Multi-stage Thresholded Region Classification for Whole-Body PET-CT Lymphoma Studies

Lei Bi[1], Jinman Kim[1], Dagan Feng[1,2], and Michael Fulham[1,3,4]

[1] School of Information Technologies, University of Sydney, Australia
[2] Med-X Research Institute, Shanghai Jiao Tong University, China
[3] Department of Molecular Imaging, Royal Prince Alfred Hospital, Australia
[4] Sydney Medical School, University of Sydney, Australia

Abstract. Positron emission tomography computed tomography (PET-CT) is the preferred imaging modality for the evaluation of the lymphomas. Disease involvement in the lymphomas usually appear as foci of increased Fluorodeoxyglucose (FDG) uptake. Thresholding methods are applied to separate different regions of involvement. However, the main limitation of thresholding is that it also includes regions where there is normal FDG excretion and FDG uptake (NEUR) in structures such as the brain, bladder, heart and kidneys. We refer to these regions as NEURs (the normal excretion and uptake (of FDG) regions). NEURs can make image interpretation problematic. The ability to identify and label NEURs and separate them from abnormal regions is an important process that could improve the sensitivity of lesion detection and image interpretation. In this study, we propose a new method to automatically separate NEURs in thresholded PET images. We propose to group thresholded regions of the same structure with spatial and texture based clustering; we then classified NEURs on PET-CT contextual features. Our findings were that our approach had better accuracy when compared to conventional methods.

1 Introduction

Fluorodeoxyglucose positron emission tomography computed tomography (FDG PET-CT) is regarded as the imaging modality of choice for the evaluation staging, assessment of response / relapse of the lymphomas, where sites of disease usually display increased FDG uptake and the co-registered CT provides anatomical localization [6] [13]. A semiquanitative measure of FDG uptake is referred to as a standard uptake value (SUV), which is a radiotracer concentration normalized by patient mass [13]. The SUV is commonly used to describe regions of abnormal FDG uptake relative to other structures and SUV thresholding is the most common method to identify these in patients with lymphoma. Some investigators have proposed methods to calculate the threshold such as 50%SUV$_{max}$ or a SUV=2.5 [13]. A consequence of these methods is that when applied globally to the entire image, the FDG excretion by the kidneys and the normal high FDG such as cerebral uptake are delineated together with sites of

P. Golland et al. (Eds.): MICCAI 2014, Part I, LNCS 8673, pp. 569–576, 2014.

disease. Further, NEURs are often fragmented into a number of regions in a single structure, which make image interpretation more problematic. The ability to identify and label NEURs and separate them from sites of disease will improve lesion detection, interpretation and visualization.

In this study, we propose a multi-stage method to automatically label NEURs from thresholded PET images. PET-CT images were used to derive contextual image features with high discriminative attributes by taking advantage of the high PET sensitivity and anatomical localization data from CT. We used a spatial and texture based clustering algorithm to group the thresholded regions belonging to the same structure and then classified these grouped regions into one of the NEUR classes according to combined contextual features derived from PET-CT images.

1.1 Related Work

Our study relates to image classification techniques that attempt to separate and label different structures using image contextual features. We define related work into three main categories:

Abnormality detection research that attempted to detect only one type of abnormality, such as for liver tumors [11]. These methods rely on the selection of appropriate image features to separate abnormal and normal regions; they typically require segmentation to derive prior knowledge of the abnormalities, which adds complexity to the classification.

Multi-structure localization methods that detect and semantically label anatomical structures, such as the method proposed by Criminisi et al., [3]. These approaches generally only consider healthy normal structures, rather than abnormal structures.

Abnormality detection and multi-structure labeling methods label the structure and abnormalities usually in parallel. These methods rely on contextual features to separate normal from abnormal and rely exclusively on the localization of normal structures [14] [12].

Our study also uses contextual features to identify normal structures but we differ from previous work as follows: (1) we do not rely solely on contextual features because PET images have inconsistent localization information and have the inherent variability of FDG uptake among patients, NEURs are not consistent from patient to patient and, (2) we deal with whole-body PET-CT images rather than limited images of a particular region e.g. thorax or abdomen, which have greater clinical relevance than a limited assessment of the body.

2 Methods

2.1 Materials and Ground Truth Construction

Our dataset consists of 33 whole-body PET-CT studies from 10 lymphoma patients provided by the Department of Molecular Imaging, Royal Prince Alfred

(RPA) Hospital, Sydney; each patient had multiple scans (3 patients with 2 scans, 3 scans and 4 scans, each; 1 with 6 scans) during diagnosis and treatment of their lymphomas. All studies were acquired using a Siemens Biograph TruePoint PET-CT scanner (Siemens Medical Solutions, Hoffman Estates, IL, USA) with a PET resolution of 168×168 pixels at $4.07mm^2$ and CT resolution of 512×512 pixels at $0.98mm^2$ and slice thickness of $3mm$. The bed and linen were removed from CT by adaptive thresholding and image subtraction from a bed template [9].

Training data and ground truth data were constructed using the PET Response Criteria in Solid Tumors (PERCIST) thresholding method on each PET image (see Section 2.3). The resulting binary mask, consisting of NEUR was then manually labeled as belonging to the brain, bladder, heart, left kidney, right kidney or other structures. The other class contained regions of increased FDG uptake (identified from the clinical report) related uptake in brown fat and lymph node inflammation. A total of 503 thresholded regions were manually labeled and included 42 brain, 32 bladder, 35 heart, 73 left kidney, 75 right kidney and 246 other regions.

2.2 Multi-stage Classification Framework

Fig.1 shows the overview of the proposed classification framework; there are 4 main components: the PET image was thresholded based on PERCIST and its counterpart CT image was pre-processed to detect the bony skeleton (Section 2.3). The skeleton was then removed from the PET image and the remaining pixels were then grouped into individual regions via connected thresholding. A spatial and texture based clustering were then applied to group the fragmented regions into a structure (Section 2.4) prior to a contextual features based classification for NEURs labelling (Section 2.5).

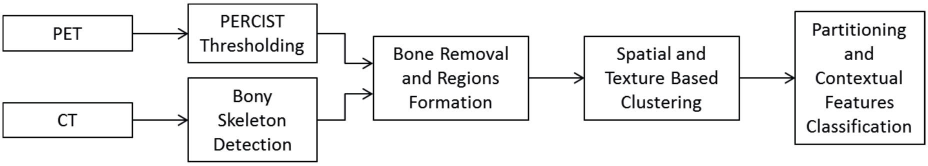
PET
PERCIST Thresholding
CT
Bony Skeleton Detection
Bone Removal and Regions Formation
Spatial and Texture Based Clustering
Partitioning and Contextual Features Classification

Fig. 1. Overview of our proposed multi-stage classification framework

2.3 Automatic PERCIST thresholding and Bony Skeleton Detection

PERCIST is a robust method for calculating the SUV threshold based on the combined use of a SUV normalized with lean body mass (SUV$_{LBM}$) together with a reference region of interest (ROI) [13]. We adopted the automated PERCIST calculation in Bi et al., [1] to generate a binary mask $T_{PERCIST}$. Here, the reference ROI was a sphere of diameter $3cm$ that was placed within the right lobe of the liver. We segmented the bony skeleton from CT and then removed these structures from the PERCIST thresholded PET image. A binary skeleton $T_{skeleton}$ mask was generated using a threshold of > 150 Hounsfield Units (HU) [7] on the CT image. $T_{skeleton}$ was subtracted from $T_{PERCIST}$. A morphological filter was applied on the resulting binary mask to remove noise.

2.4 Spatial and Texture Based Clustering

Thresholding methods typically result in a structure, e.g. the kidney, being fragmented into many regions. Such fragmentation increases the complexity of label classification (Section 2.5) since each region only partially represents a structure. Thus we grouped these fragmented regions, prior to classification, by identifying groups of similar structures according to their spatial location and texture image features.

Density-based spatial clustering (DBSC) was applied to find a number of clusters from estimated density distributions of regions in the dataset [4]. Formally, DBSC can be defined as a clustering algorithm based on the concept of density reachability (density-connected) where a region R is *directly density-reachable* from region R' if R' has at least κ number of neighbor regions (including R) residing within a given distance ϵ. R is further considered as *density-reachable* from R' if there is a sequence of regions $R_1, \cdots, R_n$ with $R_1 = R$ and $R_n = R'$, where R_{i+1} is *directly density-reachable* from R_i. Therefore, a density-based cluster is the maximum set of *density-reachable* (including directly and non-directly) regions. DBSC starts from a random region and iteratively visits all the regions. To avoid the false clustering of regions where only the spatial distance to each other was used, we incorporated a texture feature similarity between the regions denoted as:

$$D(R, R') = \omega_s \cdot min(\| p - p' \|) + \omega_t \cdot \sum_{i=1}^{n}(\| f_i(R) - f'_i \|_2) \tag{1}$$

where p and p' are the voxels spatial locations $p \in R$, $p' \in R'$ and f, f' representing the texture features. Four texture features (mean, standard deviation, skewness and kurtosis), from the PET and CT images, were used to measure the similarities between the two regions; these features were selected for their proven performances in representing these images [11] [12]. To reduce the variability of FDG uptake across the PET scans when calculating texture similarity, we normalized the FDG uptake into SUV$_{LBM}$. ω_s and ω_t are the weights associated with the spatial distance and texture feature similarity terms. We set the minimum number of neighbor regions as $\kappa = 1$. This ensured that all regions may become a cluster and no fragmented regions were discarded. An equal weight was set to spatial distance and texture similarity. We calculated the distance $\epsilon = 10$ from the training data (plotting all the distances for individual region to its neighbors and then finding the distance that is able to group the maximum number of regions while having the minimum inhomogeneity within the cluster).

2.5 Whole-Body Partitioning and Contextual Features Classification

Prior to NEURs classification, whole-body PET-CT images were partitioned into three sections to reduce the search space: above lungs (AL), lungs (LA) and below lungs (BL). The lung structures were automatically segmented using an established adaptive thresholding method [7] to provide a coarse estimate of the sections.

Our classification was based on contextual features, which included combinations of region-level textures (RLT), scale-invariant features transform (SIFT) [10], and histogram of oriented gradients (HOG) [5]. RLT features were the same as in section 2.4 plus the addition of the average location in transverse, coronal and saggital planes (represented in percentages). RLT were used to describe the regions in a descriptive statistical manner representing a likelihood of a region at a spatial context. SIFT was used to describe the local features and can be considered to return important properties (key points) of the regions. The SIFT is robust for classification in different image scales or noise levels, which is a desired property for PET-CT. We sampled key points over the thresholded regions and a default 128 dimensional feature vector was used to represent each of these key points [10]. The HOG are similar to the SIFT, but they differ in that HOG compute on an overlapping squared cell, from which the edge orientations are measured. We used the same approach suggested by Felzenszwalb et al., [5] to set cell size equal to 8, with 9 directions in each cell. The HOG were sampled by using the cell over the thresholded regions and were represented via a 31 dimensional feature vector.

We used two separate bag-of-words (BoW) histograms to summarize the SIFT and HOG features, individually. Each of BoW histograms had 200 bins (100 bins for PET and CT). The two histograms, together with RLT features, were trained separately with a radial basis function (RBF) kernel to non-linearly map the data into a higher dimension space. This helps to make the training data more separable in a computationally efficient way, where a linear kernel usually has poor performance in a non-linear classification task while a polynomial kernel is computationally expensive [8]. The RBF kernel parameters were optimized with a default grid search analysis method in the LIBSVM described by Chang et al [2]. These features were then fitted into three separate multi-class support vector machine (SVM) (one-against-one) for classification, such that each SVM was optimized for different features. The probability score of region R to be classified as label m was calculated as the weighted combination of all the features defined as:

$$\boldsymbol{P}(R) = \sum_{\varphi \in F} \gamma_{\varphi} \cdot \boldsymbol{\rho}_{\varphi}(R), F = \{RLT, SIFT, HOG\} \tag{2}$$

where $m \in \{Brain, Bladder, Heart, L.kidney, R.kidney, Other\}$, φ is the contextual feature and $\boldsymbol{\rho}_{\varphi}(R)$ is a probability matrix of different labels for given R and it is the output from SVM. γ_{φ} is the associate weight. We we used equal weights for this combination to avoid bias. The final labelling of region R was based on the matrix label with the highest probability.

3 Results and Discussion

We compared the labels assigned by our method with the labels of the ground truth. We used leave the same patient out cross-validation approach in our evaluation (leaving out all scans from the same patient to remove bias). In Fig. 2 we depict our classification results on 4 randomly selected patient studies. Our approach was able to separate NEUR classes; in Fig. 2(b) the kidneys are fragmented and in Fig. 2(a) there are multiple sites of disease in the left axilla and at the base of the left neck.

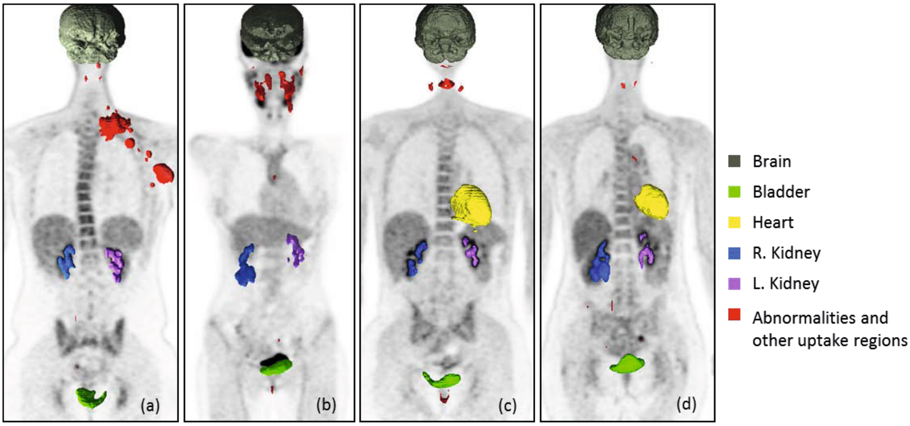

Fig. 2. Classification results from 4 randomly selected studies rendered on PET

We compared our method to two other approaches. The first was a conventional SVM method, similar to the work proposed by Wu et al., [14], where image features were extracted over regions from both PET and CT and fitted with an SVM. The second was based on a whole-body image partition (Section 2.5) and SVM (denoted as P+SVM), which resembles the approach in Song et al., [12]. The results are summarized in Table 1. Our method had higher classification accuracy, which we attribute to the grouping process; 5/5 studies had multiple heart fragments that were correctly grouped and there were 7/8 for the brain. P+SVM performed better compared to SVM, which was likely due to P+SVM restricting the search space during classification. The bladder was consistently classified by all methods, which was likely to be due to the bladder typically having the highest FDG value and in our data, without any fragmentation into multiple regions. The errors were mainly in the misclassification of the other regions. Two right kidney regions were wrongly classified as bladder, caused by

Table 1. Classification results of our method compared to a conventional SVM (SVM) and SVM applied to whole-body image partitions (P+SVM)

		Prediction (%)					
Methods (Overall)	Ground Truth	Other	Brain	Bladder	Heart	L.Kidney	R.Kidney
SVM **(79.93%)**	Other	**89.43**	-	0.41	1.22	4.47	4.47
	Brain	21.43	**73.81**	-	-	4.76	-
	Bladder	6.25	-	**93.75**	-	-	-
	Heart	40.00	-	-	**57.14**	2.86	-
	L.Kidney	36.99	-	1.37	-	**61.64**	-
	R.Kidney	30.67	-	1.33	-	-	**68.00**
P+SVM **(89.01%)**	Other	**91.87**	-	0.41	5.28	1.22	1.22
	Brain	23.81	**76.19**	-	-	-	-
	Bladder	6.25	-	**93.75**	-	-	-
	Heart	28.57	-	-	**71.43**	-	-
	L.Kidney	8.22	-	-	-	**91.78**	-
	R.Kidney	8.00	-	1.33	-	-	**90.67**
Our Method **Grouping+P+SVM** **(93.84%)**	Other	**93.90**	1.22	0.81	0.81	0.81	2.44
	Brain	9.52	**90.48**	-	-	-	-
	Bladder	6.25	-	**93.75**	-	-	-
	Heart	2.86	-	-	**97.14**	-	-
	L.Kidney	5.48	-	-	-	**94.52**	-
	R.Kidney	4.00	-	2.67	-	-	**93.33**

Table 2. Classification results using SIFT, HOG or RLT image features alone

	Prediction (%)					
Feature (Overall)	Other	Brain	Bladder	Heart	L.Kidney	R.Kidney
SIFT (84.10%)	80.08	97.62	93.75	97.14	76.71	86.67
HOG (85.69%)	93.09	90.48	93.75	85.71	61.64	78.67
RLT (88.67%)	93.90	78.57	78.13	71.43	90.41	88.00

sites of disease that involved the kidneys but this is a rare occurrence since the diseased regions would need to have similar contextual and spatial features from PET-CT.

We assessed the importance of the individual image features in the classification of NEURs by applying our method with only a specific feature of SIFT, HOG or RLT. In the results in Table 2, individual feature resulted in better classification of certain structures; indicating that heart can be better represented by SIFT and kidneys by RLT features for instance. When compared to the combined features in Table 1, the combination was able to make best use of the properties from all feature extraction algorithms.

4 Conclusion

In this study, we propose a new multi-stage classification method to classify and label regions of FDG excretion and normal uptake automatically from

PET-CT images. Our experiments with 33 clinical lymphoma PET-CT cases demonstrated that our approach had higher accuracy when compared to conventional methods. We suggest our approach will improve image interpretation and visualization.

References

1. Bi, L., Kim, J., Wen, L., Feng, D.D.: Automated and robust percist-based thresholding framework for whole body pet-ct studies. In: EMBC 2012, pp. 5335–5338. IEEE (2012)
2. Chang, C.-C., Lin, C.-J.: Libsvm: a library for support vector machines. ACM TIST 2(3), 27 (2011)
3. Criminisi, A., Shotton, J., Bucciarelli, S.: Decision forests with long-range spatial context for organ localization in ct volumes. In: MICCAI Workshop on Probabilistic Models for Medical Image Analysis (2009)
4. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, vol. 96, pp. 226–231 (1996)
5. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE. T Pattern. Anal. 32(9), 1627–1645 (2010)
6. Freudenberg, L., Antoch, G., Schütt, P., Beyer, T., Jentzen, W., Müller, S.P., Görges, R., Nowrousian, M.R., Bockisch, A., Debatin, J.F.: Fdg-pet/ct in restaging of patients with lymphoma. Eur. J. Nucl. Med. Mol. I. 31(3), 325–329 (2004)
7. Hu, S., Hoffman, E.A., Reinhardt, J.M.: Automatic lung segmentation for accurate quantitation of volumetric x-ray ct images. IEEE. T. Med. Imaging. 20(6), 490–498 (2001)
8. Kakar, M., Olsen, D.R.: Automatic segmentation and recognition of lungs and lesion from ct scans of thorax. Comput. Med. Imag. Grap. 33(1), 72–82 (2009)
9. Kim, J., Hu, Y., Eberl, S., Feng, D., Fulham, M.: A fully automatic bed/linen segmentation for fused pet/ct mip rendering. In: Society of Nuclear Medicine Annual Meeting Abstracts, vol. 49, p. 387. Soc. Nuclear Med (2008)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision. 60(2), 91–110 (2004)
11. Pescia, D., Paragios, N., Chemouny, S.: Automatic detection of liver tumors. In: ISBI 2008, pp. 672–675. IEEE (2008)
12. Song, Y., Cai, W., Kim, J., Feng, D.D.: A multistage discriminative model for tumor and lymph node detection in thoracic images. IEEE. T. Med. Imaging. 31(5), 1061–1075 (2012)
13. Wahl, R.L., Jacene, H., Kasamon, Y., Lodge, M.A.: From recist to percist: evolving considerations for pet response criteria in solid tumors. J. Nucl. Med. 50(Suppl. 1), 122S–150S (2009)
14. Wu, B., Khong, P.-L., Chan, T.: Automatic detection and classification of nasopharyngeal carcinoma on pet/ct with support vector machine. IJCARS 7(4), 635–646 (2012)