Adapting content-based image retrieval techniques for the semantic annotation of medical images

Ashnil Kumar[a,c,*], Shane Dyer[b], Jinman Kim[a,c], Changyang Li[a,c], Philip H. W. Leong[b,c], Michael Fulham[a,c,d,e], Dagan Feng[a,c,f]

[a]*School of Information Technologies, University of Sydney, Australia*
[b]*School of Electrical and Information Engineering, University of Sydney, Australia*
[c]*Institute of Biomedical Engineering and Technology, University of Sydney, Australia*
[d]*Department of Molecular Imaging, Royal Prince Alfred Hospital, Sydney, Australia*
[e]*Sydney Medical School, University of Sydney, Australia*
[f]*Med-X Research Institue, Shanghai Jiao Tong University, China*

Abstract

The automatic annotation of medical images is a prerequisite for building comprehensive semantic archives that can be used to enhance evidence-based diagnosis, physician education, and biomedical research. Annotation also has important applications in the automatic generation of structured radiology reports. Much of the prior research work has focused on annotating images with properties such as the modality of the image, or the biological system or body region being imaged. However, many challenges remain for the annotation of high-level semantic content in medical images (e.g., presence of calcification, vessel obstruction, etc.) due to the difficulty in discovering relationships and associations between low-level image features and high-level semantic concepts. This difficulty is further compounded by the lack of labelled training data. In this paper, we present a method for the automatic semantic annotation of medical images that leverages techniques from content-based image retrieval (CBIR). CBIR is a well-established image search technology that uses quantifiable low-level image features to represent the high-level semantic content depicted in those images. Our method extends CBIR techniques to identify or retrieve a collection of labelled images that have similar low-level features and then uses this collection to determine the best high-level semantic annotations. We demonstrate our annotation method using retrieval via weighted nearest-neighbour retrieval and multi-class classification to show that our approach is viable regardless of the underlying retrieval strategy. We experimentally compared our method with several well-established baseline techniques (classification and regression) and showed that our method achieved the highest accuracy in the annotation of liver computed tomography (CT) images.

Keywords:
image annotation, content-based image retrieval, computed tomography, liver, ImageCLEF

1. Introduction

1.1. Motivation and Aims

Medical imaging is a fundamental component of modern healthcare with roles in patient diagnosis, treatment planning, and assessment of response to therapy. A direct consequence of this is the rise in medical imaging informatics research, including content-based image retrieval [1, 2], modality-classification and case-based retrieval [3], classification [4, 5], and annotation [5–7]. Semantic image annotation is also emerging as a research question, in which the main research challenge is to detect subtle differences in low-level image features and to relate them to higher-level labels derived from a standard terminology. Ultimately the goal is to apply the annotation technologies for the automatic generation of structured imaging reports [8, 9].

Annotation is also considered to be a prerequisite for semantic medical search engines that enable radiologists to find medical images, reports, and associated publications more efficiently [7]. Automatic semantic annotation is needed because it is difficult, time-consuming and expensive to manually annotate the rich contents of these items. The annotation and image markup use case of the caBIG project [10], which described a software library that could be used for the annotation of large collections of images, provides an example of the ponderous nature of manual annotation processes. Wennerberg et al. [7] improved the efficiency of this manual annotation process using an ontology modularisation tool that identifies and ranks fragments of an ontology that are relevant to the annotation task; this relevance is based upon the specific domain (e.g., lymphoma) and hierarchical relationships of terms already an-

[*]Corresponding author. Address: School of Information Technology Building J12, University of Sydney, NSW 2006, Australia; Tel.: +61 2 9036 9805; Fax: +61 2 9351 3838

Email addresses: ashnil.kumar@sydney.edu.au (Ashnil Kumar),
sdye9175@uni.sydney.edu.au (Shane Dyer),
jinman.kim@sydney.edu.au (Jinman Kim),
changyang.li@sydney.edu.au (Changyang Li),
philip.leong@sydney.edu.au (Philip H. W. Leong),
michael.fulham@sydney.edu.au (Michael Fulham),
dagan.feng@sydney.edu.au (Dagan Feng)

notated. However, these manual annotation approaches require physicians to subjectively determine the labels that are relevant to a particular image based on the physicians' expertise and prior experience.

In contrast, automatic image annotation is conducted on the basis of quantifiable image features. The combination of features present in each image suggests the annotations that are relevant. Many existing approaches described in the summary paper by Deselaers et al. [11] only annotated the images with the properties of the image, such as the image modality, body orientation, body region and biological system being examined. Setia et al. [5] extracted local feature descriptors from the most salient (interesting) points on each image to capture the geometric relationships present in the image; a hierarchical classification method was used to annotate each image by the image properties listed earlier. In a similar application, Tommasi et al. [6] proposed a method that extracted global and local features using three classification strategies that emphasised feature fusion at different stages of the annotation process. Ko et al. [12] presented a method that utilised a random forest classifier together with a predefined body relation graph to identify and annotate the body region shown in the image.

A more difficult objective is to annotate the images with clinically relevant *content*, such as the presence of calcification, mass effect, etc. In the general (i.e., non-medical) domain, image annotation tasks have moved rapidly from object identification to sentence generation, where the aim is to describe the images through words, in the same way in which a human witness might describe a scene that they have observed; several such methods have been described in a recent summary paper [13]. Kulkarni et al. [14] used computer vision based object detection to construct a graph of the objects and labeled the graph based upon statistics mined from large corpora of descriptive text; the labels and graph relationships could then be used to generate descriptive sentences.

One of the major hurdles in achieving this objective for medical images is that there are likely to be thousands of semantic labels to learn and often very few labeled training samples [15]. Thus a major challenge of such research is the development of categorisation and annotation techniques that are less hindered by lack of training samples [16]. To reduce problems caused by lack of training data, Gimenez et al. [17] avoided classification methods and instead annotated liver CT images using logistic regression, through the least absolute shrinkage and selection operator (LASSO). However, their method only annotated binary semantic outcomes that could be presented by positive or negative observations, e.g., whether or not a lesion was homogeneous. In a follow-up study, Depeursinge et al. [18] learned semantic terms describing the visual appearance of liver lesions derived from a linear combination of multi-scale wavelet features. This allowed their method to model each annotation at the the most relevant image scale. The method predicted the probability that a particular semantic description (e.g., irregular lesion margin) was applicable to the lesion in the image but did not annotate the effects on anatomical structures, e.g., the proximity of the lesion to the hepatic vasculature.

The recognition of image content also falls within the scope of another important area of medical imaging informatics research called content-based image retrieval (CBIR) [1]. In CBIR, low-level visual features such as intensity, texture, shape, and the spatial arrangement of objects are used to determine which images are similar to a given query [19]. A key challenge for CBIR is the *semantic gap*, which is the difference between machine-computed similarity and a human's interpretation of similarity [19]. Many different CBIR algorithms have been investigated for this purpose; a summary can be found in the recent review by Kumar et al. [2]. Well-established CBIR techniques are therefore designed to relate low-level image features to higher-level semantic concepts. We hypothesise that the problem of automatic semantic image annotation could be addressed in a related fashion, by adapting the ability of CBIR techniques to leverage low-level image features in the search for images with similar high-level semantic concepts.

Thus in this paper, we present a method for the automatic annotation of medical images that is derived from CBIR techniques. Given an image to annotate, we propose to identify or retrieve a collection of semantically similar images that have already been labelled and use this collection to determine the best semantic annotations for the unlabelled image. Our annotation method is designed for limited training data compared to the number of annotations that need to be automatically recognised. We suggest that the technique would be applicable regardless of the underlying retrieval strategy and therefore we describe two ways of identifying the best annotations: either through multi-class classification and nearest-neighbour search, both of which are well-established CBIR methods. We evaluated our work on the annotation of liver CT images by comparing our annotation method to several other well-established techniques. We also compared our method to the state-of-the-art techniques submitted to the Imaging track of the Conference and Labs of the Evaluation Forum (ImageCLEF) [20] Liver Annotation Challenge [21]; the outcomes were reported at the CLEF workshop [22]. In this paper we expand upon the report by including: (i) detailed definitions of the classification and nearest neighbour methods for annotation, and (ii) a more comprehensive evaluation, which includes comparison with well-established techniques that were not submitted to the ImageCLEF Liver Annotation Challenge.

1.2. Terminology and Notation

We employ the following terminology in the remainder of this paper. A *question* refers to a specific annotation task, i.e., an element of the structured report that needs to be automatically filled. A *label* is an annotation that could possibly be assigned to a question. An *answer* is the label that our method automatically assigns to the question based on the analysis of the image features; the answer is chosen from a set of labels that are unique to each question. The term *query* refers to a single un-annotated image volume that will be annotated using our approach.

We also use the following notation. Let Ω be a question and $\mathcal{L}_\Omega$ be the set of labels for Ω with $|\mathcal{L}_\Omega| = l$. During annotation, we also let $\mathcal{L}^+_\Omega \subseteq \mathcal{L}_\Omega$ denote a possible set of answers (needed only in case of ties) and $L \in \mathcal{L}^+_\Omega$ denotes the final answer. Note that since only one label is chosen as the final answer (i.e., $|L| = 1$) then $L = \mathcal{L}^+_\Omega \Leftrightarrow \left|\mathcal{L}^+_\Omega\right| = 1$ (there were no ties).

2. Materials and Methods

2.1. Dataset

We used a public dataset of volumetric (3D) computed tomography (CT) images of the liver from the ImageCLEF 2014 Liver Annotation Challenge [21]. The dataset contained 50 CT volumes cropped to the region around the liver; the volumes had varied resolutions (x: 190–308 pixels, y: 213–387 pixels, slices: 41–588) and pixel spacings (x, y: 0.674–1.007mm, slice: 0.399–2.5mm). A mask of the liver pixels and the bounding box for a selected lesion were provided for each image.

The data also included a set of 60 well-established image features (with a total dimensionality of 458) that had been extracted from the images in the dataset. We refer to these as computer generated (CoG) features and they are further described in Section 2.3.

The dataset also contained 73 ground truth annotations. These annotations were independently determined by an experienced clinician based upon the semantic labels in the ONtology of Liver for Radiology (ONLIRA) [23]. The ontology described the different sets of possible labels $\mathcal{L}_\Omega$ for different questions Ω. Several label sets have one or two special labels with the following meaning:

- **other**: none of the other labels in $\mathcal{L}_\Omega$ fully answer the question.
- **N/A**: the question Ω is not relevant to this image.

Only 65 of the 73 annotations were relevant for the task [21]; questions with unbounded labels (e.g., measurements) were not included as part of the evaluation. There were 145 unique labels among these 65 questions; the median diversity was 2 possible labels and the maximum diversity was 10 labels.

2.2. Method Overview

Our aim was to derive the annotations of a query based upon similarity to other images. We adapted two state-of-the-art approaches, classification of image similarity using support vector machines (SVMs) [24] and weighted nearest-neighbour (WNN) search [25] to show that our method is applicable regardless of the underlying retrieval strategy.

Multi-class SVM classification was used because SVMs are effective in high-dimensional spaces where the dimensionality is higher than the sample size. SVMs are also versatile as they can use different kernel functions for different classification tasks. These characteristics of SVMs make them appropriate for our particular dataset, in which there were a wide variety of annotation labels with a small number of training samples. The WNN method was used because it uses the most similar samples for labelling, which is important when the distribution of samples is not known. Thus, it gives higher annotation priority to more similar samples than to those (dissimilar samples) that are further away. This was appropriate for our dataset, where different annotations had varying distributions of labels. In addition, WNNs are also useful in cases where there are no clear class boundaries, such as between related terms with very subtle differences.

Table 1: Visual Features

Partition	Feature Groups	Dimensionality
Liver	- Intensity - Size	3
Lesion	- Intensity - Location - Shape - Size - Texture	353
Vessel	- Size	2
All Lesions	- Counts - Intensity - Size - Texture (Haar only)	88
Global	SIFT BoVF	1000

We applied similar image annotation processes for both methods. Visual features were first extracted from the images. The classification or similarity model was then trained using the extracted visual features; feature selection was also performed in this phase. After the model is developed, new queries can then be annotated. This is done by comparing the features extracted from these images to the model as described in Sections 2.4 and 2.5.

2.3. Image Features

Table 1 summarises the features that were used in our experiments. These features are detailed in the following subsections.

2.3.1. Dataset Features

The dataset contained computed generated (CoG) image features that had already been extracted from the liver, the hepatic vasculature, and the marked (primary) lesion. Image features were also extracted from all lesions in the image and accumulated into a global feature value (e.g., mean intensity). The features described 3D object shape properties (e.g., volume, surface area, sphericity, solidity, convexity, Hu shape invariants [26]), texture information (e.g., Haralick [27], Gabor [28], Tamura [29], Haar [30]), and pixel intensity information. The ImageCLEF 2014 task documentation [21] provides a detailed list of the CoG image features.

We cleaned the CoG feature data by removing feature dimensions with missing values (i.e., given a not-a-number or NaN value) or that were used to scale other features and had no variation across all samples. These feature dimensions were excluded from all samples:

1. *Group:* Lesion. *Feature:* Anatomical Location (5 dimensions). *Reason:* All dimensions were missing (a NaN value) for one of the images.

2. *Group:* Lesion. *Feature:* Hu Moments (3 dimensions). *Reason:* All dimensions were missing (a NaN value) for one of the images.
3. *Group:* Lesion. *Feature:* Histogram (only the first two dimensions). *Reason:* These features were the upper and lower bounds and were the same for all samples. They were not needed after normalisation.
4. *Group:* All Lesions. *Feature:* HistogramOfAllLesions (only the first two dimensions). *Reason:* These features were the upper and lower bounds and were the same for all samples. They were not needed after normalisation.

Removing feature dimensions with missing values allowed us to keep the same sample size (50) in contrast to removing the entire sample, which would have made our limited training dataset even smaller. The cleaned CoG feature set had a total dimensionality of 446.

2.3.2. Bag of Visual Features

We also created a bag of visual features (BoVF) representation of the image. This was derived from scale invariant feature transform (SIFT) descriptors [31] extracted from key points detected in the 2D slices of the CT images. The SIFT descriptors were extracted from all of the axial slices on an image. We randomly sampled 5% of the descriptors extracted from the training dataset. We generated a visual codebook by grouping the subsampled descriptors using k-means clustering with $k = 1000$. This value of k has been successfully used in other medical image retrieval projects on diverse imaging data, e.g., x-rays, magnetic resonance, CT, etc. [32]. The subsampled clustering process was much faster than clustering on the full set of descriptors and created codebooks of similar quality [33]. The visual code for each cluster was the cluster centroid, i.e., the mean of all descriptors within that cluster.

We then assigned a single visual code to every key point in an image. For each key point, we calculated the Euclidean distance between its descriptor and all of the visual codes (cluster centroids) in the codebook; the codebook entry with the lowest distance (i.e., most similar descriptor features) was assigned as the visual code for the key point. This process was repeated for all of the key points in all of the axial slices in an image. We then created a BoVF descriptor using a k-bin histogram representing the frequency of the visual codes in that image [32]. The visual codes from all of the axial slices were pooled into a single descriptor.

2.3.3. Feature Normalisation

The features we extracted had a variety of ranges. This difference in range would cause some feature dimensions to have a higher influence on distance computation (see Section 2.5) compared to others. We therefore normalised the features to the range $[0, 1]$ by linearly scaling them to a random variable with zero mean and unit variance and shifting the values so they were within the desired range. Let x be the value of a feature f, and μ_f and σ_f be the mean and standard deviation of f in the dataset. The normalised value $\tilde{x}$ of x was determined as follows:

$$\tilde{x} = \frac{\left(x - \mu_f\right)/3\sigma_f + 1}{2} \tag{1}$$

According to Aksoy and Haralick [34], this equation normalises 99% of the feature values to the range $[0, 1]$; normalised values lower than 0 were set to 0, while those higher than 1 were set to 1. This normalisation scheme was shown to be effective in prior image retrieval research [35].

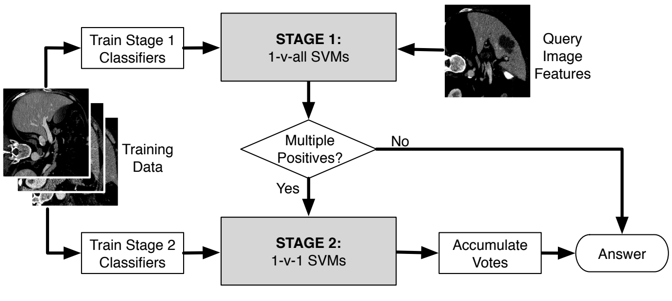

Figure 1: An overview of the classification scheme for annotation.

2.4. Retrieval Strategy: Two-Stage SVM Classification

SVMs [24] are a well-established classification technique with many applications in medical CBIR [36, 37]. SVMs are supervised learning models that can be used for binary classification. An SVM divides labelled training data into two categories and classifies new samples into one of these categories. Multi-class problems are usually solved with banks of multiple SVMs.

We adapted such an approach for our multi-class, multi-label annotation problem as shown in Figure 1. The core idea was to divide the problem into two stages for each question. The first stage identified a collection of labels that represented groups of images with similar image features ($\mathcal{L}_\Omega^+$). The second stage was used to evaluate each of the elements of this collection to select the best answer ($\mathcal{L}$).

Each stage in our classification approach comprised a bank of several SVM classifiers. For every label $A \in \mathcal{L}_\Omega$, we trained an A-vs-rest (1-vs-all) SVM classifier, hence forming l 1-vs-all SVMs. We also trained A-vs-B (1-vs-1) SVMs for every pair of labels $A, B \in \mathcal{L}_\Omega$ where $A \neq B$, forming a total of $\left(l^2 - l\right)/2$ 1-vs-1 SVMs. For every question, our first stage was composed of the 1-vs-all SVMs and the second stage was composed of the 1-vs-1 SVMs. This two stage approach was repeated separately for each question.

After training, we annotated the queries using the following procedure. We first extracted the same features from the query as described in Section 2.3. The query image was then classified using the first stage. If only one of the 1-vs-all SVMs returned a positive classification (i.e., there was no tie) then the label corresponding to that classifier was adopted as the answer. If the classifiers in the first stage assigned multiple labels (i.e., multiple 1-vs-all classifiers returned positive results) then the second stage was activated. The output of the first stage was the set of labels given a positive response by their associated

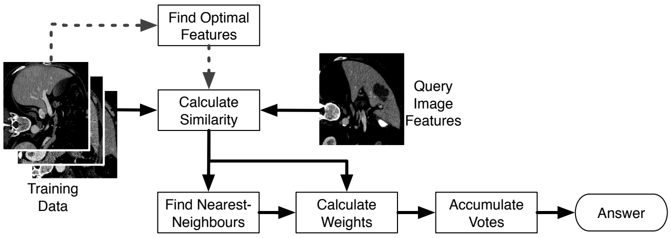

Figure 2: An overview of the WNN scheme for annotation. The dashed arrows indicate an optional feature selection process.

1-vs-all SVM, i.e., $\mathcal{L}^+_\Omega$. During the second stage, we classified the query using the 1-vs-1 classifiers for all the labels in $\mathcal{L}^+_\Omega$ (i.e., the 1-vs-1 classifiers for the tied labels). A majority voting scheme was used to select the answer.

Two tiebreaker situations remained after both classification stages. The ties included the case where the first stage did not return a positive label and when there was a tie in the vote during the second stage (multiple labels had the highest majority vote). In both of these situations, we set the answer to “other”. For such ties in questions Ω where “other” $\notin \mathcal{L}_\Omega$, we selected the label “N/A” if it was available or “false” for questions that expected a boolean answer.

Our design of the two stage classification scheme was due to the unbalanced training dataset. We expected that the classifiers for labels with few samples would have relatively low accuracy and thus the two stage approach introduced further discriminative power, especially in the case of ties.

2.5. Retrieval Strategy: Weighted Nearest-Neighbours

Our WNN approach for annotation used the most similar training images to select the answers for the query. The core idea was that the query would be annotated with the same label assigned to a collection of images with similar features. An overview of the method is shown in Figure 2. We created two variations of this approach: the first variation used the entire feature space to determine the nearest-neighbours while the second variation used forward sequential feature selection [38] to define a unique optimal feature space for each individual Ω. Thus, in the first variation, the process in Figure 2 was only performed once; in the second variation, the process was repeated for each Ω individually.

To locate the nearest-neighbours, we calculated the dissimilarity (s) of the features of each training image from the features of the query, using the Euclidean distance:

$$s(Q,T) = \sqrt{\sum_{i=0}^{d} (q_i - t_i)^2} \tag{2}$$

where Q was the feature vector of the query image (Q), T was the feature vector of a training image (T), q_i was the i-th dimension of Q, t_i was the i-th dimension of T, and d was the dimensionality of the feature set. Under this formulation, higher values of s indicated greater dissimilarity; $s(Q,T) = 0$ implied that Q and T were exactly similar.

The set of n images in the training set with the lowest values of s were chosen as the nearest neighbours because the most similar images are generally expected to be retrieved within the first few results (referred to as early precision) [32]. Let $S = \{s_1, ..., s_n\}$ be the dissimilarity values of these images sorted in ascending order. A weighted voting scheme was used to select the answer for each question using this set of dissimilarity values. The weighted vote v_i for the i-th most similar image was given by:

$$v_i = \frac{s_1 + \epsilon}{s_i + \epsilon} \tag{3}$$

where $s_i \in S$ was the dissimilarity value of the i-th most similar image and $\epsilon = 1.18 \times 10^{-38}$ was used to avoid divisions by zero.

The weighted vote V_A for a label $A \in \mathcal{L}_\Omega$ was given by:

$$V_A = \sum_{i=1}^{n} \lambda_{A,i} v_i \tag{4}$$

where

$$\lambda_{A,i} = \begin{cases} 1 & \text{if } A \text{ is the label of } T_i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

and T_i was the i-th nearest-neighbour (training image).

In the case of a tie (multiple labels having the same weighted vote), we set the answer to “other”. When “other” $\notin \mathcal{L}_\Omega$, we selected the label “N/A” if it was available or ”false” for questions that expected a boolean answer.

We accounted for the small training dataset by weighting the value using Equation 3. Neighbours with a higher similarity would thus have a stronger vote compared to images with a lower similarity. The weighting scheme ensured that the emphasis was placed on the labels of the neighbours that were closest to the query, even if there were a larger number of neighbours of a different label that were further away. This is in contrast to a majority voting scheme, where labels that had a higher frequency in the dataset would have a higher chance to be selected as the answer (depending on the value of n).

3. Evaluation

3.1. Experimental Procedure

We used 10-fold cross-validation to evaluate the accuracy of our annotation methods on Dataset One. We evaluated the following variations of our methods:

- Two-Stage SVM classification with the following kernels:
  - linear
  - polynomial
  - quadratic
  - radial basis function (RBF)
  - multilayer perceptron (MLP)
- WNN search
- WNN search with sequential feature selection

For WNN search, we repeated our experiments with cross-validation to find the optimal value of $n = 20$. We compared our methods to the following well-established techniques from the literature:

- One-Stage (standard) SVM classification with the following kernels:
  - linear
  - polynomial
  - quadratic
  - radial basis function (RBF)
  - multilayer perceptron (MLP)
- LASSO-based regression used in prior annotation research [17]

We calculated the accuracy of the annotation by comparing the answer given by each approach to the ground truth.

Each method was tested using three different feature sets:

- **Feature Set 1:** The CoG features after cleaning (*total dimensionality* = 446).
- **Feature Set 2:** SIFT BoVF (*total dimensionality* = 1000). The visual codebook (Section 2.3) was generated separately for each fold on the training data only so there was no bias towards the test data.
- **Feature Set 3:** Feature Set 1 + Feature Set 2 (*total dimensionality* = 1446).

3.2. Results

Table 2 summarises the annotation accuracy of our method (two-stage SVM with different kernels, nearest-neighbour search) in comparison with baseline methods (one-stage SVM and LASSO regression); the mean and standard deviation were calculated over all 10 folds. Figure 3 shows the mean accuracy of the baseline methods for individual questions and Figure 4 shows the same information for the proposed methods. The results demonstrate that WNN search has the highest mean accuracy and consistently high accuracy across most questions (48 questions with accuracy ≥ 90%). All variations of our proposed methods also have a higher accuracy than the LASSO method used in prior work. Note that the baseline LASSO method only converged for Feature Set 1 and only when using the unnormalised feature set and thus only the results for this situation is shown in Table 2 and Figure 3.

3.3. Outcomes of ImageCLEF Annotation Challenge

We also submitted our method to the ImageCLEF Liver Annotation challenge where it was independently evaluated and compared to other methods on a different test dataset. In 2014, our method outperformed all other methods, including generalised coupled tensor factorisation (GCTF) [39] and ensembles of different classifiers (Ensemble) [40]. Our method also scored higher than the single additional method submitted in 2015, which used a combination of Random Forests with shape, texture, and lesion features [41]. For more details, we refer interested readers to the papers cited in this section and the summary papers [21, 42].

3.4. Discussion

Table 2 show that the proposed WNN search achieved the highest annotation accuracy, outperforming the baseline and other proposed methods. These results for WNN search were consistent across all three feature sets. The high results achieved by the WNN search was due to the way in which the labels were prioritised; the labels of images that were more similar to the query were given more importance than those of images that were less similar, even if the latter collection was larger. A relatively simple approach like WNN retrieval was ideal as the annotation would be dependent on a subset of major discriminating image features. WNN also performs well in the situation where items of a class are clustered very closely together within the feature space even if there are no clear class boundaries, which can be the case when the differences are subtle or the labels are semantically related (e.g., "obliterated" vein lumen versus "partially obliterated").

In our WNN approach, we used a distance-based normalisation for scoring different labels (see Equation 3). The vote varies depending on the similarity of the image thereby allowing us to account for subtle differences between very similar images. A contrasting method is rank normalisation, in which the vote is dependent on the rank of the neighbouring image. However, weights that are based on rank alone cannot determine whether the two images are subtly or distinctly different. In contrast, a distance-based normalisation would have very similar weights for subtly different images and very different weights for distinctly different images.

The accuracy of the WNN method with feature selection was lower than that of WNN method with no feature selection. These results are counter-intuitive as the expectation was that feature selection would significantly improve the accuracy of the annotation. Similarly, the low accuracy of the LASSO method, which performs variable selection by shrinking parameter estimates (coefficients of the regression) closer to zero, was also unexpected. The lower accuracy suggests that there was less information for distinguishing between subtle cases. One explanation for this could be that the dataset contained very few subtle cases meaning there were very few samples from which to derive optimal features or regression coefficients. A more diverse training dataset would include more subtle cases and we suggest this would have a positive impact on the accuracy of methods that use some form of feature selection. At our current diversity (145 unique labels) WNNs with feature selection has approximately 1% lower accuracy than WNNs without feature selection; this is not a statistically significant difference (according to the Student's *t*-test) and using a dataset with an appropriate number of labels for the full diversity of liver annotations would improve the accuracy when using methods with feature selection.

We also discovered that our proposed two-stage SVM classification method had higher accuracies than one-stage SVMs

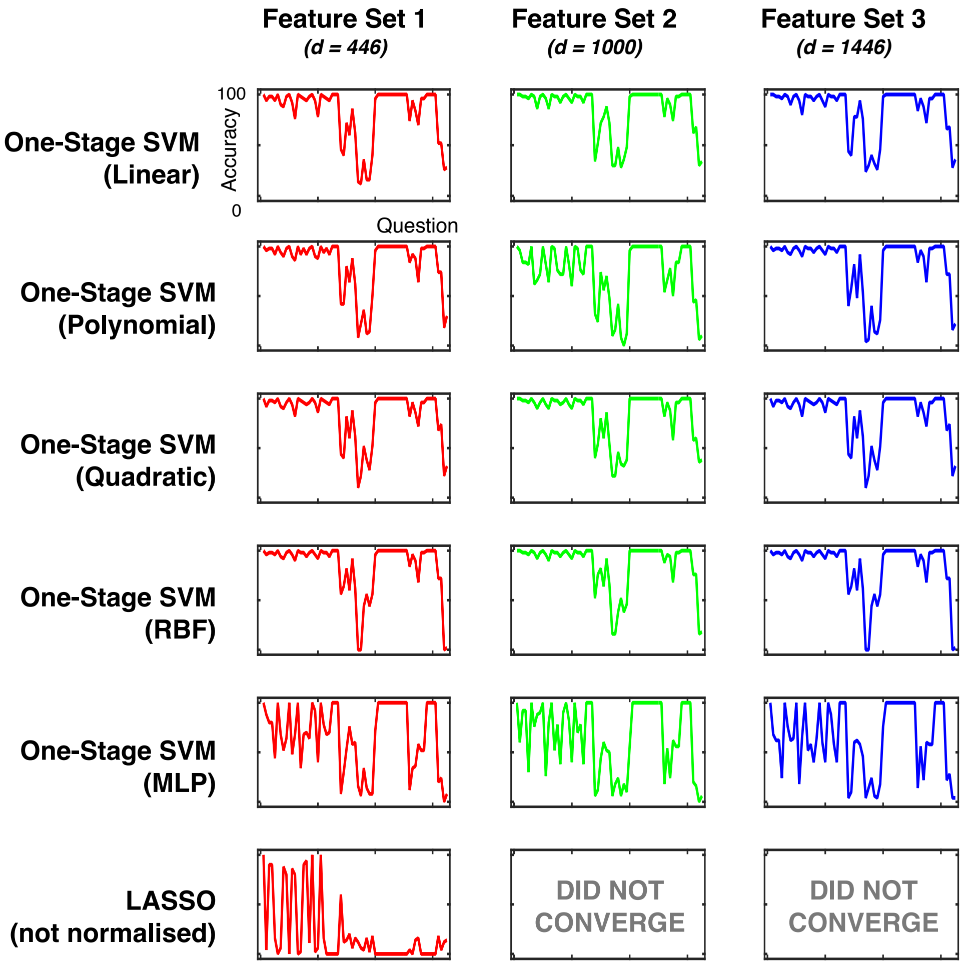

Figure 3: Accuracy of baseline annotation methods on different questions.

Table 2: Accuracy (%) of Different Methods (highest mean accuracy underlined)

Method		Feature Set					
		1: CoG		2: SIFT		3: CoG + SIFT	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
baseline	one-stage linear SVM	82.18	9.69	85.54	8.09	85.35	8.74
	one-stage polynomial SVM	83.38	8.77	73.97	11.97	83.20	7.78
	one-stage quadratic SVM	84.46	9.05	84.52	7.97	84.28	8.60
	one-stage RBF SVM	84.86	7.35	86.12	8.02	84.86	7.35
	one-stage MLP SVM	66.40	14.15	65.75	13.25	63.29	14.10
	LASSO (unnormalised)	25.07	4.37	-	-	-	-
proposed	two-stage linear SVM	84.00	9.96	86.98	8.12	86.49	8.89
	two-stage polynomial SVM	84.95	8.94	77.60	12.97	84.83	8.27
	two-stage quadratic SVM	85.72	9.18	85.91	8.04	85.94	8.79
	two-stage RBF SVM	85.75	7.82	87.26	8.03	85.75	7.82
	two-stage MLP SVM	70.55	15.05	69.78	14.55	67.38	15.48
	WNN	87.75	8.74	88.74	7.93	87.75	8.83
	WNN with feature selection	86.62	9.50	87.54	8.92	87.35	8.66

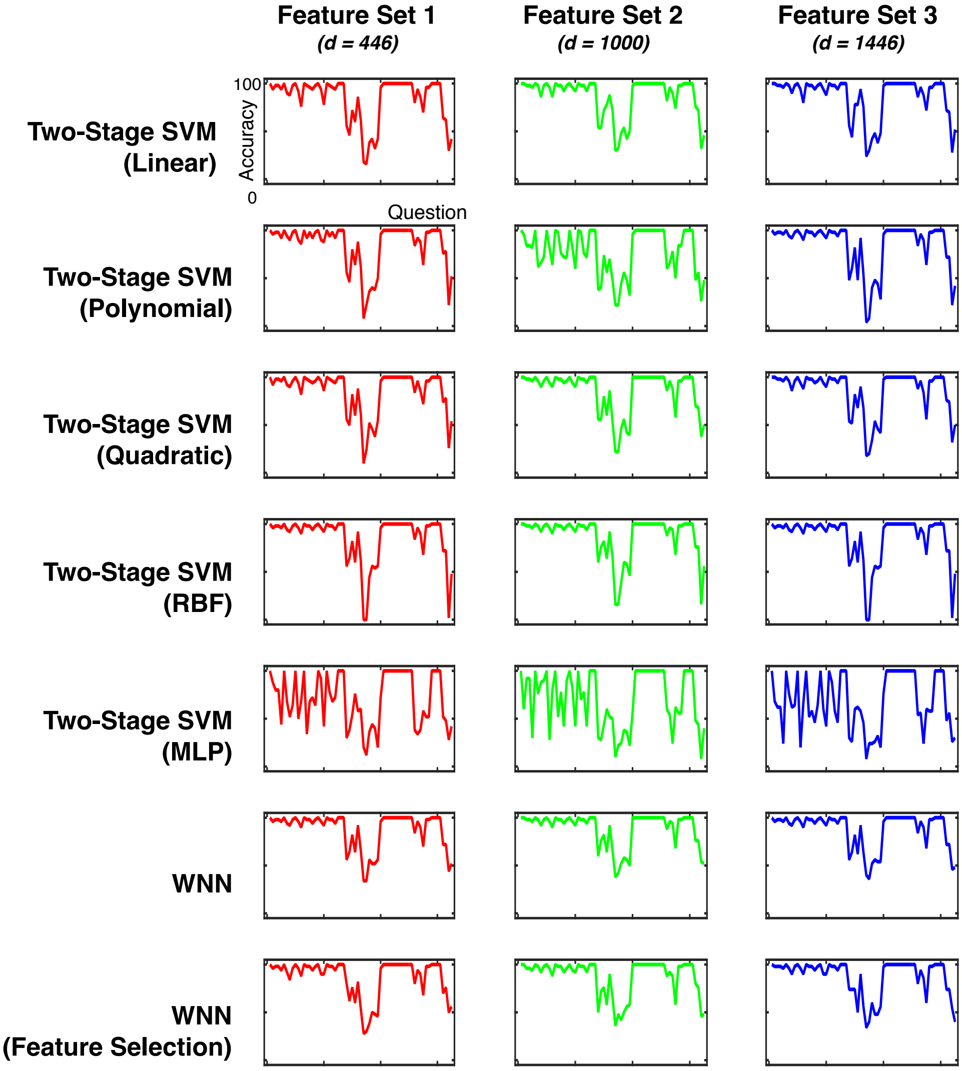

Figure 4: Accuracy of proposed annotation methods on different questions.

using the same kernels (Table 2). These results are consistent across all kernels and all three Feature Sets. This improvement can be attributed to our second stage of 1-vs-1 SVMs breaking ties whenever the first stage 1-vs-all SVMs indicates that multiple labels could be assigned to the question. The second stage of SVMs are trained to discriminate specific pairs of labels and are better able to decide which one is the better answer given the image features of the query. In this manner, the first stage retrieves the collection of most likely labels while the second stage selects the best answer from this collection. An advantage of our approach compared to using a 1-vs-1 approach for *all* questions is the number of SVM classifications that need to be performed. A 1-vs-1 approach would need $\frac{l^2-l}{2}$ SVMs for every question (Section 2.4), requiring over 700 SVM classifications for our set of questions.

It is interesting to note that one-stage RBF SVMs had the same accuracy when using Feature Set 1 as when using Feature Set 3. This was also true for two-stage RBF SVMs. Deeper analysis showed that the accuracy was the same for every fold and every question individually. This is a surprising finding and suggests that the addition of CoG features hinders the higher accuracy achieved by SIFT BoVF alone (Feature Set 2). The other findings in Table 2 reveal that no one feature set was universally better than the others. These outcomes indicate the possibility that ensemble systems using different combinations of features and image recognition methods could be a feasible method for incrementally enhancing accuracy of the annotation. It is interesting to note that, while Feature Set 3 worked best only in combination with two-stage quadratic SVMs it generally had the second highest accuracy for almost all the other methods; this suggests that Feature Set 3 is more robust to the choice of method. Feature Set 2 achieved ≥ 80% accuracy for most of the methods that we evaluated. This is because SIFT features represent local features extracted from key points and are therefore better able to encapsulate the most important visual characteristics of each image. Another advantage of SIFT BoVF (used in Feature Sets 2 and 3) is that its extraction is not dependent upon any form of segmentation. As such Feature Set 2 can be applied in completely automatic annotation approaches.

The accuracy of our approach is consistently low for questions related to the area and characteristics of the lesions, as evidenced by the dip in accuracy in the centre of the charts in Figures 3 and 4. Many of these questions had large label sets (up to 10 labels) and as such there were some labels that had only one or two samples in the entire dataset. In this scenario, it is realistic to expect a lower accuracy when annotating these very rare cases. The effect of the small dataset meant that significance testing would not be insightful as it would be heavily distorted by the low accuracy of these very rare cases. However, it is interesting to note that two-stage SVMs with polynomial and MLP kernels had slightly higher accuracy than other methods for these questions indicating their potential for annotating even these rare cases. We could therefore achieve higher accuracies by optimising each question separately, using unique annotation methods and features for different questions, instead of the current universal annotation technique. It would also be possible to improve the accuracy of these annotations by adapting and integrating methods optimised specifically for annotating lesion characteristics [18].

4. Conclusions

In this paper we provide a new concept for extending CBIR methods to automatically annotate liver CT images, by deriving the annotations from the most semantically relevant images within the already labelled collection. Our methods had higher annotation accuracy on small datasets when compared to several baseline techniques. This was because the underlying CBIR technologies we extended were effective in high dimensional spaces where the dimensionality was larger than the sample size. Our work also scored the best results at the ImageCLEF Liver Annotation Challenge.

Our methods enable the annotation of the *semantic* content of the image and not simply annotation of the image modality or body region, which was prevalent in earlier work [11]. Our methods annotate more than just binary observations [17] and are also capable of annotating the characteristics of the anatomical structures in the image [18]. We have released an implementation of our method[1] to encourage future research in semantic image annotation using our work as a baseline.

In future work, we plan to improve our approach by an optimised fusion of methods in which each question is annotated using the best performing method and feature set combination. We will examine data augmentation techniques to boost the number of training samples to ensure that annotations by our method are significantly more accurate compared to baseline methods. As part of this, we will also introduce methods to make our approach robust to missing values in the data. We will also investigate recent CBIR work that incorporates complementary non-image information [43], e.g., the semantic distance [44] between related terms in an ontology. We will also examine emerging deep learning techniques that have shown great promise in image recognition and classification [45, 46]. We have already begun adapting deep learning methods for modality [47] and body region [48] annotation.

[1] `http://sydney.edu.au/~engineering/it/~ashnil/code/liverannot.html`

Acknowledgments

The authors would like to thank the ImageCLEF organisers for their independent evaluation of our annotation method.

This work was supported in part by Australian Research Council grants.

This research was supported by the Faculty of Engineering and Information Technologies, The University of Sydney under the Faculty Research Cluster Program.

References

[1] H. Müller, N. Michoux, D. Bandon, A. Geissbuhler, A review of content-based image retrieval systems in medical applications—clinical benefits and future directions, International Journal of Medical Informatics 73 (1) (2004) 1 – 23.

[2] A. Kumar, J. Kim, W. Cai, D. Feng, Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data, Journal of Digital Imaging 26 (6) (2013) 1025–1039.

[3] J. Kalpathy-Cramer, A. G. S. de Herrera, D. Demner-Fushman, S. Antani, S. Bedrick, H. Müller, Evaluating performance of biomedical image retrieval systems—An overview of the medical image retrieval task at ImageCLEF 2004–2013, Computerized Medical Imaging and Graphics 39 (2015) 55–61.

[4] H. Pourghassem, H. Ghassemian, Content-based medical image classification using a new hierarchical merging scheme, Computerized Medical Imaging and Graphics 32 (8) (2008) 651 – 661.

[5] L. Setia, A. Teynor, A. Halawani, H. Burkhardt, Grayscale medical image annotation using local relational features, Pattern Recognition Letters 29 (15) (2008) 2039 – 2045.

[6] T. Tommasi, F. Orabona, B. Caputo, Discriminative cue integration for medical image annotation, Pattern Recognition Letters 29 (15) (2008) 1996 – 2002.

[7] P. Wennerberg, K. Schulz, P. Buitelaar, Ontology modularization to improve semantic medical image annotation, Journal of Biomedical Informatics 44 (1) (2011) 155 – 162.

[8] D. L. Weiss, C. P. Langlotz, Structured Reporting: Patient Care Enhancement or Productivity Nightmare?, Radiology 249 (3) (2008) 739–747.

[9] F. M. Hall, The Radiology Report of the Future, Radiology 251 (2) (2009) 313–316.

[10] D. Channin, P. Mongkolwat, V. Kleper, K. Sepukar, D. Rubin, The caBIG$^{\text{TM}}$ Annotation and Image Markup Project, Journal of Digital Imaging 23 (2) (2010) 217–225.

[11] T. Deselaers, T. M. Deserno, H. Müller, Automatic medical image annotation in ImageCLEF 2007: Overview, results, and discussion, Pattern Recognition Letters 29 (15) (2008) 1988 – 1995.

[12] B. C. Ko, J. H. Lee, J.-Y. Nam, Automatic medical image annotation and keyword-based image retrieval using relevance feedback, Journal of Digital Imaging 25 (4) (2012) 454–465.

[13] A. Gilbert, L. Piras, J. Wang, F. Yan, E. Dellandrea, R. Gaizauskas, M. Villegas, K. Mikolajczyk, Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation Task, in: CLEF 2015 Working Notes, vol. 1391 of *CEUR Workshop Proceedings*, 2015.

[14] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. Berg, T. Berg, BabyTalk: Understanding and Generating Simple Image Descriptions, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (12) (2013) 2891–2903.

[15] G. Wang, D. Forsyth, D. Hoiem, Comparative object similarity for improved recognition with few or no examples, in: IEEE Conference on Computer Vision and Pattern Recognition, 3525–3532, 2010.

[16] C. Lampert, H. Nickisch, S. Harmeling, Attribute-Based Classification for Zero-Shot Visual Object Categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (3) (2014) 453–465.

[17] F. Gimenez, J. Xu, Y. Liu, T. Liu, C. Beaulieu, D. Rubin, S. Napel, Automatic Annotation of Radiological Observations in Liver CT Images, in: AMIA Annual Symposium Proceedings, 257–263, 2012.

[18] A. Depeursinge, C. Kurtz, C. Beaulieu, S. Napel, D. Rubin, Predicting Visual Semantic Descriptive Terms From Radiological Image Data: Preliminary Results With Liver Lesions in CT, IEEE Transactions on Medical Imaging 33 (8) (2014) 1669–1676.

[19] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1349 –1380.

[20] B. Caputo, H. Müller, J. Martinez-Gomez, M. Villegas, B. Acar, N. Patricia, N. Marvasti, S. Üsküdarlı, R. Paredes, M. Cazorla, I. Garcia-Varea, V. Morell, ImageCLEF 2014: Overview and analysis of the results, in: CLEF proceedings, vol. 8685 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 192–211, 2014.

[21] N. Marvasti, N. Kökciyan, R. Türkay, A. Yazıcı, P. Yolum, S. Üsküdarlı, B. Acar, ImageCLEF Liver CT Image Annotation Task 2014, in: CLEF 2014 Working Notes, vol. 1180 of *CEUR Workshop Proceedings*, 329–340, 2014.

[22] A. Kumar, S. Dyer, C. Li, P. H. W. Leong, J. Kim, Automatic Annotation of Liver CT Images: the Submission of the BMET Group to ImageCLEFmed 2014 , in: CLEF 2014 Working Notes, vol. 1180 of *CEUR Workshop Proceedings*, 428–437, 2014.

[23] N. Kokciyan, R. Turkay, S. Uskudarli, P. Yolum, B. Bakir, B. Acar, Semantic Description of Liver CT Images: An Ontological Approach, IEEE Journal of Biomedical and Health Informatics, 18 (4) (2014) 1363–1369.

[24] B. Scholkopf, A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA, USA, 2002.

[25] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13 (1) (1967) 21–27.

[26] M.-K. Hu, Visual pattern recognition by moment invariants, IRE Transactions on Information Theory 8 (2) (1962) 179 –187.

[27] R. M. Haralick, K. Shanmugam, I. Dinstein, Textural Features for Image Classification, IEEE Transactions on Systems, Man and Cybernetics 3 (6) (1973) 610–621.

[28] A. Jain, F. Farrokhnia, Unsupervised texture segmentation using Gabor filters, in: IEEE International Conference on Systems, Man and Cybernetics, 14–19, 1990.

[29] H. Tamura, S. Mori, T. Yamawaki, Textural Features Corresponding to Visual Perception, IEEE Transactions on Systems, Man and Cybernetics 8 (6) (1978) 460–473.

[30] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: IEEE Conference on Computer Vision and Pattern Recognition, 511–518, 2001.

[31] D. G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[32] X. Zhou, R. Stern, H. Müller, Case-based fracture image retrieval, International Journal of Computer Assisted Radiology and Surgery 73 (2012) 401–411.

[33] O. A. Penatti, F. B. Silva, E. Valle, V. Gouet-Brunet, R. da S. Torres, Visual word spatial arrangement for image retrieval and classification, Pattern Recognition 47 (2) (2014) 705–720.

[34] S. Aksoy, R. M. Haralick, Feature normalization and likelihood-based similarity measures for image retrieval, Pattern Recognition Letters 22 (5) (2001) 563 – 582.

[35] A. Kumar, J. Kim, L. Wen, M. Fulham, D. Feng, A graph-based approach for the retrieval of multi-modality medical images, Medical Image Analysis 18 (2) (2014) 330–342.

[36] M. Rahman, S. Antani, R. Long, D. Demner-Fushman, G. Thoma, Multimodal Query Expansion Based on Local Analysis for Medical Image Retrieval, in: B. Caputo, H. Müller, T. Syeda-Mahmood, J. Duncan, F. Wang, J. Kalpathy-Cramer (Eds.), Medical Content-Based Retrieval for Clinical Decision Support, vol. 5853 of *Lecture Notes in Computer Science*, 110–119, 2010.

[37] U. Avni, H. Greenspan, E. Konen, M. Sharon, J. Goldberger, X-ray Categorization and Retrieval on the Organ and Pathology Level, Using Patch-Based Visual Words, IEEE Transactions on Medical Imaging 30 (3) (2011) 733–746.

[38] P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection, Pattern Recognition Letters 15 (11) (1994) 1119–1125.

[39] B. Ermis, T. Cemgil, N. Marvasti, B. Acar, Liver CT Annotation via Generalized Coupled Tensor Factorization, in: CLEF 2014 Working Notes, vol. 1180 of *CEUR Workshop Proceedings*, 421–427, 2014.

[40] A. Spanier, L. Joscowicz, Towards Content-Based Image Retrieval: From Computer Generated Features to Semantic Descriptions of Liver CT Scans, in: CLEF 2014 Working Notes, vol. 1180 of *CEUR Workshop Proceedings*, 421–427, 2014.

[41] I. Nedjar, S. Mahmoudi, M. A. Chikh, K. Abi-yad, Z. Bouafia, Automatic Annotation of Liver CT Image: ImageCLEFmed 2015, in: CLEF 2015 Working Notes, vol. 1391 of *CEUR Workshop Proceedings*, 2015.

[42] N. Marvasti, M. del Mar Roldán, S. Üsküdarlı, J. F. Aldana, B. Acar, Overview of the ImageCLEF 2015 Liver CT Annotation Task, in: CLEF 2015 Working Notes, vol. 1391 of *CEUR Workshop Proceedings*, 2015.

[43] C. Kurtz, A. Depeursinge, S. Napel, C. F. Beaulieu, D. L. Rubin, On combining image-based and ontological semantic dissimilarities for medical image retrieval applications, Medical Image Analysis 18 (7) (2014) 1082–1100.

[44] A. Budanitsky, G. Hirst, Evaluating WordNet-based Measures of Lexical Semantic Relatedness, Computational Linguistics 32 (1) (2006) 13–47.

[45] I. Arel, D. Rose, T. Karnowski, Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier], IEEE Computational Intelligence Magazine 5 (4) (2010) 13–18.

[46] D. Ciresan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 3642–3649, 2012.

[47] D. Lyndon, A. Kumar, J. Kim, P. H. W. Leong, D. Feng, Convolutional Neural Networks for Medical Classification, in: CLEF 2015 Working Notes, vol. 1391 of *CEUR Workshop Proceedings*, 2015.

[48] D. Lyndon, A. Kumar, J. Kim, P. H. W. Leong, D. Feng, Convolutional Neural Networks for Medical Clustering, in: CLEF 2015 Working Notes, vol. 1391 of *CEUR Workshop Proceedings*, 2015.