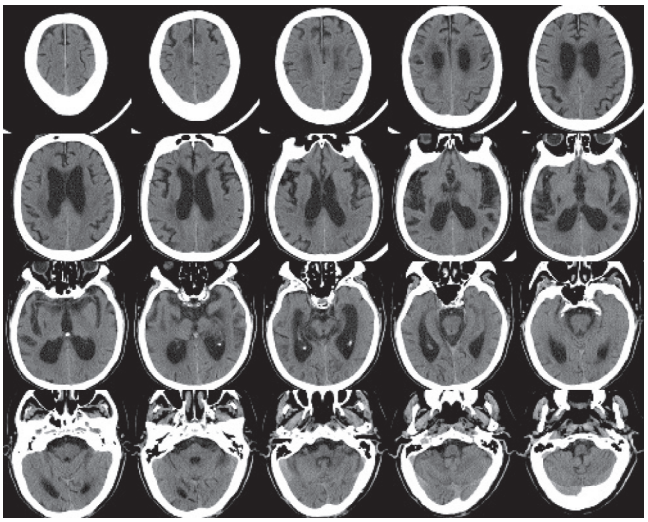
where $k' = 96$, $k = 3$, $i' = 11$, and $j' = 11$ for the first Conv layer. In other words, each convolutional operator generates K dimensional map of y by Eq. (5). For example, for block 1, if $x_0 = (200,200,3)$ with the original image size, then feature map has a size of $x_1 = (48,48,96)$ after layer-1 convolutional operator. The calculation of the size of feature maps follows the rule that is set out in Eq. (6).

$$x_{i+1}^{size} = \left(\frac{x_i^{size} - F_i + 2 \times Pad}{Stride} + 1\right) \tag{6}$$

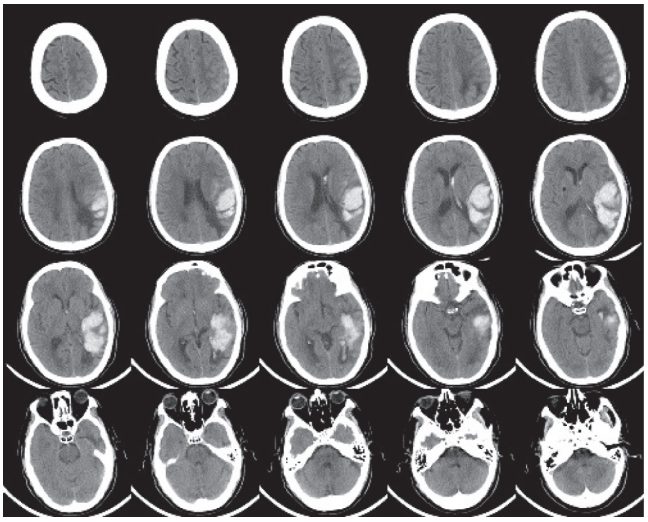
For example, in Fig. 2 top graph layer 1, the parameters are set to be $F_0 = (11 \times 11)$ along both x and y directions, $Pad = 0$, $Stride = 4$, leading to the size of x_1 being $48 \times 48 \times 96$ (i.e., $48 = (200 - 11)/4 + 1$). Since the images are in grey level, the 3rd channel representing RGB colours is ignored at this paper.

Additionally, each component or pixel of a feature map is subject to a non-linear gating process to legitimise the processed data. In doing so, the simplest approach of rectified linear unit (ReLU) is applied as conveyed in Eq. (7) that thresholds the data with zero.

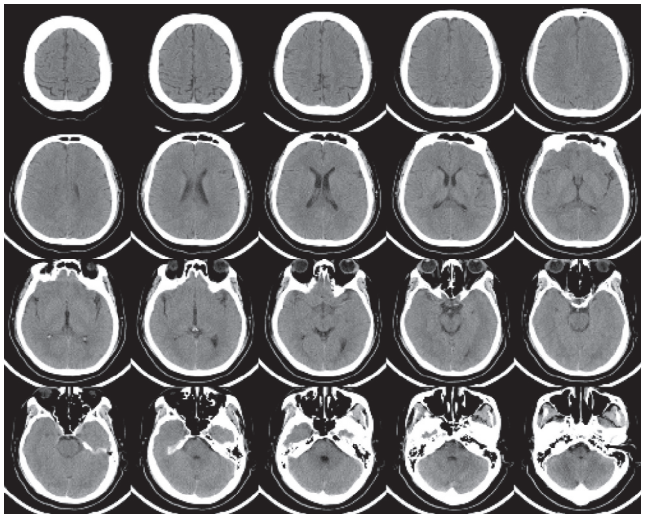
$$y_{ijk} = max(0, x_{ijk}) \tag{7}$$



(a) Alzheimer's



(b) Lesion



(c) Normal

Fig. 1 – The montage view of the middle 20 slices of CT images from subjects with Alzheimer's disease (top), lesion (middle), and of normal (bottom) respectively.

This operator however does not change the size of each feature map.

To down size the feature map, pooling is employed to coalesce nearby feature values into one downsided samplings and reduce the influence of noise while operating on each individual feature channel. The most commonly used choice of pooling remains to be max-pooling to select the largest component within a neighbourhood as manifested in Eq. (8).

$$y_{ijk} = max\{y_{i'j'k} : i \le i' < i+p, j \le j' < j+q\} \tag{8}$$

whereas the downsize rate is controlled by pooling stride (P-S). As a result, the output size of layer 1 in Fig. 2 top graph has a dimension of (24,24,96) where $24 = \frac{48}{\texttt{Pool\_Stride}}$ that is computed using Eq. (6) where $\texttt{Pool\_Stride} = 2$.

Another important operator is dropout to deal with overfitting in the CNN networks. As such, randomly dropout units (along with their connections) from the neural network during training stage are selected and discarded. The dropout rate in this study is set to be 0.5, i.e. half of the data units.

Once each layer of forward processing is completed, backward process proceeds to ensure that the parameters of feature maps of $w = (w_1, \dots, w_L)$ are learned in such a way that the overall function of $z = f(x, w)$ sustains a minimum loss, $\ell(z, \hat{z})$, where $z = (z_1, \dots, z_i, \dots)$ corresponds with the output value of x_i and $\hat{z}_i$ the ground truth of x_i in the training datasets. Therefore the loss function can be determined in Eq. (9).

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(z_i, f(x_i, w)) \tag{9}$$

There exists a number of algorithms to minimise L. In this research, the approach of *gradient descent* is employed which quantifies the gradient of L at a current solution w^t and then updates the latter along the direction of fastest descent of L as revealed in Eq. (10).

$$w^{t+1} = w^t - \eta_t \frac{\partial f}{\partial w}(w^t) \tag{10}$$

where η_t refers to the learning rate that is usually pre-defined and is within the range of (0, 1). In this way, parameters of w can be solved using training datasets.

Most importantly, while filter sizes can be of any size within the limit of data sizes, they are chosen manually in advance. In addition, the dimension of the output layer at the end of CNN architecture must be $1 \times 1 \times 3$, which reduces the full input image into a single vector of class scores (in our case, class number is 3), arranged along the depth dimension and can be computed using Eq. (6) together with the values of pooling stride. For example, the output sizes of all 7 layers of 2D CNN laid out in Fig. 2 entail sizes of (24,24,96), (6,6,256), (6,6,384), (6,6,256), (1,1,192), (1,1,96) and (1,1,3) respectively when the input sizes are of (200,200).

In 3D CNN, the implementation of each block contains two operators of convolution and pooling respectively as depicted in Fig. 2 bottom graph. Analogous to 2D CNN, 3D max-pooling operator is implemented by taking a block with $k \times m \times n$ dimension as an input and yielding a single value

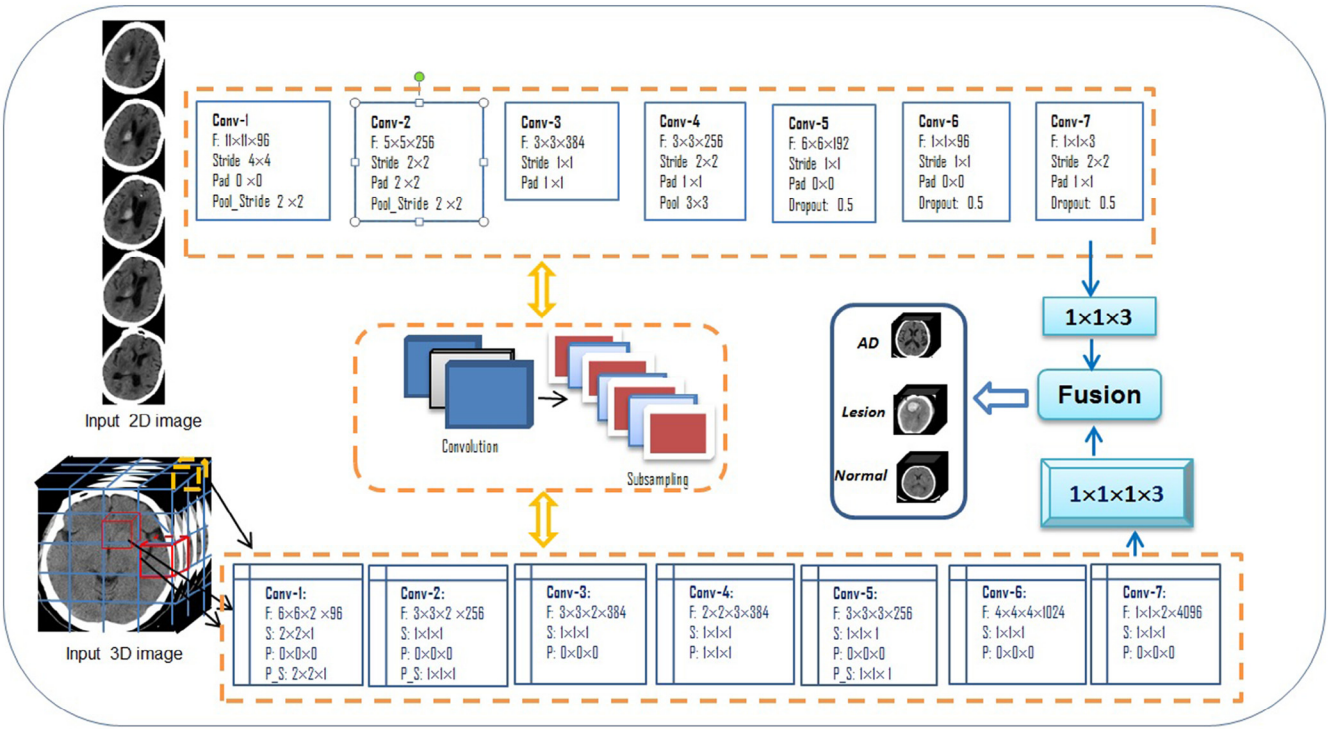


Fig. 2 – The fusion of both 2D and 3D CNNs for CT images.

that holds the maximum of the block. Similarly, given the input size of (40,40,10), the output sizes of the seven layers are of (9,9,9,96), (7,7,8,256), (5,5,7,384), (6,6,7,256), (4,4,5,1024), (1,1,2,4096), and (1,1,1,3) respectively.

2.3. *Classification and fusion*

In 2D form, image slices are applied to train the 2D CNN model, whereas in 3D form, small cubes ($40 \times 40 \times 10$) are utilised. The classification at each network is performed using Softmax classifier [15] that determines a score of normalised class probabilities. Eq. (11) defines mathematically the Softmax function.

$$f_i(z) = \frac{e^{z_j}}{\sum_k e^{z_k}} \tag{11}$$

where the function takes a vector of arbitrary real-valued scores (in z) and compresses it to a vector of values between zero and one that sum to one. The obtaining of the class scores f involves the calculation of cross-entropy loss that is formulated in Eq. (12).

$$L_i = -f_{y_i} + \log \sum_j e^{f_j} \tag{12}$$

where the notation f_j refers to the j_{th} element of the vector of class scores f [15].

After the establishment of class scores at each of 2D and 3D networks individually, the fusion takes place linearly. For the category of normal subject, a datum is classified as normal only if more than 95% of the slices of its dataset are labelled as *Normal* as well as all cubes (or boxes) have to be labelled as *Normal*, whereas for classes of AD and lesion, the majority voting ascertain the final classification scores.

2.4. *Classification based on hand-crafted features of 3D SIFT*

In this research, comparison with two hand-crafted approaches for feature detection is also conducted, including scale invariant feature transform (SIFT) [16] and KAZE [17].

As demonstrated in Fig. 3, the implementation of 3D SIFT engages in the process of SIFT feature detection through the application of Difference of Gaussian (DoG) operators and sparse coding [18,19], codebook generation by the employment of the paradigm of Bag of visual Words (BoW) [20,21], image representation using the created codebook and finally classification based on support vector machines (SVM).

2.5. *Implementation of 3D KAZE features for classification*

While SIFT can extract features that maintain scale invariant, it lacks sophistication in locating boundary and structural details, which is particularly important for medical images where precise delineation of a region is in need most of the time [22]. This has led to the investigation of another hand-crafted approach of KAZE technique. By deciding to apply a nonlinear scale space through the consolidation of nonlinear diffusion filtering [23], KAZE circumvents the shortcomings that SIFT presents.

To extend this technique to 3D, in this study, the detection of KAZE features engages in the processes of 3D Gaussian smoothness, calculation of conductivity, creation of nonlinear

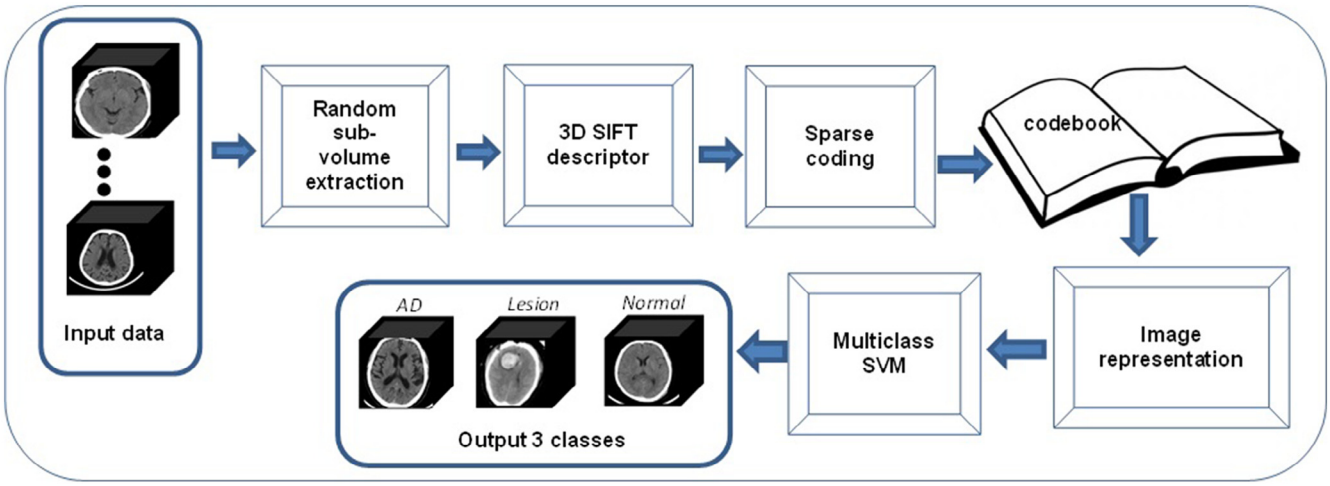


Fig. 3 – The flowchart of classification of CT images applying 3D SIFT features.

scale spaces, extraction of features and finally coarse-to-fine suppression. Consequently, the gradients of feature points can be detected by the application of either Eq. (13) or Eq. (14) or both at two different levels.

$$g_1(|\nabla(x,y,z)|) = \exp\left(-\left(\frac{|\nabla(x,y,z)|}{K}\right)^2\right) \quad (13)$$

$$g_2(|\nabla(x,y,z)|) = \frac{1}{1+\left(\frac{|\nabla(x,y,z)|}{K}\right)^2} \quad (14)$$

where K indicates the contrast parameter to control the smooth level, which can be determined automatically based on the intensity distribution levels of CT images, ∇ the gradient operator and $|.|$ the absolute value.

3. Results

Based on the formulae entailed above, both experimental and evaluation results are attained. Table 2 provides detailed numbers of images that are utilised in each category in this evaluation. In total, there are 285 datasets in 3D form consisted of 4076 2D slices. They are divided into 2 sub-groups while applying CNN approach, which are training and testing respectively.

For training, 180 datasets are assigned and selected randomly containing 2700 in 2D and 3795 in 3D, whereas the remaining 105 datasets are reserved as test sets. Consequently, the confusion matrix for the testing data is given in Table 3 that presents the accuracy rates of classification for the three classes, which gives 85.2%, 80.0%, and 95.3% for Alzheimer's, lesion, and normal respectively, with an average of 87.62%.

Evaluation with the two hand-crafted techniques of 3D KAZE and 3D SIFT also takes place with confusion matrixes given in Tables 4 and 5 respectively. In summary, the average accuracy rate is 83.2% for 3D KAZE and 85.2% for 3D SIFT. When applying these two approaches, one-verse-all strategy is appointed during the training/testing stage due to the concern of small number of datasets. In this way, one data set (N = 1) is randomly selected as a test datum whereas the rest (N = 284) remains as training sets throughout the whole database. The data obtained in both Tables 4 and 5are the averaged results

Table 2 – The data sets (in subject number) for both training and testing applied in each network. The numbers in brackets are the total number of 2D slices and 3D blocks in that category.

	Alzheimer's	Lesion	Normal	Total
Training	30	80	70	180
2D frame	700	700	1300	2700
3D boxes	775	1160	1860	3795
Test	27	35	43	105
2D frames	300	247	829	1376
3D boxes	675	840	1161	2676
Total	57	115	113	285
2D frames	850	947	2129	4076
3D boxes	1450	2000	3021	6471

Table 3 – The confusion matrix of testing results of three clusters using improved CNN networks.

	Alzheimer's	Lesion	Normal	Accuracy rate (%)
Alzheimer's	23	4	0	85.2
Lesion	3	28	4	80.0
Normal	2	0	41	95.3
Average				87.62

Table 4 – The confusion matrix for 3D KAZE approach.

	Alzheimer's	Lesion	Normal	Accuracy rate (%)
Alzheimer's	44	5	8	77.2
Lesion	8	93	14	80.9
Normal	8	5	100	88.5
Average				83.15 ± 0.35

Table 5 – The confusion matrix for 3D SIFT approach.

	Alzheimer's	Lesion	Normal	Accuracy rate (%)
Alzheimer's	36	6	15	63.2
Lesion	4	102	9	88.7
Normal	3	6	105	92.9
Average				85.26 ± 1.60

Table 6 – Comparison between CNN and hand-crafted approaches.

Methods	Accuracy rate (%)
2D SIFT	85.61 ± 1.10
2D KAZE	86.31 ± 1.04
3D SIFT	85.26 ± 1.60
3D KAZE	83.15 ± 0.35
2D CNN	86.32
3D CNN	87.62

of 10 rounds of training-test processes. Subsequently, the standard deviations of these 10 rounds for KAZE and SIFT lie within the range of ±0.35% and ±1.60% respectively.

While each CT dataset is of 3D form, the depth between every 2 slices along the depth-direction is relatively large (~3.0–5.0 mm). Hence it is reasonable to accept that CT images can also be considered on a 2D basis. Accordingly, both 2D versions of SIFT and KAZE for feature detection are evaluated and compared. Towards this end, Fisher vector [24] coupled with VLFeat library [25] is applied to represent features. The final classification score for each subject is a linear combination of scores of histograms for all the slices. As discussed above, for the normal class, each subject is clustered as *normal* unless all but one slice are labelled as *Normal* (i.e. ≥95%). Table 6 presents the comparison results and indicates that the 2D forms of hand-crafted approaches of both SIFT and KAZE appear to outperform their 3D counterparts, especially for KAZE with a margin of 3% differences, which could be explained away by the fact that this group of CT data do have substantial low depth resolution. In addition, the proposed synergy of 2D/3D CNN approach appears to have achieved the best result so far with 87.6% accuracy whereas both 2D CNN and 2D KAZE deliver similar performance standing in second place. Considering the fact that CNN is renowned for performing better with larger datasets whereas this study has a small disposal of samples (N = 285), the good performance that CNN based approaches have confirmed the potential that deep learning techniques possess for classification of CT images. It is envisaged that more datasets will be collected in the future to take the findings forward in order to benefit a wider communities, including patients, clinicians and academia.

4. Conclusion and discussion

This research concerns with the classification of CT images into three categories of Alzheimer's, lesion, and normal through the application and elaboration of a deep learning neural network, CNN. Although the category of lesion comprises the largest number of the dataset (N = 115), not every 2D slice or 3D block contains lesion signature information, e.g. tumour. As a result, the lesion group has the smallest number of images with 947 slices, whilst AD and normal groups hold 1000 and 2129 images respectively in 2D form. Although the differences in number may not be significant, in particular between AD and lesion groups, the classification results appear to be in line with the number of data that each group has. For example, the normal subject group that dominates the datasets with a total of 2129 image slices produces the highest accuracy rate of 95.3%. Similar trend also appears in the application of 3D KAZE and 3D SIFT approaches. Therefore the direct conclusion remains that more data will achieve better classification results. In addition, the results given by KAZE tend to be more robust with less variations than by SIFT, especially 3D KAZE that sustains only ±0.35% standard deviations.

In addition, while CT brain data are in three dimensional, the larger thickness (~3–5 mm) between slices in this collection has led to the classification results in 3D form worse than in 2D form. The fusion therefore takes place to take the advantage of assimilation of both 2D and 3D networks while preserving differentiating characteristics of CT images in both forms. Similarly, it appears that the 3D versions of both hand-crafted approaches of SIFT and KAZE are relegated from their 2D form with worse performance. In the future, this phenomenon will be further explored with a collection of more CT data and higher spatial resolutions.

At present, although nearly every patient with suspected AD undertakes CT scan, mainly for ruling out the other explanations, CT data have not been included in the diagnosis process yet. With the encouragement of classification results obtained from deep learning techniques, it is anticipated that in the future, a learnable expert system will be in the pipeline to assist clinicians to corroborate Alzheimer's disease while integrating all the available data (e.g. memory test, lab test, CT images, etc.), leading to the improvement of current diagnosis rate of AD (~33%). Although CT images present a number of diagnostic features of AD as detailed in Ref. [26], for example, the neuro-pathologic changes in the temporal lobe, including focal atrophy of the subiculum and entorhinal cortex, manual quantification of these key measurements does introduce considerable errors, which can lead to decisions being problematic [4]. This research intends to shed light on the significant contribution that unsupervised deep learning techniques can make towards classification of CT data, which in the next stage will be followed by the measurement of those AD diagnostic features precisely using CNN, allowing CT images to measure up to their potentials and benefit patients, caregivers and society as a whole.

Acknowledgements

This work constitutes part of project WIDTH that is funded by European Union Seventh Framework Programme under Marie Curie Scheme (IRSES, No 269124). Their financial support is gratefully acknowledged.

REFERENCES

[1] BBC. The dementia timebomb. http://www.bbc.co.uk/science/0/21878238, Retrieved: 31 March 2016.
[2] UK Government. Dementia. https://www.gov.uk/government/news/, Retrieved: 31 March 2016.
[3] Y. Zhang, E. Londos, L. Minthon, C. Wattmo, H. Liu, L.O. Wahlund, Usefulness of computed tomography linear measurements in diagnosing Alzheimer's disease, Acta Radiol. 49 (1) (2008) 91–97.
[4] A.R. Oksengaard, M. Haakonsen, R. Dullerud, K. Engedal, K. Laake, Accuracy of ct scan measurements of the medial temporal lobe in routine dementia diagnostics, Int. J. Geriatr. Psychiatry 18 (2003) 308–312.
[5] K. Fukushima, Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cyb. 36 (1980) 193–202.
[6] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, IEEE 86 (11) (1998) 2278–2324.
[7] Y. LeCun, F.J. Huang, L. Bottou Learning methods for generic object recognition with invariance to pose and lighting. Proceedings of Computer Vision and Pattern Recognition (CVPR), 2:II97–104, 2004.
[8] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[9] M. Ranzato, F.J. Huang, Y. Boureau, Y. LeCun Unsupervised learning of invariant feature hierarchies with applications to object recognition. Proceedings of Computer Vision and Pattern Recognition (CVPR), pages 1–8, 2007.
[10] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE 35 (1) (2015) 221–231.
[11] A. Vedaldi, K. Lenc Matconvnet convolutional neural networks for matlab. Proceedings of the 23rd ACM international conference on Multimedia, pages 689–692, 2015.
[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, F. Li Large-scale video classification with convolutional neural networks. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014), pages 1725–1732, 2014.
[13] X. Gao Feature-wise representation for both still and motion 3d medical images. IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), pages 1–4, 2014.
[14] C. Jongen, J.P.W. Pluim, P.J. Nederkoorn, M.A. Viergever, W. Niessen, Construction and evaluation of an average ct brain image for inter-subject registration, Comput. Biol. Med. 34 (8) (2004) 647–662.
[15] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.
[16] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[17] P.F. Alcantarilla, A. Bartoli, A.J. Davison Kaze features. Eur. Conf. on Computer Vision (ECCV), pages 214–227, 2012.
[18] Y. Qian, R. Hui, X. Gao, Retrieval of 3d surgical path based on sparse coding for image-guided neurosurgery, Signal Process. 93 (6) (2013) 1673–1683.
[19] J. Yang, K. Yu, Y. Gong, T. Huang Linear spatial pyramid matching using sparse coding for image classification. IEEE Conference on Computer Vision and Pattern Recognition, page 1794–1801, 2009.
[20] J. Sivic, A. Zisserman Kaze features. Video Google: A Text Retrieval Approach to Object Matching in Videos, pages 1470–1477, 2003.
[21] Y. Qian, L. Wang, C. Wang, X. Gao, The synergy of 3d sift and sparse codes for classification of viewpoints from echocardiogram videos, in: H. Greenspan, H. Muller, T. Syeda-Mahmood (Eds.), Medical Content-Based Retrieval for Clinical Decision Support, Springer, New York, 2012, pp. 68–79.
[22] W. Li, Y. Qian, M. Loomes, X. Gao The application of kaze feature to the classification of echocardiogram videos. MRMD 2015, LNCS, 9059:61–72, 2015.
[23] F. Catte, P.L. Lions, J.M. Morel, T. Coll, Image selective smoothing and edge detection by nonlinear diffusion, SIAM J. Numer. Anal. 29 (1992) 182–193.
[24] F. Perronnin, J. Sanchez, T. Mensink Improving the fisher kernel for large-scale image classification. ECCV 2010 (Part IV), 6314:143–156, 2010.
[25] A. Vedaldi, B. Fulkerson Vlfeat: An open and portable library of computer vision algorithms. Proceedings of the 18th annual ACM international conference on Multimedia, 2010.
[26] A.E. George, M.J. de Leon, L.A. Stylopoulos, J. Miller, A. Kluger, G. Smith, et al., Ct diagnostic features of Alzheimer disease: importance of the choroidal/hippocampal fissure complex, AJNR Am. J. Neuroradiol. 11 (1) (1990) 101–107.