Leveraging Mid-Level Semantic Boundary Cues for Automated Lymph Node Detection

Ari Seff, Le Lu, Adrian Barbu, Holger Roth, Hoo-Chang Shin, and Ronald M. Summers

Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD 20892-1182, USA.

Abstract. Histograms of oriented gradients (HOG) are widely employed image descriptors in modern computer-aided diagnosis systems. Built upon a set of local, robust statistics of low-level image gradients, HOG features are usually computed on raw intensity images. In this paper, we explore a learned image transformation scheme for producing higher-level inputs to HOG. Leveraging semantic object boundary cues, our methods compute data-driven image feature maps via a supervised boundary detector. Compared with the raw image map, boundary cues offer mid-level, more object-specific visual responses that can be suited for subsequent HOG encoding. We validate integrations of several image transformation maps with an application of computer-aided detection of lymph nodes on thoracoabdominal CT images. Our experiments demonstrate that semantic boundary cues based HOG descriptors complement and enrich the raw intensity alone. We observe an overall system with substantially improved results ($\sim$78% versus 60% recall at 3 FP/volume for two target regions). The proposed system also moderately outperforms the state-of-the-art deep convolutional neural network (CNN) system in the mediastinum region, without relying on data augmentation and requiring significantly fewer training samples.

1 Introduction

Quantitative assessment of lymph nodes (LNs) is routine in the daily radiological workflow. When measuring greater than 10 mm in short-axis diameter on an axial computed tomography (CT) slice, LNs are generally considered clinically relevant or actionable [13], indicative of diseases such as lung cancer, lymphoma, or inflammation. Manual detection of enlarged LNs, critical to determining disease progression and treatment response, is a time-consuming and error-prone process. Thus, there has been active research in recent years to develop accurate computer-aided lymph node detection (CADe) systems. A challenging object class for recognition, LNs exhibit substantial variation in appearance/location/pose as well as low contrast with surrounding anatomy on CT scans. Recent work on LN CADe has varied according to the feature types and learning algorithms used for training. [1, 8] utilize direct 3D information from CT scans, performing boosting-based feature selection over a pool of 50–60 thousand 3D Haar wavelet features. Due to the curse of dimensionality (analyzed in [14]), such approaches can result in systems with limited sensitivity (e.g. 60.9% at 6.1 FP/scan for mediastinal LNs in [8]). Circumventing 3D feature computation during

LN classification, [14] implements a shallow hierarchy of linear models operating on 2D slices or views of LN candidate volumes of interest (VOIs) with histograms of oriented gradients (HOG) [3] features. Also using 2D (or 2.5D) views, the state-of-the-art performance is reported by [12] via a 5-layer, deep convolutional neural network (70% and 83% sensitivity at 3 FP/scan for mediastinal and abdominal LNs respectively).

In computer vision, edge detection serves as a valuable component in object detection tasks. Originally developed for use with natural images, the state-of-the-art edge detection methods [5,9] exploit the typical structures found in small edge patches such as straight lines and Y-junctions. [5] treats edge detection as a structured learning problem, using a random forest to predict a local edge annotation for each extracted patch from input images. While also using a random forest, [9] instead develops a multi-class classification approach, first clustering patches of ground truth edge annotations to define distinct classes of contours and then attempting to predict the cluster membership of input patches. In this work, our core hypothesis is that we can leverage the output response of semantic LN contour detection (built upon [9]) as mid-level object boundary maps, serving as enhanced input for HOG computation. By linking LN contour detection with LN detection itself, our proposed system will improve as the accuracy of state-of-the-art object contour detection methods improves.

Operating on 2D views (orthogonally sampled slices) of LN candidate volumes of interest (VOIs), our proposed method utilizes radiologist-annotated LN boundaries to first cluster small patches centered on LN boundaries into distinct contour classes. We then train a random forest [2] to classify the contour class membership of extracted LN candidate patches using sketch tokens [9]. Hybrid, mid-level feature maps are constructed by taking the per-voxel sums and maximums of the resulting contour class probabilities. In this manner, HOG is computed both on hybrid feature maps, which contain enhanced semantic objectness cues, and the CT intensity channel. A mixture-of-templates model (separate templates for modeling LNs of different size ranges) is efficiently implemented via a linear SVM, and the resulting 2D view confidence scores are averaged to obtain candidate-level classifications. Our experiments demonstrate that our new method leads to substantially improved performance over intensity-based HOG alone [14] and outperforms the state-of-the-art deep CNN system [12] on mediastinal LN detection, e.g. 78% vs. 70% recall at 3 FP/scan evaluated on the same benchmark data set. Our empirical study shows that HOG, when coupled with enriched hybrid image feature maps, can surprisingly be as effective as deep CNN. To the best of our knowledge, leveraging semantic object-label boundary cues for computer-aided diagnosis has not been previously studied.

2 Methods

Our lymph node detection system assumes we have a set of LN candidates generated within each target region. To facilitate benchmarking, we employ the publicly available LN detection datasets [12, 14]. There are 90 CT scans with ∼1,000/3,200 true/false positive (TP/FP) mediastinal LNs and 86 scans with ∼1,000/3,500 TP/FP abdominal LNs. Multiple TPs may correspond to the same LN. We also follow the view sampling procedure from [14]. For each generated candidate V, we extract 2D views or slices

$\{v_i\}$ of size 45×45 voxels, sufficient to cover the size of most LNs with additional spatial context. Sampling at 0, 1, 2, 3, and 4 voxels away from the candidate centroid bi-directionally in each of the three orthogonal coordinate planes (axial, coronal, and sagittal) yields 27 views $\{v_i\}$ per V. To label the views, we simply transfer the label of V to each v_i: +1 if located inside any LN ground truth segmentation, -1 otherwise.

Defining Lymph Node Contour Classes. Computing our hybrid image feature maps (which will serve as input to HOG) begins with developing a lymph node contour detection system. To this end, we adapt the recent work on sketch tokens [9] to our CT imaging domain. However, in contrast to that work, where the objective is to detect the contours of any object category in natural images, we aim to identify semantic LN boundary contours. The substantial variation of LN shapes implies a wide spectrum of boundary contour appearances. Seeking to capture this wide distribution, we first cluster local LN edge patches into distinct sketch token contour classes. The CT scans in each target region's dataset were examined by a board-certified radiologist[1], who manually segmented any enlarged LNs encountered. Thus for each 2D slice of a CT scan, we have corresponding ground truth tracings of any LN boundaries present (Fig. 1).

Fig. 1. Manual annotation of four abdominal lymph nodes on an axial CT slice.

After VOI decomposition of every LN candidate into 2D views of size 45×45 voxels, we have a corresponding set of binary images $\{S\}$ delineating the manually labeled LN boundaries. Following the notation of [9], we extract patches s of size 15×15 voxels from the images $\in \{S\}$. A patch s_i is extracted if its center voxel is labeled as LN boundary. Approximately 1.7 million such patches are extracted in the training folds during our cross-validation experiments. Daisy descriptors [16] are then computed to compensate for subtle shifts in the manual boundary label placements across CT slices. Next, we perform k-means clustering on the Daisy descriptors, leading to $k = 150$ sketch token classes. Fig. 2 displays example patch cluster means for contours from LNs and colon polyps [15] (shown for comparison). Large variation in the sketch tokens is evident across LNs as well as colon polyps, a smaller-sized object class. Clustering-based labeling attempts to assign LN boundary patches into k classes for better detection.

Contour Detection. After defining the LN contour classes, we aim to detect their presence on candidate LN 2D views. Training labels for 15×15 patches are

[1] The LN 3D segmentation mask datasets will be made publicly available. Visit http://www.ariseff.com for info.

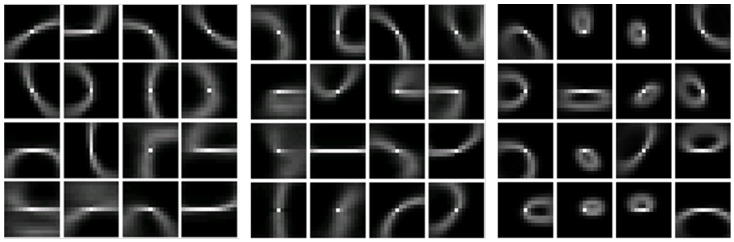

Fig. 2. Examples of sketch tokens learned from the manual tracings by radiologists for mediastinal LNs (left), abdominal LNs (middle), and colon polyps (right).

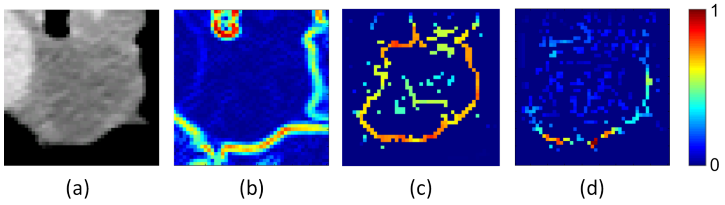

Fig. 3. (a) CT 2D View, (b) gradient transform, (c) $SumMap$, and (d) $MaxMap$ (scaled for illustration), for a true mediastinal LN candidate. Note how the simple gradient transform (b) does not delineate the boundaries of the LN as strongly as the $SumMap$ (c) derived from the supervisedly learned mid-level contour detection.

assigned as follows: If centered on a boundary pixel, patches are labeled according to their sketch token cluster membership (out of k choices); otherwise, they are labeled as negative. Similarly to [9], we compute multiple feature channels per patch [4]. These include 3 gradient magnitude channels using Gaussian blurs of $\sigma = 0$, 1.5 and 5 pixels and 8 oriented gradient channels. Because CT images are grayscale, we refrain from computing the CIE-LUV color space channels which would be relevant for natural images. Self-similarity features, useful for detecting texture-based contours, are computed on each gradient channel over a 5×5 grid leading to $\binom{5\cdot5}{2}=300$ features per channel. Thus, for a 15×15 patch, we have $15\cdot15\cdot11=2475$ channel features and $300\cdot11=3300$ self-similarity features for a total of 5775 features per patch.

We train a random forest, an efficient method for multi-class classification, to detect the $k+1$ LN contour classes [2]. Randomly sampling 1,000 patches per positive sketch token class and 2 negative patches per training image provides a decent balance between positive and negative training samples for each decision tree. 25 trees are trained whose leaf nodes denote the probability of a patch belonging to each class. Each tree uses a randomly selected subset of size $\sqrt{F}$ from F total available features for training.

Classification Using Boundary Input for HOG. A set of k sketch token class probability values are evaluated at every pixel for each 2D CT view. We construct the following mid-level, semantic representations as subsequent input for HOG computation. The first representation we compute is the sum of the sketch token probabilities at each pixel in an image. Such a map can be interpreted as the total positive probability of each pixel residing on a true lymph node boundary. We also compute a map representing the maximum sketch token probability at

each pixel because any true boundary pixel should fit well into at least one of the 150 contour classes (the reason for clustering positives into $k = 150$ classes in a "divide and conquer" manner). Letting t_{ij} denote the probability that a patch centered at pixel i belongs to a particular contour class j, and t_{i0} the probability of the negative background class, we derive the following two boundary probability cue maps: $SumMap_i = \sum_{j=1}^{k} t_{ij} = 1 - t_{i0}$; $MaxMap_i = \max_{1 \le j \le k} t_{ij}$ where k is the number of sketch token classes. Fig. 3 shows these learned feature maps for a mediastinal LN candidate. Compared to a simple image gradient transform, $SumMap$ more accurately highlights the LN's boundary.

HOG can now be computed on each derived feature map in addition to the raw intensity CT image. The HOG descriptor divides an input image into square cells and delineates the quantized distribution of local intensity gradient magnitudes and orientations for each cell. 31 features are calculated per cell [7], which are then normalized within blocks of adjacent cells. Using the same parameters as [14], the 45×45-pixel 2D views are divided into square 5×5-pixel cells, yielding 25 cells and $25 \cdot 31 = 775$ features for each map. We test various concatenations of these feature sets in Sec. 3 for performance evaluation. For robust linear classification (non-linear kernels exhibit poor generalization with limited datasets), we train an L2-regularized, L2-loss linear SVM [6], treating each 2D view as an independent instance and averaging their confidence scores to obtain the candidate-level predictions.

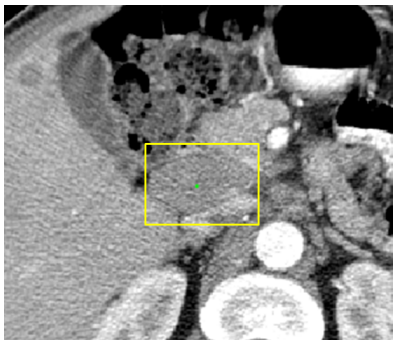

Fig. 4. A large abdominal lymph node that the single template model misses, but the mixture model detects. Larger LNs are especially clinically relevant.

Mixture-of-Templates Model by Size Gating. Enlarged LNs can vary greatly in size, reaching as large as 55 mm in short-axis diameter in the abdominal LN dataset. Although increasingly rare above 20 mm, very large LNs are especially clinically relevant. Thus it is crucial that LN CADe accurately identifies them. A single template of "HOG + Hybrid input" approach (modeling all LNs of varying sizes) will favor the detection of moderately enlarged LNs which are more common. Fig. 4 shows a typical large abdominal LN missed by a single template approach. Addressing this imbalance in the training/testing datasets, we extend our model by training two classifiers via a variation of size gating [11]. With a 15 mm size threshold (calibrated as the median ground truth LN size), one classifier is trained using all positives linked to LNs ≥ 15 mm and another is trained with the rest. The negative candidate set does not change. In testing, confidence scores output by each size-gated SVM are first scaled according to the corresponding range of training scores, making the classifiers' scores more comparable. For any instance, the mixture-of-templates model then reassigns the maximum of the two scaled scores as its final confidence. No LN size information is required in testing.

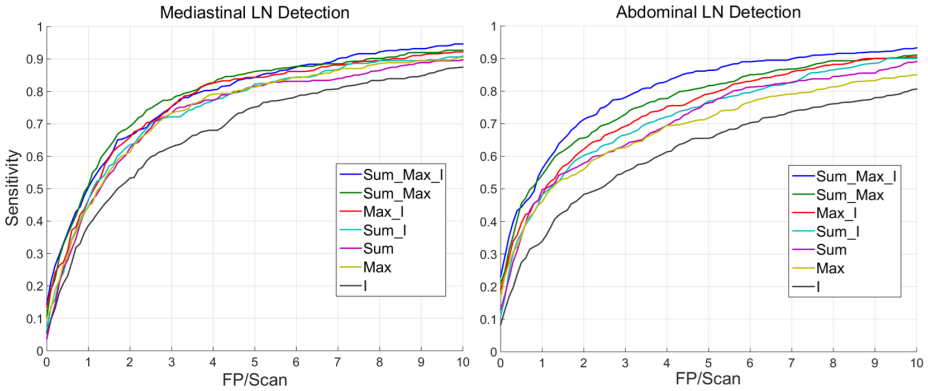

Fig. 5. Performance comparison of LN detection models trained on the seven integrated feature sets. For example, "Sum_Max_I" indicates the model trained on concatenated $HOG(SumMap)$, $HOG(MaxMap)$ and $HOG(Intensity)$ features. Six-fold cross-validation FROC curves are shown for both the mediastinal (left) and abdominal (right) target regions.

3 Evaluation & Discussion

Data & Protocol. To facilitate comparisons with other work, we evaluate our methods on the publicly available lymph node CT datasets used by [12, 14]. There are 90 patients with 389 mediastinal LNs, and 86 patients with 595 abdominal LNs. We train and test models for each target region separately. Performing a six-fold cross-validation for the combined LN contour detection/LN detection, we randomly split each group of patients into 6 disjoint sets. For each fold, models are trained on five sets and and tested on the remaining set. Training the contour detection random forest (trees are parallelized) and subsequent linear SVM for a single fold takes $\sim$40 minutes. Testing on a single patient scan, including 2D view sampling and feature computation (not counting candidate generation), takes less than 5 seconds.

Performance. The three feature sets, $HOG(SumMap)$, $HOG(MaxMap)$ and $HOG(Intensity)$, are evaluated as single template models using all seven possible feature set integrations (Fig. 5) with the free-response operating characteristic (FROC). All six feature integrations that include at least one boundary cue map outperform HOG on raw intensity alone, in the full range of the FROC curves. The top performing integrations at low FP rates, Sum_Max for the mediastinum and Sum_Max_I for the abdomen, exhibit 24%–39% greater recall than the baseline HOG (e.g. 78% versus 63% at 3 FP/scan for mediastinal LNs; 78% versus 56% at 3 FP/scan for abdominal LNs). Furthermore, this performance is comparable to the state-of-the-art deep learning results [12], moderately outperforming in the mediastinum while only slightly lower for the abdomen. In detail, comparing with [12], we achieve sensitivities of 78% vs. 70% at 3 FP/scan and 88% vs. 84% at 6 FP/scan in the mediastinum, and sensitivities of 78% vs. 83% at 3 FP/scan and 89% vs. 90% at 6 FP/scan in the abdomen. The mixture-of-templates models are also evaluated using the top performing feature sets calibrated from the single template models. Fig. 6 shows

the improvement in large malignant LN detection when the mixture model is used in the abdomen, e.g., 94% vs. 78% sensitivity at 6 FP/scan for LNs > 20 mm. We observe similar performance improvement for large mediastinal LNs when the mixture-of-templates model is employed.

Discussion. The proposed method significantly outperforms the recent work using HOG with a CT intensity map alone [14] which clearly demonstrates the merits of utilizing semantic object-level boundary cues for automated LN detection. This improvement is at the cost of annotated LN segmentation, required only at training and not in testing. The sketch tokens object boundary detector [9] is very robust and generalizable at a 15×15-pixel patch scale. The more recent structured forest edge detector [5] can be exploited as well. Comparing with the state-of-the-art deep CNN representation [12], our overall system is also a multi-layer pipeline with comparable/moderately better FROC curves in abdominal/mediastinal LN detection, respectively. The dense pixel-level semantic object boundary response map is especially critical for the performance gain over [12, 14], but is non-trivial for a deep CNN, trained for direct LN recognition, to implement. CNNs are still mostly decision/classification models. While the newest fully convolutional neural networks can compute the output class support probability map, it is at a coarse (10–20$\times$ downsampled) spatial resolution [10] (thus not sufficient in our scenario). Instead we plan to investigate the feasibility of using our multi-channel hybrid image feature maps for direct CNN training as future work.

4 Conclusion

We propose a novel method to leverage hybrid image feature maps based on mid-level object boundary cues for computer-aided lymph node detection. The learned maps can be used in place of or in addition to raw CT intensity images as input to HOG feature computation. Evaluation of our approach for LN detection in two target regions demonstrates that the mid-level information supplied by the new representations both enhances and complements typical intensity-based HOG for this complex object recognition task. Our method achieves substantially improved results over baseline HOG systems [14] and moderately outperforms the state-of-the-art deep CNN system [12] in mediastinal LN detection.

Acknowledgments. This work was supported by the Intramural Research Program of the National Institutes of Health Clinical Center.

References

1. Barbu, A., Suehling, M., Xu, X., Liu, D., Zhou, S.K., Comaniciu, D.: Automatic detection and segmentation of lymph nodes from CT data. IEEE Trans. Med. Imaging 31(2), 240–250 (2012)
2. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. vol. 1, pp. 886–893 (2005)

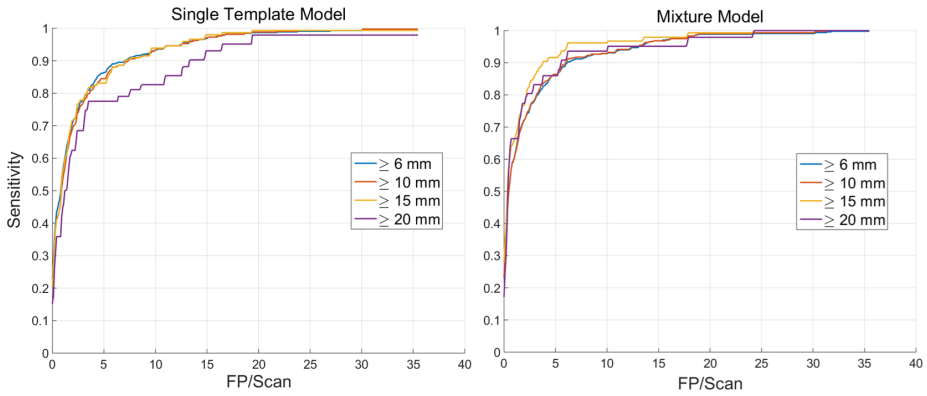

Fig. 6. Performance comparison of the single template model (left) and mixture-of-templates model (right) on abdominal LN detection. Note the substantially improved detection of malignant LNs greater than 20 mm in short-axis diameter by the mixture model.

4. Dollar, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: Proc. BMVC. pp. 1–11 (2009), doi:10.5244/C.23.91
5. Dollár, P., Zitnick, C.L.: Structured forests for fast edge detection. In: ICCV. pp. 1841–1848. IEEE (2013)
6. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
7. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. on Pat. Ana. and Mach. Intell. 32(9), 1627–1645 (2010)
8. Feulner, J., Zhou, S.K., Hammon, M., Hornegger, J., Comaniciu, D.: Lymph node detection and segmentation in chest CT data using discriminative learning and a spatial prior. Medical Image Analysis 17(2), 254–270 (2013)
9. Lim, J.J., Zitnick, C.L., Dollár, P.: Sketch tokens: A learned mid-level representation for contour and object detection. In: CVPR. pp. 3158–3165. IEEE (2013)
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CoRR abs/1411.4038 (2014)
11. Lu, L., Bi, J., Wolf, M., Salganicoff, M.: Effective 3d object detection and regression using probabilistic segmentation features in ct images. In: CVPR. pp. 1049–1056. IEEE (2011)
12. Roth, H., Lu, L., Seff, A., Cherry, K., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.: A new 2.5d representation for lymph node detection using random sets of deep convolutional neural network observations. In: MICCAI. pp. 520–527 (2014)
13. Schwartz, L., Bogaerts, J., Ford, R., Shankar, L., Therasse, P., Gwyther, S., Eisenhauer, E.: Evaluation of lymph nodes with recist 1.1. Euro. J. of Cancer 45(2), 261 – 267 (2009)
14. Seff, A., Lu, L., Cherry, K., Roth, H., Liu, J., Wang, S., Hoffman, J., Turkbey, E., Summers, R.: 2d view aggregation for lymph node detection using a shallow hierarchy of linear classifiers. In: MICCAI. pp. 544–552 (2014)
15. Summers, R.M., Yao, J., Pickhardt, P.J., Franaszek, M., Bitter, I., Brickman, D., Krishna, V., Choi, J.R.: Computed tomographic virtual colonoscopy computer-aided polyp detection in a screening population. Gastroenterology 129, 1832–1844 (2005)
16. Tola, E., Lepetit, V., Fua, P.: DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo. IEEE Trans. on Pat. Ana. and Mach. Intell. 32(5), 815–830 (2010)