Automatic Annotation of Liver CT Images: the Submission of the BMET Group to ImageCLEFmed 2014

Ashnil Kumar[1,3], Shane Dyer[2], Changyang Li[1,3], Philip H. W. Leong[2,3], and Jinman Kim[1,3]

[1] School of Information Technologies, University of Sydney, Australia
[2] School of Electrical and Information Engineering, University of Sydney, Australia
[3] Institute of Biomedical Engineering and Technology, University of Sydney, Australia
{ashnil.kumar,changyang.li,philip.leong,jinman.kim}@sydney.edu.au

Abstract. In this paper we present the strategies that were designed and applied by the Institute of Biomedical Engineering and Technology (BMET) team to the liver image annotation task of ImageCLEF 2014. This was the first year this challenge was held and as such our strategies form the basis for future exploration in this area. The major challenge of the liver annotation task was limited training data, with some annotation labels having very few positive samples and other labels having no samples at all. We propose two strategies for annotating the liver images given the unbalanced nature of the training dataset. Our first method uses multi-class classification scheme where each label has a classifier that is trained to separate it from the other labels. Our second method uses the similarity scores from an image retrieval algorithm as weights for a majority voting scheme, thereby reducing the inherent bias towards labels that have a high number of samples. We also investigate the performance of our methods using different feature sets. In total, BMET submitted 8 runs to the ImageCLEF liver annotation task. All of our runs achieved high scores ($> 90\%$) during evaluation. We also achieved the highest score out of all submissions to the ImageCLEF 2014 liver annotation task.

Keywords: SVM, Image Retrieval, Multi-class Classification, Image Annotation, Liver, Computed Tomography

1 Introduction

ImageCLEF [1] is the image retrieval track of the Cross Language Evaluation Forum (CLEF). In the past, one of the major focuses of ImageCLEF [2–8] has been medical imaging, with tasks ranging from modality-classification to case-based retrieval. In 2014, for the first time, the objective of the ImageCLEF medical imaging task was the automatic annotation of medical images [9]. The aim of the challenge was to generate a structured report based on an analysis of the image features in computed tomography (CT) images of the liver, with the goal of detecting subtle differences in image features and to annotate them using a standard terminology.

One of the major challenges was the limited amount of training data compared to the number of annotations that needed to be recognised. In particular, there were some annotations that did not occur at all in the dataset. Similarly, there were also instances where all of the training samples had the same annotation. Our methods were thus designed to account for the unbalanced dataset by reducing the bias towards annotations with more samples.

In this paper, we outline our submission to ImageCLEF 2014 liver annotation challenge. We present two methods for the annotation of liver CT images, one based on multi-class classification and another using image retrieval. Our aim was to perform the annotation using visual features only. As such, we do not utilise any of the information in the ONtology of the LIver for RAdiology (ONLIRA) [10], such as the semantic distance between related terms [11]. Our method treats the challenge as multiple multi-class classification problems. We adapted well-established classification and retrieval techniques to investigate their perform on the annotation of liver CT images. We envision that this will establish a baseline for improvements in future iterations of the challenge.

The following terminology is employed 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 a possible annotation that can be assigned to a question. An *answer* is the label that is assigned to the question.

2 Materials

The training 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 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 was provided for each image in the training dataset. The training 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. The answers to 73 questions was provided for each training image.

The test dataset contained 10 CT volumes, with varied resolutions and pixel spacings, cropped to the region around the liver. The test data also included a mask of the liver pixels, a bounding box for a lesion and a set of 60 well-established image features (with a total dimensionality of 458) for each image in the dataset.

3 Methods

3.1 Overview

Our aim was to investigate annotation using variations of two methods: support vector machine (SVM) classification [12] and content-based image retrieval [13]. The specific variations that we used were:

- **Method 1:** Two stage classification using SVMs with linear kernels.
- **Method 2:** Two stage classification using SVMs with radial basis function (RBF) kernels.
- **Method 3:** Content-based image retrieval.
- **Method 4:** Content-based image retrieval with feature selection.

The two-stage classification method is described in Section 3.3 and the image retrieval method is described in Section 3.4.

We also investigated the effect of expanding feature set on all these methods. The feature set expansion is described in Section 3.2.

Our method was applied to a subset of the annotations; seven questions where the label sets were unbounded (e.g., measurements in millimetres or counts of particular objects) were excluded. We also excluded one question that accepted multiple labels as the answer. In total, we answered 64 of the 73 questions.

3.2 Feature Sets

We used two feature sets for the annotation challenge. These were:

- **Feature Set 1:** The well-established features included with dataset after cleaning as described below (*total dimensionality* = 446).
- **Feature Set 2:** Feature Set 1 + a bag-of-visual-words (BoVW) features, constructed as described below (*total dimensionality* = 1446).

Dataset Features The image features in the dataset included features extracted from the liver, the hepatic vasculature of the liver, and the selected lesion. Features extracted globally across all lesions were also included. The features described object shape properties (e.g., volume, surface area, sphericity, solidity, convexity, Hu shape invariants [14]), texture information (e.g., Haralick [15], Gabor [16], Tamura [17], Haar [18]), and pixel intensity information.

We cleaned the feature data by removing feature dimensions that had a not-a-number (NaN) value or that were used to scale other features. We removed the Anatomical Location feature (5 dimensions) of the lesion since one training image had NaN values for this feature. The Hu Moments feature (3 dimensions) of the lesion was also removed for the same reason. We also removed the first two dimensions (upper and lower bounds) of the Histogram feature of the lesion and the HistogramOfAllLesions (a feature extracted across all lesions). The cleaned feature set had a dimensionality of 446. Readers are directed to the task documentation for more information about the image features.

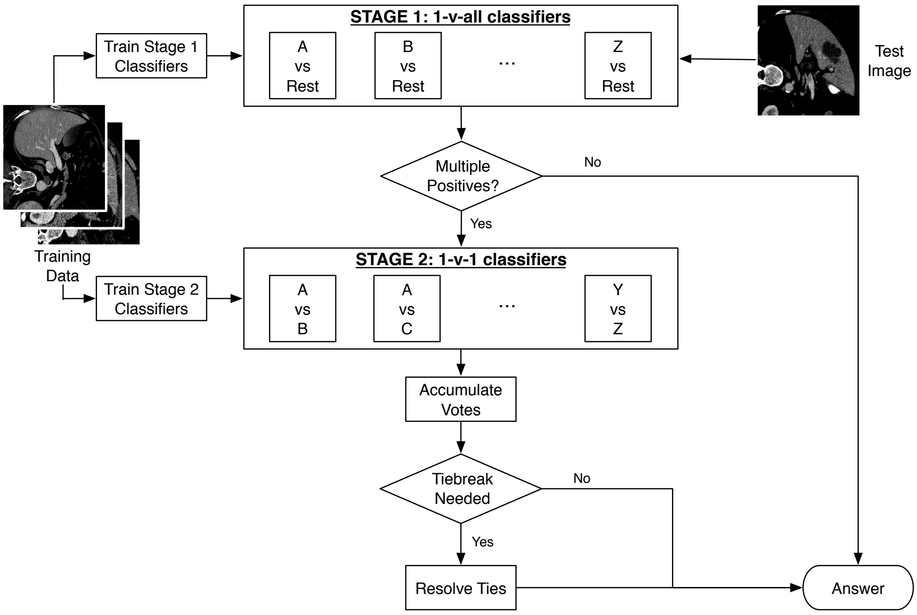

Fig. 1: An overview of the classification scheme for annotation.

Bag-of-Visual-Words We extracted Scale Invariant Feature Transform (SIFT) descriptors [19] from key points detected in the 2D slices of the CT images. There were a total of 4,524,946 descriptors extracted from the training dataset and 433,846 descriptors extracted from the test dataset.

We created a visual codebook from the SIFT descriptors extracted from the training dataset. We randomly sampled 5% of the descriptors and grouped the subsampled data using k-means clustering with $k = 1000$. The cluster centres were treated as the visual words in the codebook. A single visual word was assigned to every descriptor in both datasets; this assignment was determined by finding the visual word whose cluster centre with the minimum Euclidean distance from the descriptor. A BoVW descriptor was then created for every image using a 1000-bin histogram of the visual words in that image [20].

3.3 Annotation using Two Stage Classification

Our two stage classification approach for image annotation is shown in Figure 1. Each stage consisted of a bank of several SVM classifiers. This two stage approach was repeated separately for each question. Due to the unbalanced training dataset, we expected that the classifiers for labels with low samples would have relatively low accuracy. For this reason, we used the two stage approach to introduce further discriminative power, especially in the case of ties.

Let Ω be a question. Also let $\mathcal{L}_\Omega$ be the set of labels for Ω with $|\mathcal{L}_\Omega| = l$. For every label $A \in \mathcal{L}$, we trained a A-vs-rest (1-vs-all) SVM classifier, hence forming l 1-vs-all classifiers. We also trained A-vs-B (1-vs-1) SVMs for every pair of labels $A, B \in \mathcal{L}$ where $A \neq B$, forming a total of $\left(l^2 - l\right)/2$ 1-vs-1 classifiers. For every question, our first stage was composed of the 1-vs-all classifiers and the second stage was composed of the 1-vs-1 classifiers.

After the classifiers have been trained, our annotation process proceeded as follows. An un-annotated image (from the test dataset) was 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.

Let $\mathcal{L}^+ \subseteq \mathcal{L}$ be the set of labels given positive responses by the first stage of classifiers. During the second stage, we classified the un-annotated image using the 1-vs-1 classifiers for all the labels in $\mathcal{L}^+$ (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 were completed. 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", noted in the task description as the label selected when the radiologist was unsure of the correct annotation. 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.

During training, we tested our algorithm on various SVM kernels (linear, quadratic, radial basis function (RBF), multilayer perceptron, polynomial) and parameters using 10-fold cross validation. We discovered that the best overall accuracy was achieved by the RBF kernel with scaling factor equal to 1. There were only five questions in which the RBF kernel was beaten by other kernels and in each of these cases the difference was not significant. We therefore selected the commonly-used linear kernel and the RBF kernels for our classification approach (Methods 1 and 2, respectively).

3.4 Annotation using Image Retrieval

Our image retrieval based approach for annotation used the most similar training images to select the answers for an un-annotated image. While the classification approach (Section 3.3) trained separate classifiers for each question, the retrieval approach attempted to answer all of the questions together. An overview of the method is shown in Figure 2.

We defined the similarity of the the unannotated image (U) and a training image (T) as the Euclidean distance between their respective feature vectors:

$$s\left(U, T\right) = \sqrt{\sum_{i=0}^{d} \left(u_i - t_i\right)^2} \tag{1}$$

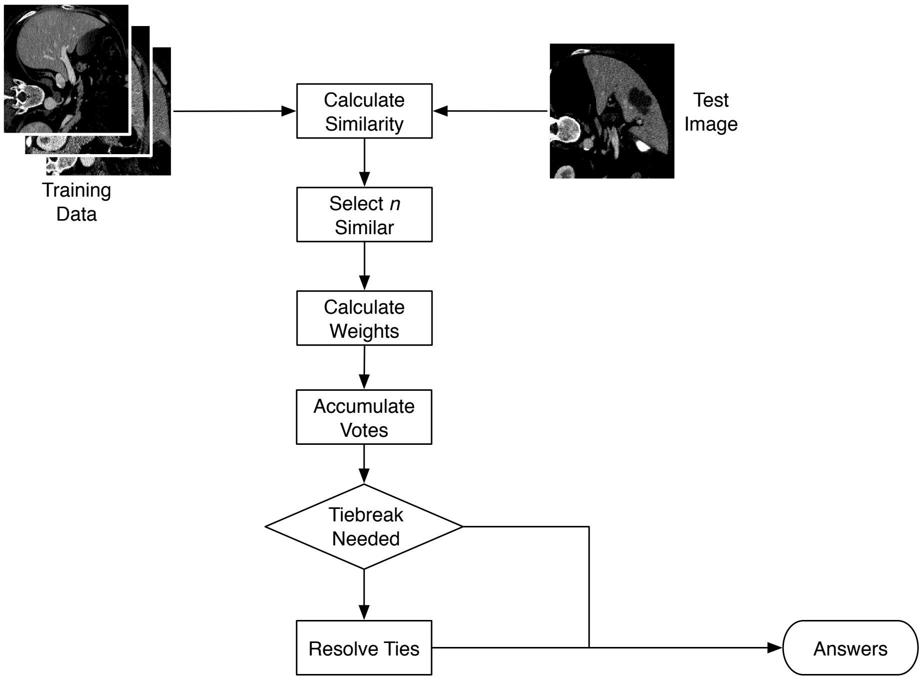

Fig. 2: An overview of the retrieval scheme for annotation.

where u_i was the i-th feature in the feature vector of U, t_i was the i-th feature in the feature vector of T, and d was the dimensionality of the feature set (see Section 3.2). Under this formulation, lower values of s indicated greater similarity with $s(U,T) = 0$ implying that U and T were exactly similar.

We selected the n most similar images from the training data. Let $S = \{s_1, ..., s_n\}$ be the similarity values of these images (sorted from most similar to least similar). A weighted voting scheme was used to select the answer for each question. The weights for the voting scheme were determined from the set of similarity values. The weight w_i for the i-th most similar image was calculated as:

$$w_i = \frac{c \times s_i}{s_1} \tag{2}$$

where c was a constant scaling factor and $s_i \in S$ was the similarity value of the i-th most similar image. In our experiments, we empirically set $n = 10$ and $c = 10$. Our weighting scheme adjusted the value of the vote based on the calculated similarity. Images with a higher similarity would thus have a stronger vote compared to images with a lower similarity. The weighting was necessary due to the unbalanced nature of the dataset. If a majority voting scheme was used then the 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).

Table 1: Summary of Results[1]

Run	Method	Feature Set	Completeness	Accuracy	Score
1	1	1	0.98	0.89	0.935
2	1	2	0.98	0.90	0.939
3	2	1	0.98	0.89	0.933
4	2	2	0.98	0.90	0.939
5	3	1	0.98	0.91	0.947
6	3	2	0.98	0.87	0.927
7	4	1	0.98	0.91	0.947
8	4	2	0.98	0.87	0.926

In the case of a tie (multiple labels having the same weighted vote), we set the answer to "other". When "other" was not part of the label set, we selected the label "N/A" if it was available or "false" for questions that expected a boolean answer.

We also designed an alternate question-based version of our retrieval scheme for annotation; this is noted as Method 4 in Section 3.1. In the alternate scheme, we applied sequential feature selection [21] to use the most discriminating features for each question during the similarity calculation (Equation 1). This ensured that optimised features were used to retrieve the most similar images, i.e., image retrieval was performed using the features most suited for answering a particular question.

4 Results and Discussion

We submitted eight runs to the ImageCLEF 2014 liver annotation challenge. The runs were created using a combination of the four methods listed in Section 3.1 and the two feature sets listed in Section 3.2. The runs were evaluated on *completeness*, the percentage of questions that were answered, and *accuracy*, the percentage of completed questions with a correct answer. Only 65 questions formed part of the evaluation; questions with unbounded labels (e.g., measurements) were not evaluated as part of the 2014 challenge. Table 1 shows the mean completeness and accuracy of our eight runs.

There were 20 registered participants for the ImageCLEF 2014 liver annotation challenge. However, only three groups (including ours) submitted runs for evaluation. A comparison of the groups is shown in Table 2. In 2014, our submission achieved the highest score of all the participants.

The results show that all of our runs achieved high scores (> 0.92). We achieved a completeness score of 0.98 for every run because we always answered 64 of the 65 questions. The question that we excluded from our submission was

[1] These results are from `http://www.imageclef.org/2014/liver#Results`.

Table 2: Comparison of Results[1]

Group	Completeness	Accuracy	Score
BMET (our group)	0.98	0.91	0.94
CASMIP	0.95	0.91	0.93
piLabVAVlab	0.51	0.39	0.45

one that accepted multiple labels as the answer. We could answer this question and achieve a perfect completeness score by removing the tiebreaker from both of our annotation methods. This would be an exception for only this particular question, i.e., all other questions would still go through a tiebreaker process if necessary.

In general, there were no large differences between the two variants of the classification method. That is, the accuracy of two classification methods (Method 1 and Method 2) were approximately the same. This suggests that the choice of kernel was not a major factor in the overall accuracy of the annotation. This outcome was contrary to the training stage (as stated in Section 3.3) where our selection of the RBF kernel (Method 2) was due to its higher accuracy compared to the other kernels. We attribute this difference to the unbalanced training dataset, which may not have reflected the labels of the test dataset. However, our high scores (> 0.93) demonstrate that our classification approach for annotation performs well despite the unbalanced training dataset.

The score of the retrieval method with feature selection (Method 4) was equal to or less than that of the retrieval method with no feature selection (Method 3). This result is counter-intuitive as the expectation is that feature selection would improve the accuracy of the annotation. One explanation for this could be that the feature selection reduces the similarity scores calculated during retrieval (since fewer features are used), which in turn negatively impacts the weighted vote by returning voting power to labels with a larger number of training samples.

It is interesting to note that the classification methods performed best when using the expanded feature set (Feature Set 2) while the retrieval methods performed best when using the normal feature set (Feature Set 1). This suggests that one of the major considerations in the annotation of the liver is the combination of features and methods. That is, the findings indicate that one cannot choose a method for annotation without considering which features will be used, and vice versa.

5 Conclusions

This paper described the methods and results of the BMET group's submission to the liver annotation task of ImageCLEF 2014. Our eight runs investigated different combinations of methods and feature sets. While all of our runs achieved high scores they also revealed the areas in which our method could be optimised.

Our future work will investigate building associations between image features and ONLIRA terms to create classifiers for labels with no samples in the training dataset [22].

Acknowledgments. The authors would like to thank the organisers of the ImageCLEF 2014 liver annotation task for their assistance in procuring the detailed results.

References

1. Caputo, B., Müller, H., Martinez-Gomez, J., Villegas, M., Acar, B., Patricia, N., Marvasti, N., Üsküdarlı, S., Paredes, R., Cazorla, M., Garcia-Varea, I., Morell, V.: ImageCLEF 2014: Overview and analysis of the results. In: CLEF proceedings. Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2014)
2. Müller, H., Deselaers, T., Deserno, T., Kalpathy-Cramer, J., Kim, E., Hersh, W.: Overview of the ImageCLEFmed 2007 medical retrieval and medical annotation tasks. In Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D., Peñas, A., Petras, V., Santos, D., eds.: Advances in Multilingual and Multimodal Information Retrieval. Volume 5152 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2008) 472–491
3. Müller, H., Kalpathy-Cramer, J., Kahn, C., Hatt, W., Bedrick, S., Hersh, W.: Overview of the ImageCLEFmed 2008 medical image retrieval task. In Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G., Kurimo, M., Mandl, T., Peñas, A., Petras, V., eds.: Evaluating Systems for Multilingual and Multimodal Information Access. Volume 5706 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2009) 512–522
4. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Radhouani, S., Bakke, B., Kahn, C., Hersh, W.: Overview of the CLEF 2009 medical image retrieval track. In Peters, C., Caputo, B., Gonzalo, J., Jones, G., Kalpathy-Cramer, J., Müller, H., Tsikrika, T., eds.: Multilingual Information Access Evaluation II. Multimedia Experiments. Volume 6242 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2010) 72–84
5. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., de Herrera, A.G.S., Tsikrika, T.: Overview of the CLEF 2011 medical image classification and retrieval tasks. In: CLEF (Notebook Papers/Labs/Workshop). (2011)
6. Müller, H., de Herrera, A.G.S., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., Eggel, I.: Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In: CLEF (Online Working Notes/Labs/Workshop). (2012)
7. de Herrera, A.G.S., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., Müller, H.: Overview of the ImageCLEF 2013 medical tasks. Working notes of CLEF (2013)
8. Kalpathy-Cramer, J., de Herrera, A.G.S., Demner-Fushman, D., Antani, S., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems—an overview of the medical image retrieval task at ImageCLEF 2004–2013. Computerized Medical Imaging and Graphics (2014) doi: 10.1016/j.compmedimag.2014.03.004.
9. Marvasti, N., Kökciyan, N., Türkay, R., Yazıcı, A., Yolum, P., Üsküdarlı, S., Acar, B.: ImageCLEF Liver CT Image Annotation Task 2014. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. (2014)

10. Kokciyan, N., Turkay, R., Uskudarli, S., Yolum, P., Bakir, B., Acar, B.: Semantic description of liver CT images: An ontological approach. IEEE Journal of Biomedical and Health Informatics (2014) 10.1109/JBHI.2014.2298880.
11. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of lexical semantic relatedness. Computational Linguistics **32**(1) (2006) 13–47
12. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA (2002)
13. Kumar, A., Kim, J., Cai, W., Feng, D.: Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data. Journal of Digital Imaging **26**(6) (2013) 1025–1039
14. Hu, M.K.: Visual pattern recognition by moment invariants. Information Theory, IRE Transactions on **8**(2) (1962) 179 –187
15. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics **3**(6) (1973) 610 –621
16. Jain, A., Farrokhnia, F.: Unsupervised texture segmentation using gabor filters. In: IEEE International Conference on Systems, Man and Cybernetics. (1990) 14–19
17. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. IEEE Transactions on Systems, Man and Cybernetics **8**(6) (1978) 460–473
18. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (2001) 511–518
19. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
20. Zhou, X., Stern, R., Müller, H.: Case-based fracture image retrieval. International Journal of Computer Assisted Radiology and Surgery **7** (2012) 401–411
21. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern Recognition Letters **15**(11) (1994) 1119 – 1125
22. Lampert, C., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(3) (2014) 453–465