Computer Aided Diagnosis Using Multilevel Image Features on Large-Scale Evaluation

Le Lu, Pandu Devarakota, Siddharth Vikal, Dijia Wu, Yefeng Zheng, and Matthias Wolf

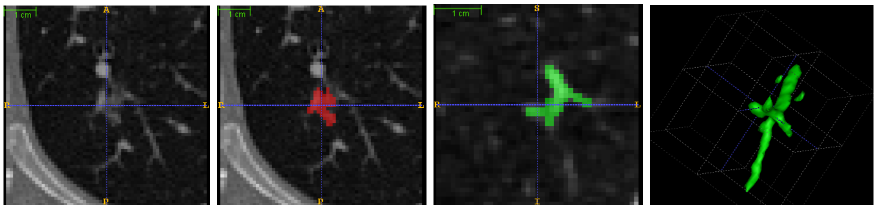
Siemens Medical Solutions USA; Siemens Corporate Research
(le.lu@nih.gov; yefeng.zheng@siemens.com; mwolf@siemens.com)

Abstract. Computer aided diagnosis (CAD) of cancerous anatomical structures via 3D medical images has emerged as an intensively studied research area. In this paper, we present a principled three-tiered image feature learning approach to capture task specific and data-driven class discriminative statistics from an annotated image database. It integrates voxel-, instance-, and database-level feature learning, aggregation and parsing. The initial segmentation is proceeded as robust voxel labeling and thresholding. After instance-level spatial aggregation, extracted features can also be flexibly tuned for classifying lesions, or discriminating different subcategories of lesions. We demonstrate the effectiveness in the lung nodule detection task which **handles all types of solid, partial-solid, and ground-glass nodules using the same set of learned features**. Our hierarchical feature learning framework, which was extensively trained and validated on large-scale multiple site datasets of 879 CT volumes (510 training and 369 validation), achieves superior performance than other state-of-the-art CAD systems. The proposed method is also shown to be applicable for **colonic polyp detection, including all polyp morphological subcategories**, via 770 tagged-prep CT scans from multiple medical sites (358 training and 412 validation).

1 Introduction

Lung cancer is the leading deadly cancer in western population, but similar to colon cancer, it is highly preventable if lung nodules can be detected early. Therefore, image-interpretation-based cancer detection using 3D computer tomography (CT) has emerged as a common clinical practice, and computer-aided detection tools for enhancing radiologists' diagnostic performance and effectiveness have been developed in the last decade. The key for radiologists to accept the clinical usage of a CAD system is the highest possible true positive (TP) detection sensitivity with the desirably low false positive (FP) rate per patient. In this paper, we exploit a new method of multilevel (discriminatively trained) image feature learning, as the key to achieve this goal.

Many CAD algorithms highly depend on delicate lesion image segmentation algorithms to delineate the boundary of lesion tissue from its normal context surroundings in 2D/3D images. A collection of drastically different methods [1–3] have been proposed. The technical analogies of all methods are based on *analyzing low-level, surface geometric and volumetric intensity patterns*, and *exploiting strong spatial regularization* (e.g., prior shape fitting, Markov-Gibbs random field) to optimize the binary segmentation accuracy. On the other end, [4] utilizes analytical shape and appearance priors



(a) (b) (c) (d)

Fig. 1. Examples of training data and voxelwise annotation (Red for nodule; Green for vessel). (a) CT image displaying a Ground-Glass nodule; (b) its corresponding annotation; (c) CT image of a branching vessel with its annotation overlaid; (d) 3D volume rendering of the same structure to demonstrate its spatial complexity.

and Markov-Gibbs random field; [1] designs elaborate region-growing criteria separating nodule growth from normal tissues; and [3] empowers morphological approaches and convexity models. CAD detection bias or dependency on segmentation (e.g., under- or over-segmentation often occurred, segmentation failures) may not be desirable as discussed in [5].

A Bayesian voxel labeling approach for lung nodule detection was thus proposed [5], avoiding explicit segmentation. Nevertheless, their four types of probabilistic formulations of nodule, vessel, vessel junction and outlier are based on medical literature description or general knowledge, but not verified from a large amount of data. For example, the nodule model is chosen as a solid ellipsoid with similar concentric ellipsoidal isosurfaces, and the vessel model represents a section of a solid torus with similar concentric isosurfaces. Their detailed model parameters are heuristically decided from common medical prior knowledge, for example, 15mm for the average maximum radius of a pulmonary vessel. The nodule model in [5] is invalid for partial-solid and ground-glass nodules which nevertheless have great importance for CAD to perform well, as it is more ambiguous for radiologists. The solid parenchymal nodule model may also have trouble with other contextual types (e.g., juxta-pleural, pleural tail, and vascular [6]). The Bayesian closed-form mathematical representation of nodule, vessel, vessel junction, and outliers may be questionable whether the analytic models can describe well the tremendous anatomical appearance variations in hospital-size datasets. For example, the vessel junction is assumed to be a bifurcation structure in [5] whereas, from Fig. 1 (d), the vascular anatomy describes four branches.

In this paper, we follow voxel classification framework [7, 8], but our voxel-level labeling is data-driven and statistically learned from the annotated lesion image masks [6,9] on a number of CT scans. Only voxel probability assignment and thresholding (e.g., suppressing background clutter) are employed to obtain a lesion class-probability response map. Descriptive feature extraction based on instance-level spatial aggregation is then performed, without explicit segmentation optimization. Our system is trained using Bayesian multiple instance relevance vector machine (MILRVM) classifier [10] as a building-block through database-level selective training. The validation results of lung CAD on 879 CT scans collected from 10+ medical sites in US, Asia, Europe, are very promising. To the best of our knowledge, we are the first to report a unified, high-performance classification framework of detecting all solid, partial-solid, and ground-glass lung nodules, using the same set of supervisedly learned image features. This approach can be seamlessly applied to colonic polyp detection as well.

Our paper leverages discriminatively trained, higher-order image appearance model to label voxel class (e.g., polyp/nodule/vessel) probability map, with **mostly no segmentation level optimization**. Simple probability thresholding and connected component-based pruning are used. Then we extract a group of statistics or generative features in the dual space of learned probability and CT intensity volumes. These features are effective to fit our new hierarchical classification models, and yet flexible for different tasks. The proposed hierarchical feature learning approach is generic for identifying colon polyp, lung nodule, or vessel problems, in a unified and principled data-driven manner. We validate its effectiveness by evaluating the impacts on large-scale colon and lung CAD system performances (879 and 770 volumes respectively). The results are very encouraging and significantly outperform the recent state-of-the-arts [1, 2, 5, 11–15].

2 Materials and Methods

Candidate Generation: CAD systems generally contain two stages: ROI (region of interest, or instance) candidate generation, and ROI feature learning & classification. Candidate generation (CG) is to rapidly identify anomalous regions with high sensitivity (e.g., $> 99\%$) but low specificity, e.g. $150 \sim 200$ candidates per scan, with typically zero to several true positives, in screening population. For convenience of comparison, we use the standard local shape index [16] method for extracting polyp candidates, and an efficient multiscale 3D Difference of Gaussians (DOG) filter to find potential nodule instances. In this paper, our focus is the later phase. The proposed method consists of voxel-level, instance-level, and database-level hierarchical feature learning, spatial aggregation, and final classification.

2.1 Supervised Probabilistic Voxel Map Labeling $\wp$ in ROI

Nodule, vessel, and polyp features were learned from 209, 56, and 427 training instances (i.e., voxel-annotated subvolumes in 3 dimensions of $41 \sim 83$ voxels), respectively. Examples are illustrated in Fig. 1 and Fig. 2. Training data represent a vast variety of lesions and vessels, with different intensity patterns, shape morphologies, adjacent contextual structures, and sizes ($\geq 3mm$, highly correlated with actionability of a given lesion), as long as they are clinically relevant. CT images are isotropically resampled with the resolution of $1mm$. The training nodule or vessel instances are first cropped as subvolume ROI from the whole CT volume, based on ground-truth. Three clinical experts were asked to manually segment nodule and vessel using a “voxel-painting” tool where **Red** represents nodule and **Green** for vessel voxels. Polyp segmentation is defined as extracting the polyp surface area, via a closed 3D curve boundary on a colonic wall surface. A semi-automatic annotation tool was built for editing computer generated polyp segmentation contours, similar to [17]. The colon isosurface can be returned by running the Marching Cubes algorithm [18] to separate colon lumen from soft-tissue. After annotation, we obtained a set of labeled or masked voxels in each instance ROI for lesion or vessel. In total, there are 122215, 60335, or 190638 voxels to be treated as positive samples for training class probability of nodule, vessel, or polyp, respectively. Unmasked

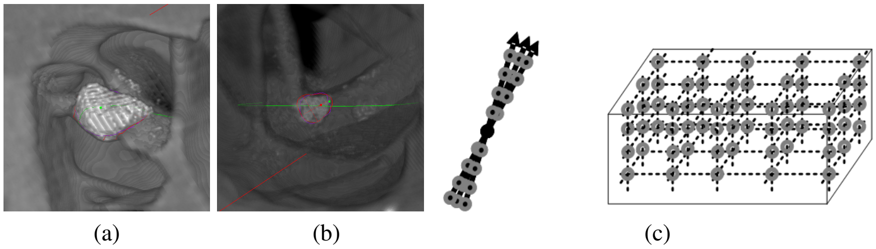


Fig. 2. (a,b) Examples of two polyps with annotated 3D contours and obtained probability map (brighter intensity indicates higher probability). Labeling noise is observed. (c) Steerable sampling grid patterns of labeling voxels on 3D surface (colon polyp) or in 3D volume (lung nodule and vessel) [19].

voxels located at least $3mm$ away from labeled ones were initially treated as negatives, with resampling. The numbers of negative training samples are in general $4 \sim 6$ times the sizes of positive sets.

Training & Features: We build the voxel-level classifier using probabilistic boosting tree (PBT[1]) training [20], coupled with 3D axis-sampled steerable features for polyp *surface voxels*, and 3D box-sampled feature patterns centered at nodule/vessel *volume voxels*. This step returns three classifiers $\{PBT_n; PBT_v; PBT_p\}$ for nodule/vessel/polyp, respectively. $\{PBT_n; PBT_v; PBT_p\}$ normally have $4 \sim 5$ layers of internal nodes with $160 \sim 240$ features selected. For details of image features for training of *surface voxels* and *volume voxels*, refer to [17, 19]. PBT training is assisted by cross-validation-based model selection on determining the tree depth. Note that the bootstrapping strategy on finding hard negative samples and retraining is not found particularly helpful.

Curvature features have demonstrated to be very helpful for parsing *surface voxels* [11,12,17], but appear not to provide much additional information gain on classifying *volume voxels*, especially in the case of nodule voxel detection. The other reason is that the previous work only focuses on solitary solid nodules [5, 15] to show that curvature may be useful, whereas we train a single classifier PBT_n to handle all three types of nodules, under various anatomical contexts [6, 23]. Partial solid and ground-glass can have very weak, noisy, and non-informative curvature features. Not computing Gaussian/principal curvatures in 3D volume space also improves the computational efficiency. The per-grid feature pool number drops from 71 to 23 for boosting feature selection.

Testing & Pruning: 1), In runtime testing, for each given 3D lesion candidate ROI obtained by a candidate generation process (which is common in CAD pipeline [3, 12, 17]), we exhaustively assign each ROI voxel v with its class probability value $\wp \in [0, 1]$ by evaluating either one of $\{PBT_n; PBT_v; PBT_p\}$. Then we generate the label map for foreground (nodule, vessel, or polyp) voxels versus background ones, by simple thresholding on $\wp$-field: $L = 1$, if $\wp > \tau$; and $L = 0$, otherwise. From the training Receiver Operating Characteristic (ROC) performance of $\{PBT_n; PBT_v; PBT_p\}$, we can select

[1] PBT is a powerful two-class and multiclass discriminative learning framework [20]. Random Forests or Ferns [21,22] are also applicable for training voxel-level labeler. Our empirical experience shows that PBT can learn very similar or slightly better ROC curves with much simpler model complexity, i.e., one tree versus multiple trees per model/classifier.

the respective thresholds $\{\tau_n = 0.23; \tau_v = 0.45; \tau_p = 0.24\}$ to hold 98% sensitivity. 2), Next, a fast connected component algorithm (26-neighborhood) is used to partition the L-field into separate clusters $\{C\}$. The cluster with the largest $\sum(\wp)$ from its support L map are kept (denoted as $\mathbb{S}$ where $L = 1$) and $L = 0$ is set for all remaining voxels, assuming there is only one dominating structure from each CAD candidate ROI. 3), After connected component base non-maximum pruning, we can effectively remove false positive responses while keeping high sensitivity. Based on our empirical evaluation, highly optimized segmentation procedures [6, 11, 17, 24] may not improve the learned features with significantly better discriminativeness. For the detection purpose, our modeling of the voxel-level unary energy term (in a CRF sense) appears mostly sufficient. However the rough segmentation accuracy by detection is statistically lower than [17, 24], e.g., for size measurement (tuned by τ towards over-segmentation for high detection sensitivity).

2.2 Spatial Aggregation Image Features (SAIF) in ROI

Voxels of different lesion types are mapped into the same universal $\wp$ space. Each voxel can be represented as a tuple of $(\wp, \upsilon, x_v, y_v, z_v)$ as probability, intensity, and spatial location. We compute the following SAIF feature groups per region of interest (ROI).

Statistics of Class Probability $\{\wp\}$ and Intensity $\{\upsilon\}$:(9) This feature group computes five overall statistics of $\{\wp\}$ in $\mathbb{S}$: $Prob_{Sum}$ is the sum of polyp-class posterior probabilities within segmentation $\sum\{\wp\}$; $Prob_{Avg}$ is the corresponding average probability $Prob_{Sum}/|\mathbb{S}|$, and its second to fourth order moments (i.e., standard deviation $Prob_{Std}$, skewness $Prob_{Skw}$ and kurtosis $Prob_{Kts}$). Similarly, we compute the $1st \sim 4th$ order moments for the set of intensity distribution $\{\upsilon\}$.

3D Ellipsoid Shape Descriptor:(10) For the 3D voxel mass $\mathbb{S}$ per ROI, we first estimate its centroid and covariance matrix in *volumetric* coordinates:

$$[\bar{x}, \bar{y}, \bar{z}] = \frac{\sum_{\mathbb{S}}[x_v, y_v, z_v] \times \wp}{\sum_{\mathbb{S}} \wp} \tag{1}$$

$$\text{CoMat} = \frac{\sum_{\mathbb{S}}(\varDelta X)^T(\varDelta X) \times \wp}{\sum_{\mathbb{S}} \wp} \tag{2}$$

where $\varDelta X = [x_v, y_v, z_v] - [\bar{x}, \bar{y}, \bar{z}]$ Then, Singular Value Decomposition is used to calculate three Eigen-values of CoMat: R_1, R_2, R_3 that geometrically maps to the three radii if fitting the mass of $\mathbb{S}$ as an ellipsoid. The covariance matrix CoMat models the *3D volumetric spatial distribution* of underlying lesion or vessel confidence/probability in 3D CT images. Apart from standard Ellipsoid fitting, $\wp$ is used as a weight factor in Eq. 1,2, to reflect per-voxel class probability. Assuming $R_1 \geq R_2 \geq R_3$, six other features $(R_1 \times R_2 \times R_3, (R_1 \times R_2 \times R_3)^{1/3}, R_1 \times R_2, R_1/R_2, R_2/R_1, R_1/R_3, R_3/R_1)$ are computed for feature expansion purpose. More sophisticated 3D shape features, such as "plateness", "stickness" and "ballness" [25], do not have superior performance than our ellipsoid based shape descriptor, from our empirical evaluation.

Multiscale Intensity Histogram Features:(16) By using the $\wp$-weighted covariance matrix CoMat, we search all voxels $\{[\upsilon, x_v, y_v, z_v]\}$ within the $\sqrt{M}$ *Mahalanobis* distance, originating from the ellipsoid centroid $[\bar{x}, \bar{y}, \bar{z}]$, i.e.,

$$\text{MHD}(\upsilon) = (\varDelta X)(\text{CoMat})^{-1}(\varDelta X)^T \leq M \tag{3}$$

where $\varDelta X = [x_v, y_v, z_v] - [\bar{x}, \bar{y}, \bar{z}]$. M is set as $2, 4, 6, 8$, corresponding to the fitted 3D object ellipsoids of multiple spatial scales and keeping their radii aspect ratios. A domain-knowledge based CT intensity binning of $[0, 350); [350, 950)$; $[950, 1100)$;$[1100, 4095]$ is used to construct an intensity histogram $IH_k, k = 0, 1, 2, 3$ for each ellipsoid. Binning stands for air, soft tissue, fat and bone structures. Thus a total of 16 MIH features are calculated (4 bins by 4 scales) to model the intensity patterns in multiscale contexts. Deriving histograms in the image *Gradient* domain is also feasible [26] (left for future work).

Boundary Gradient-Shape Statistics Features:(12) Given the $\wp$-thresholded voxel set (as a 3D point cloud) in the ROI, we first extract all boundary voxels $\{b\}$ with neighbors both inside and outside (i.e., $L = 1$ or 0). 26-neighborhood is used. The following three measurements are then computed.

$$NDist(b) = \|[x_b, y_b, z_b] - [\bar{x}, \bar{y}, \bar{z}]\| / \boldsymbol{D} \tag{4}$$

$$NGrad(b) = \|\nabla(x_b, y_b, z_b)\| / \boldsymbol{G} \tag{5}$$

$$Ori(b) = \frac{\nabla(x_b, y_b, z_b)}{\|\nabla(x_b, y_b, z_b)\|} \circ \frac{[x_b, y_b, z_b] - [\bar{x}, \bar{y}, \bar{z}]}{\|[x_b, y_b, z_b] - [\bar{x}, \bar{y}, \bar{z}]\|} \tag{6}$$

where $\boldsymbol{D}$ is the maximum of $\|[x_b, y_b, z_b] - [\bar{x}, \bar{y}, \bar{z}]\|$ in the boundary voxel set $\{b\}$; $\boldsymbol{G}$ is the maximum $\|\nabla(x_b, y_b, z_b)\|$ from training. Therefore distance $NDist(b)$ and gradient $NGrad(b)$ measurements are normalized $\in [0, 1]$ and scale-invariant. $Ori(b)$ is the dot-product ($\in [-1, +1]$) of the centroid-to-boundary direction and its local gradient direction which encodes boundary shape information. From $\{b\}$, we treat $\{NDist(b)\}$, $\{NGrad(b)\}$, $\{Ori(b)\}$ as three empirical distributions and compute their global statistics of $1st \sim 4th$ order moments as features. These boundary gradient-shape features are related to 2D/3D ray features [27] describing irregular shapes, but more compact (without concatenating dozens of sampling directions) and truly rotation invariant for robustness.

Our spatially aggregated image features capture the joint intensity/shape and class-conditional probability $(\wp, \upsilon)$ statistics of all ROI voxels, which has not been explicitly addressed previously [1,2,5,8,12,14]. These features model the high range/order spatial interactions among voxels within ROI, as an empirical distribution. Features are also translation and rotation invariant. **Most importantly, after Sec. 2.1, SAIF definition is universal for different types of lesions (e.g., polyp and nodule) and handle all their subtypes represented in training.**

2.3 Flexible Learning on Detection using Soft Categorization

We design flexible tree-based classification hierarchies to optimize the final CAD detection performance, using image features computed for (+/-) ROIs. Bayesian multiple instance relevance vector machine (MILRVM) [10] is adopted as the main building-block on classification of ROI features and soft-categorization. MILRVM is validated with good classification accuracy and generality, with feature selection and linear fusion.

Subcategory Soft-Gating: We propose a two-layer hierarchical classification architecture, called "soft-gating" framework in a Bayesian "divide-and-conquer" or mixture-of-experts setting. 1), A basic MILRVM [10] classifier is trained using *Sessile+Pedunculated* (SP) polyps versus *Flat* polyps in colon CAD; or *Wall-attached* (WA) versus *Nonwall-attached* (NWA) nodules in lung CAD, respectively. The main idea is that polyps under different shape morphologies present large within-class variations in feature space; nodules attached to a lung peripheral wall are expected to show different characteristics [6] when compared to isolated nodules. By handling these true positive lesion can-

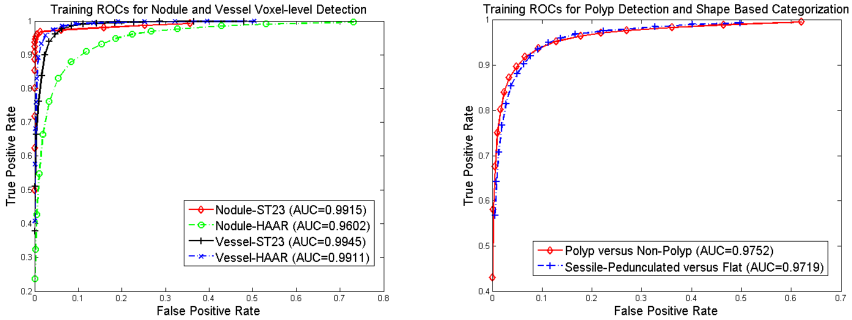


Fig. 3. ROC Curves of voxel-level learning: comparison of nodule and vessel voxel detection using 3D steerable features or Haar features with PBT training [20] **(Left)**; polyp versus non-polyp classification; Sessile-Pedunculated versus Flat polyp based shape categorization **(Right)**.

didates separately according to the defined (sub)categorization, we can obtain a more robust classifier. Learning these category attributes may be partially informative for further lesion malignancy diagnosis. Similar Area-Under-Curve (AUC) measurements of ROC curves are obtained, by comparing polyp versus non-polyp voxel classification (AUC=0.9752) or Sessile-Pedunculated versus Flat polyp categorization (AUC=0.9719). 2), The leaf classifier is trained using the annotated lesions of the targeted category only (from ground-truth) versus all negatives. This binary tree classification framework is illustrated in Fig. 4 (**Right**). Shallow tree is used here due to already hierarchically learned and strongly informative SAIF features and MILRVM.

In runtime, soft-gating means that all candidates will be passed into both left and right leaf classifiers to be evaluated, yet with a different category probability from the gating classifier (in contrast to hard-gating where candidates only run into either one of the tree branches.), such as $Pb(SP)$, $Pb(Flat)$, $Pb(W)$ and $Pb(NW)$. There is no need to normalize the pairs of weights (e.g., $Pb(SP) + Pb(Flat) = 1$) since we expect negative candidates to get low probabilities on both $Pb(SP)$ and $Pb(Flat)$. Finally, the probability of being “Polyp” or “Nodule” for any candidate is obtained in a Bayesian fusion manner:

$$Pb(Polyp) = Pb(SP) \times Pb_L(Polyp|SP) + Pb(Flat) \times Pb_R(Polyp|Flat) \quad (7)$$

$$Pb(Nodule) = Pb(W) \times Pb_L(Nodule|W) + Pb(NW) \times Pb_R(Nodule|NW) \quad (8)$$

The purposes of having a classifier which can distinguish different polyp categories (commonly Sessile and Pedunculated versus Flat polyps) are two folds: shape morphology labeling is helpful for radiologist decision making; and more importantly, it can help to have a good gate classifier dividing the whole polyp candidate population into different leaf branches, against their shape characteristics. Leaf classifiers are further trained for each branch. Combining the gate and leaf classifiers, a formed classification hierarchy with better classification accuracy and generality is normally expected, in the spirit of divide-and-conquer. Two schemes are performed and evaluated: (1) We break the positive pool of PBT_1 into two parts, according to the shape morphology labels of their rooting polyps as (Sessile and Pedunculated, +), or (Flat, -); then PBT_1^G is trained and its affiliated FPR features are used for the system Gate RVM classifier. (2) We enhance the negative training set of PBT_1^G by adding all negatives of PBT_1 for a new negative set (i.e., non-polyp voxels and flat polyp voxels). By keeping the positive training set of PBT_1^G, we train PBT_2^G which gets the generality of "gating FPR features" under control, against non-polyp voxels. Therefore, ideally only Sessile and Pedunculated polyps have high response; whereas flat polyps and polyp FPs receive low scores. An example is shown in Fig. 3.

Selective-Training False Positive Filter: The last stage of classification can be followed with an anatomy-based false positive (FP) filter classifier. We construct a new MILRVM classifier using all training nodules versus the most difficult FPs of vessels and vessel bifurcation or branches (Fig. 1 (d)) in lung CAD. Denote that the lesion (+) probability from soft-gating is $Pb(+)^c$ and $Pb(+)^f$ for FP filter. We combine them in Eq. 2.3 and obtain the final $Pb(+)$ to be thresholded for free ROC analysis.

$$Pb(+) = 1 - (1 - Pb(+)^c)(1 - Pb(+)^f) \tag{9}$$

For instance, a positive candidate with $Pb(+)^c = 0.8$ and $Pb(+)^f = 0.8$ will have final $Pb(+) = 1 - 0.2 \times 0.2 = 0.96$, while a confusing vessel FP of $Pb(+)^c = 0.8$ and $Pb(+)^f = 0.2$ gets $Pb(+) = 0.84$. We gain a new sizable classification margin $0.12 = 0.96 - 0.84$.

3 Experimental Results & Discussion

Data: We learn lung CAD image features and final classifiers based on 510 training CT cases with 717 solid nodules (SN), 124 partial solid nodules (PSN) and 91 ground-glass nodules (GGN), and evaluate the system performance on the validation dataset of 369 volumes including 462 SN, 93 PSN, and 51 GGN, respectively. In colon CAD, we have a total of 770 Tagged-prep CT scans with 239 polyps in the patient level and 416 polyps in the volume level for colon CAD study (120 patient-level or 226 volume-level polyps for validation. Each patient has two scans). Training and validation are split at patient level. Datasets were collected from 10+ hospitals from US, Europe and Asia. Various scanner vendors and screening imaging protocols are used. *We do not have access to datasets from other work that we compare here. However, given the diversified nature of data collection, our datasets are sufficiently representative for performance validation.*

Discriminative Feature Evaluation: We first study the performance of voxel-level supervised learning. For labeling nodule/vessel voxels, 3D Haar features are also feasible for PBT feature selection and boosting. However, Haar features do not apply to colonic surface voxel learning since it is not rotation-invariant and can be very memory and computationally expensive [28]. It is observed that 3D steerable features noticeably outperform 3D Haar features in learning nodule voxel classifier and perform comparably for classifying vessel voxels. Next, Fisher Discriminant Score (FDS) of any given SAIF image feature f is defined as $FD(f) = (\bar{f}^+ - \bar{f}^-)^2/(\sigma^2(f^+) + \sigma^2(f^-))$ where $\bar{f}^+$ and $\bar{f}^-$ denote the mean; $\sigma^2(f^+)$ and $\sigma^2(f^-)$ present the covariance of f distribution on positive $\{f^+\}$ and negative $\{f^-\}$ classes. FDS describes the gross two-class distribution separability or decision margin in a 1-dimensional feature space, as representing discriminativeness.

In Fig. 4, many SAIF features in Sec. 2.2 show excellent FD scores in polyp detection. The top three most informative feature groups are **Statistics of Voxel-Class Probabil-**

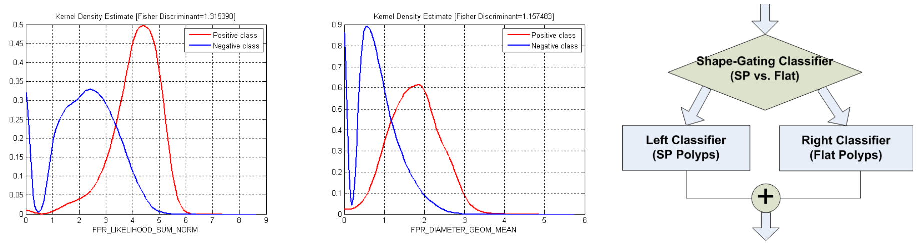


Fig. 4. Kernel Density Estimator plots of $Prob_{Sum}$ **(Left)** and Diameter_Geometry_Mean **(Middle)**, i.e., $(R_1 \times R_2 \times R_3)^{1/3}$, with the highest and $5th$ highest Fisher Discriminant (FD) scores of 1.3154 and 1.1575, respectively. **(Right:)** Shape Morphology based Soft-Gating Framework.

ity $\{\wp\}$, geometric dimensions from **3D Ellipsoid Shape Descriptor**, and surprisingly, **Boundary Gradient-Shape Statistics Features**. For example, the sum of probability feature $Prob_{Sum}$ and geometric mean diameter $(R_1 \times R_2 \times R_3)^{1/3}$ demonstrate excellent FD scores as 1.3154 and 1.1575 for single features. When size-gating is employed, boundary features show higher weights on the branch of large lesions which indicates more reliable feature computation as well. The supervisedly trained Probability features $\{\wp\}$ clearly outperform intensity feature groups. On the other hand, previous works employ and even heavily rely upon intensity information for detection and segmentation of lesions in CT images. In summary, $\{\wp\}$ $>$ Gradient $>$ Intensity. Similar observations are also found for nodule SAIF features.

CAD Performance: Integrating our hierarchically learned SAIF image features into existing system for CAD training, achieves significantly higher detection rates under the same false positive (FP) rate and in terms of free ROCs (Fig. 5). In colon CAD validation dataset, with feature integrated, our shape-gated main classifier improves the per-patient actionable polyp detection sensitivity from 82.50% [99/120] to 93.33% [112/120], along with per-volume FP rate dropping to 2.18 from 2.85. Furthermore, it also improves the performance for detecting the relatively more difficult subcategory of flat polyps from 68.29% [28/41] to 87.80% [36/41], due to the "divide-and-conquer", SP versus flat polyp gating scheme. Six instance-level features based on SP versus Flat trained voxel labeler are selected with high weights for the gating MILRVM classifier of $Pb(SP)$, $Pb(Flat)$, out of 13 chosen features. Polyp features also strongly contribute to compose the left $Pb_L(Polyp|SP)$ and right $Pb_R(polyp|Flat)$ leaf classifiers. For example, multiscale intensity histogram features help removing tagged stools and fatty polyps of no clinical significance, with higher intensity concentrations in outer bins. Spatial occupancy features eliminate tiny, spurious structures. As shown in Fig. 5 for both training and validation, the final wall versus nonwall gating + vessel FP filter consistently improves ROC performance for all three nodule categories and across datasets. Especially, harder-to-detect small solid nodules ($3 \sim 4mm$) have higher sensitivity of 78.6% [88/112] from 69.64% [78/112]; and GGN of 84.3% increased from 80.4%. High sensitivities on detecting solid small and GGN nodules provide complementary and important values where radiologists may underperform, by employing CAD as a second reader. Using learned features with high FD scores helps the classifier generality from training to validation dataset.

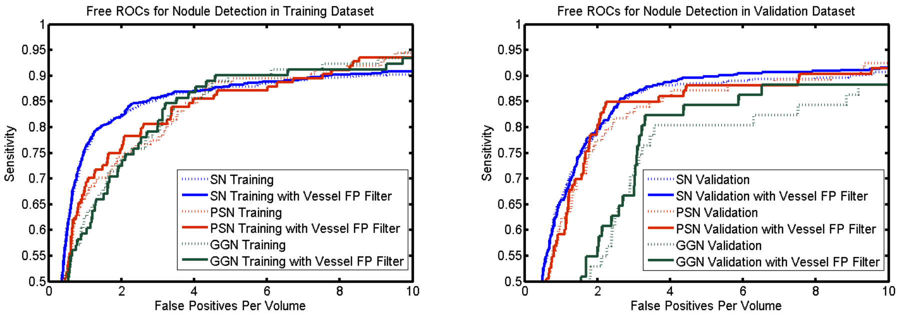


Fig. 5. Free ROC Curves of CAD system-level lung nodule detection in three categories: Solid Nodule (SN) in Blue; Partial Solid Nodule (PSN) in Red; and Ground-Glass Nodule (GGN) in Green, for both training **(Left)** and validation **(Right)** datasets. The impact of Vessel FP filter is evaluated.

Comparison: Our CAD performances compare favorably against the state-of-the-arts [1,2,5,8,12,14]. For Nodule detection, we achieve testing sensitivities of 90% (SN $\geq$ 3mm), 87.2% (PSN), 84.3% (GGN) and 78.6% (small with $3 \sim 4$ mm) at 4.1 FP/scan, while [2] reports sensitivity of 90.2% at 8.2 FP/scan; [5] obtains 90% sensitivity for non-calcified solid parenchymal nodules ($\geq$ 4mm) at 5.1 FP/Scan. [8] manually preselects 140 CT volumes with at least one ($\geq$ 4mm) GGN nodule and reports sensitivity of 73% at 1 FP/Scan ($77\% \sim 78\%$ at 4 FP/Scan) for GGN nodule detection only. For comparison on dataset scales, [5] uses 60 nodules from 50 CT scans under a single imaging protocol. [2] employs 108 thoracic CT scans using a wide range of tube dose levels which contain 220 nodules (185 solid and 35 GGN). We report results on total 879 volumes of 1179 solid, 217 partial solid and 142 ground-glass nodules, which is the largest. For Polyp detection, our testing sensitivities are 93.3% (SP$\geq$3mm) and 87.8% (Flat) at 2.18 FP/scan, while [12] achieves $85 \sim 95\%$ (SP$\geq$6mm) sensitivities at 5 FP/scan. [14] reports 90% (SP$\geq$6mm) and $75 \sim 80\%$ (flat) at 4.5 FP/scan. In colon CAD, we use 770 Tagged-prep CT scans of multiple sites, among the largest studies with [12,14].

Geometric or Probabilistic Process? A variety of drastically different techniques have been proposed for lesion detection. However, most previous work [1,2,5,11,12,15,16,23,29,30] focus on *extracting low-level, directly observable surface geometry and volumetric intensity features*: as geometric descriptors (mostly curvature based) to describe the degree of satisfying the sphericity polyp shape assumption [11,16], segmentation or geometric protrusion based polyp occupancy measurements [12], fuzzy clustering and deformable model [29], and intensity features (as mean, median, maximum, minimum, etc.) [30] or Hessian statistics for polyp detection. [1,2,5,15,23] all address nodule shape morphology modeling versus other structures. In our work, geometry and intensity information are first encoded into the voxel labeling process through PBT learning. Then translation and rotation invariant visual features are computed summarizing the joint distribution of intensity and learned lesion-class probability.

Is Data-driven Learning Required for Large Data? Probabilistic approach of modeling the shape differences between polyps and other colonic surface structures is exploited [13] in a Bayesian closed-form mathematical formulation. Similarly, [5] discusses its counterpart in nodule detection. Both work derive features based on the shape, intensity prior knowledge presented in the medical literature. From a large scale annotated polyp/nodule segmentation dataset (e.g., hundreds of volumes and lesions), their analytic models do not fit well the huge anatomical appearance variation. Our voxel-level labeling is data-driven and supervisedly learned from annotated object image masks [6,9,17] on vast datasets (for better image generality). Consequently, [5,13] report significantly inferior performance results on very limited datasets of 36 volumes and 24 polyps and 50 volumes with 60 solitary solid nodules. Apart from another learning based 3D polyp detection approach [28], our learning and parsing processes are efficiently performed on the colonic surface rather than exhaustively searching in subvolumes. Therefore the difficulty of detecting variable-posed 3D volumetric objects is completely avoided, while the required sample alignment in [28] by rotating 3D volumes is very time-consuming.

4 Conclusion

Our main contributions are four-fold. First, a flexible, hierarchical feature learning framework is presented, integrating different levels (voxel-, ROI instance- and database-level) of discriminative and descriptive information for CAD. Second, we propose Spatial Aggregation Image Features (insensitive to segmentation noises) to encode the robust statistics from voxel labeling responses within each ROI. Third, our approach provides a unified solution of detecting all types of nodules (solid, partial-solid, ground-glass) and polyps (sessile, pedunculated, flat; large, small), via the learned image features. Image appearance of different lesion categories are mapped into the universal $\wp$-probability space whereas previous work design different methods for each lesion type [5,8,12]. Last, we validate CAD performances on large-scale datasets achieving comparable or higher sensitivity at significantly lower FP rates, against the state-of-the-arts [1,2,5,8,12].

References

1. Dehmeshki, J., Amin, H., Valdivieso, M., Ye, X.: Segmentation of pulmonary nodules in thoracic ct scans: A region growing approach. IEEE Trans. Med. Imag. **27** (2008) 467–480
2. Ye, X., Lin, X., Dehmeshki, J., Slabaugh, G., Beddoe, G.: Shape based computer-aided detection of lung nodules in thoracic ct images. In: IEEE Trans. on Biomedical Engineering. (1810-20, 2009)
3. Kubota, T., Jerebko, A., Dewan, M., Salganicoff, M., Krishnan, A.: Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models. Medical Image Analysis **15(1):133-154** (2011)
4. El-Baz, A., Gimel'farb, G.: Robust medical images segmentation using learned shape and appearance models. In: MICCAI. (2009) 281–288
5. Mendonca, P., Bhotika, R., Zhao, F., Miller, J.: Lung nodule detection via bayesian voxel labeling. In: Information Processing in Medical Imaging. (134-146, 2007)
6. Wu, D., Lu, L., Bi, J., Shinagawa, Y., Boyer, K., Krishnan, A., Salganicoff, M.: Stratified learning of local anatomical context for lung nodules in ct images. In: IEEE CVPR. (2010)
7. Lo, P., Sporring, J., Ashraf, H., Pedersen, J., de Bruijne, M.: Vessel-guided airway tree segmentation: A voxel classification approach. Medical Image Analysis **14** (2010) 527–538
8. Jacobs, C., Sanchez, C., et al.: Computer-aided detection of ground glass nodules in thoracic ct images using shape, intensity and context features. MICCAI **3** (2011) 207–14

9. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: Multiclass object recognition and segmentation by jointly modeling texture, layout, and context. In: IJCV. (1:2-23, 2009)
10. Raykar, V., Krishnapuram, B., Bi, J., Dundar, M., Rao, R.: Bayesian multiple instance learning: automatic feature selection and inductive transfer. In: ICML. (2008) 808–815
11. Jerebko, A., Lakare, S., Cathier, P., Periaswamy, S., Bogoni, L.: Symmetric curvature patterns for colonic polyp detection. In: MICCAI. (2006) 169–176
12. van Wijk, C., van Ravesteijn, V., Vos, F., van Vliet, L.J.: Detection and segmentation of colonic polyps on implicit isosurfaces by second principal curvature flow. IEEE Trans. Medical Imaging (2010)
13. Melonakos, J., Mendonca, P., Bhotka, R., Sirohey, S.: A probabilistic model for haustral curvatures with applications to colon cad. In: MICCAI. (2007)
14. Slabaugh, G., Yang, X., Ye, X., Boyes, R., Beddoe, G.: A robust and fast system for ctc computer-aided detection of colorectal lesions. Algorithms **3(1)** (2010) 21–43
15. Mendonca, P., Bhotika, R., Sirohey, S., Turner, W., Miller, J., Avila, R.: Model-based analysis of local shape for lesion detection in ct scans. In: MICCAI. (688-695, 2005)
16. Paik, D., et al.: Surface normal overlap: a computer-aided detection algorithm with application to colonic polyps and lung nodules in helical ct. IEEE Trans. Med. Imag. (2004) 23(6):661–75
17. Lu, L., Barbu, A., Wolf, M., Liang, J., Bogoni, L., Salganicoff, M., Comaniciu, D.: Accurate polyp segmentation for 3d ct colonography using multi-staged probabilistic binary learning and compositional model. In: IEEE CVPR. (2008)
18. Lorensen, W., Cline, H.: Marching cubes: A high resolution 3d surface construction algorithm. ACM SIGGRAPH Computer Graphics **21(4): 163-169** (1987)
19. Lu, L., Barbu, A., Wolf, M., Liang, J., Bogoni, L., Salganicoff, M., Comaniciu, D.: Simultaneous detection and registration for ileo-cecal valve detection in 3d ct colonography. In: European Conference on Computer Vision. Volume 4. (2008) 10–15
20. Tu, Z.: Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering. In: ICCV. (2005) 1589–1596
21. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: ICCV. (1-8, 2007)
22. Schwing, A., Zach, C., Zheng, Y., Pollefeys, M.: Adaptive random forest-how many experts to ask before making a decision? In: CVPR. (1377-84, 2011)
23. Farag, A., Graham, J., Farag, A., Falk, R.: Lung nodule modeling - a data-driven approach. In: ISVC. (347-356, 2009)
24. Lu, L., Bi, J., Wolf, M., Salganicoff, M.: Effective 3d object detection and regression using probabilistic segmentation features in ct images. In: IEEE Computer Vision and Pattern Recognition. (2011) 1–8
25. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Trans. Pattern Anal. Mach. Intell. **29** (2007) 2247–2253
26. Toews, M., Wells, W.M.: Efficient and robust model-to-image alignment using 3d scale-invariant features. Medical Image Analysis (2013)
27. Smith, K., Carleton, A., Lepetit, V.: Fast ray features for learning irregular shapes. In: ICCV. (2009)
28. Tu, Z., Zhou, X., Bogoni, L., Barbu, A., Comaniciu, D.: Probabilistic 3d polyp detection in ct images: The role of sample alignment. In: CVPR. (2006) 1544–1551
29. Yao, J., Miller, M., Franaszek, M., Summers, R.: Colonic polyp segmentation in ct colongraphy-based on fuzzy clustering and deformable models. IEEE Trans on Medical Imaging **23(11)** (2004) 1344–1352
30. van Ravesteijn, V., van Wijk, C., Vos, F., Truyen, R., Peters, J., Stoker, J., van Vliet, L.: Computer aided detection of polyps in ct colonography using logistic regression. IEEE Trans. on Med. Imag. (2010)