Multiple Instance Learning of Pulmonary Embolism Detection with Geodesic Distance along Vascular Structure

Jinbo Bi
Knowledge Solutions Group, IKM
Siemens Medical Solutions
jinbo.bi@siemens.com

Jianming Liang
Computer Aided Diagnosis Group, IKM
Siemens Medical Solutions
jianming.liang@siemens.com

Abstract

We propose a novel classification approach for automatically detecting pulmonary embolism (PE) from computed-tomography-angiography images. Unlike most existing approaches that require vessel segmentation to restrict the search space for PEs, our toboggan-based candidate generator is capable of searching the entire lung for any suspicious regions quickly and efficiently. We then exploit the spatial information supplied in the vascular structure as a post-candidate-generation step by designing classifiers with geodesic distances between candidates along the vascular tree. Moreover, a PE represents a cluster of voxels in an image, and thus multiple candidates can be associated with a single PE and the PE is identified if any of its candidates is correctly classified. The proposed algorithm also provides an efficient solution to the problem of learning with multiple positive instances. Our clinical studies with 177 clinical cases demonstrate that the proposed approach outperforms existing detection methods, achieving 81% sensitivity on an independent test set at 4 false positives per study.

1. Introduction

Pulmonary embolism (PE) is a highly lethal condition that occurs when an artery in the lung becomes completely or partially blocked. In most cases, the blockage is caused by one or more blood clots that travel to the lungs from other parts of the body (e.g., legs or pelvis). While PE is not always fatal, it is nevertheless the third most common cause of death in the US, with at least 650,000 cases occurring annually [2]. The clinical challenge, particularly in an Emergency Room scenario, is to correctly diagnose patients that have a PE and then send them on to therapy. This, however, is not easy, as the primary symptom of PE is dysapnea (shortness of breath), which has a variety of causes, some of which are relatively benign. Thus, it is hard to separate out the critically ill patients suffering from PE from patients who may require a different treatment or none at all.

The two crucial clinical challenges for a physician, therefore, are to diagnose whether a patient is suffering from PE and to identify the location of the PE. Computed Tomography Angiography (CTA) has emerged as an accurate diagnostic tool for PE. However, each CTA study consists of hundreds of images, each representing one slice of the lung. Manual reading of these slices is laborious, time consuming and complicated by various PE look-alikes (false positives) including respiratory motion artifacts, flow-related artifacts, streak artifacts, partial volume artifacts, stair step artifacts, lymph nodes, and vascular bifurcation, among many others. Consequently, a computer aided detection (CAD) system is demanded to assist radiologists in detecting and characterizing emboli in an accurate, efficient and reproducible way [9, 11, 14, 10].

When PE is diagnosed, medications are given to prevent further clots, but these medications can sometimes lead to subsequent hemorrhage and bleeding since the patient must stay on them for a number of weeks after the diagnosis. Thus, the physician must review each CAD output carefully for correctness in order to prevent over-diagnosis. Because of this, the CAD system has to produce only a small number of false positives per patient scan. The goal of a PE CAD system, therefore, is to automatically identify PEs with as few false positives as possible.

We have developed a PE CAD prototype following a common paradigm which consists of three stages: candidate generation (CG), feature computation and classification. The candidate generation aims to quickly identify suspicious PE regions. Each suspicious region corresponds to a cluster of voxels, and is referred to as a PE candidate. Feature computation is to extract a set of descriptive features from each candidate, so that these features can be utilized by a classifier to identify true PE candidates and reduce the false positives (*i.e.*, candidates which do not refer to any true PEs). This paper presents a spatial multiple instance learning approach which employs the spatial vascular structural information in the form of geodesic distances among the

1-4244-1180-7/07/$25.00 ©2007 IEEE

candidates, and is capable of learning with multiple positive candidates to effectively reduce the false positives.

2. Contributions of Our Approach

1. Spatial information given by the vascular tree structure is utilized in the post-candidate-generation step.

   Because thrombi of pulmonary embolism exist only inside the pulmonary arteries, all the existing approaches to automatic PE detection (e.g., [11, 14, 10]) require vessel segmentation as the very first step, and then limit their search space for PE to the segmented vessels. However, vessel segmentation is computationally time-consuming and has been problematic in terms of accuracy which results in mis-placing potential PEs outside of the search space. Limiting the search space to inaccurate vessel segmentation imposes a strong constraint, and any mistakes in the vessel segmentation can result in missing PEs for detection.

   Our PE candidate generator searches the entire lung for any suspicious regions based on a "tobogganing" method, which utilizes the simple sliding operation and requires no vessel segmentation, resulting in high efficiency (only 30 seconds for each case on a 2.4GHz P4 PC) and high sensitivity ($\geq$ 90% per case) with modest false positive rate (50 per case). Not only does our approach produce all the suspicious PE regions (candidates), but, as a by-product, it also provides a reasonable vessel tree segmentation. This vessel segmentation may not be accurate enough to limit the search space, but it carries structural information which can be used in later stages to enhance detection accuracy. This vascular structural information has neither been utilized in the candidate generation nor in the feature computation. We propose to use it for accurate classification of the candidate clusters by imposing a soft constraint derived from the vessel tree structure. PE candidates close to each other along the vessel structure are more likely to share common characteristics than those far apart. The soft constraint is derived through the geodesic distances between the candidate clusters along the vascular structure.

2. Learning with multiple positive candidates is considered in order to effectively reduce false positives

   For clinical acceptability, it is critical to control false positive rates and detect as many true PEs as possible. A PE is a clot in the blood vessel, and thus it can be large, or have an elongated shape along the vessel, or split at the vessel bifurcation. Multiple candidate regions can exist to intersect with a single PE. Once one of the candidates is identified and highlighted to physicians, physicians can easily trace out the entire PE. Consequently, it is sufficient to detect one candidate for each PE. In the classification task, a PE is identified if at least one of the candidates associated with it is correctly labeled as a positive. PE detection rate (or sensitivity) is defined as the percentage of PEs correctly identified in all patients. A false positive is a candidate which does not overlay with any PE but falsely labeled by a classifier as a positive. The false positive rate is defined as the total number of non-PE candidates classified as positives in a patient.

   Although this problem can be regarded as a special case of multiple instance learning (MIL) problems [3, 8], it is distinct from typical MIL tasks in the sense that less ambiguity is present in our problem. In a typical MIL problem, many instances of a positive example can be fundamentally negative instances. Nevertheless, the positive instances in our problem all intersect in one way or another with segmented PEs marked by radiologists, meaning that the instances of a positive example (a PE) are all positive instances. Our goal is not to identify the truly positive instances for positive examples, but to identify at least one instance for each positive example. Although the proposed method can be applied to typical MIL problems as well, it is more efficient to deal with our circumstances.

3. Candidate and Feature Extraction

Our candidate generation method for PE detection is based on tobogganing [5]. In tobogganing, which takes its name from the processing analogy of sliding down a steep hill, a voxel v with intensity value $P(v)$ and neighbors $N(v)$ slides down to the voxel $\hat{v}$: $\hat{v} = \arg\min_{t \in N(v) \cup \{v\}} P(t)$. When multiple voxels in the neighborhood of a voxel achieve the same minimal intensity, the first voxel met may be chosen. A voxel that cannot slide to any of its neighbors is called a concentration. The voxels sliding down to the same concentration form a toboggan cluster with a unique label. We have developed a dynamic fast tobogganing algorithm, which starts from a specified location and quickly forms a toboggan cluster locally without involving any voxels beyond the outer boundary of the toboggan cluster. Each voxel is only processed once, leading to high efficiency. At the completion of tobogganing, we collect all the toboggan clusters in the search region to form PE segmentations, from which a single detection position is derived based on morphological ultimate erosion. In summary, the candidate generation approach includes the following steps: coarse lung segmentation, tobogganing, collecting toboggan clusters and identifying detection locations.

Our tobogganing algorithm has been further accelerated based on the following *a priori* knowledge of medical physics. First, in CTA images, the embolus appears as

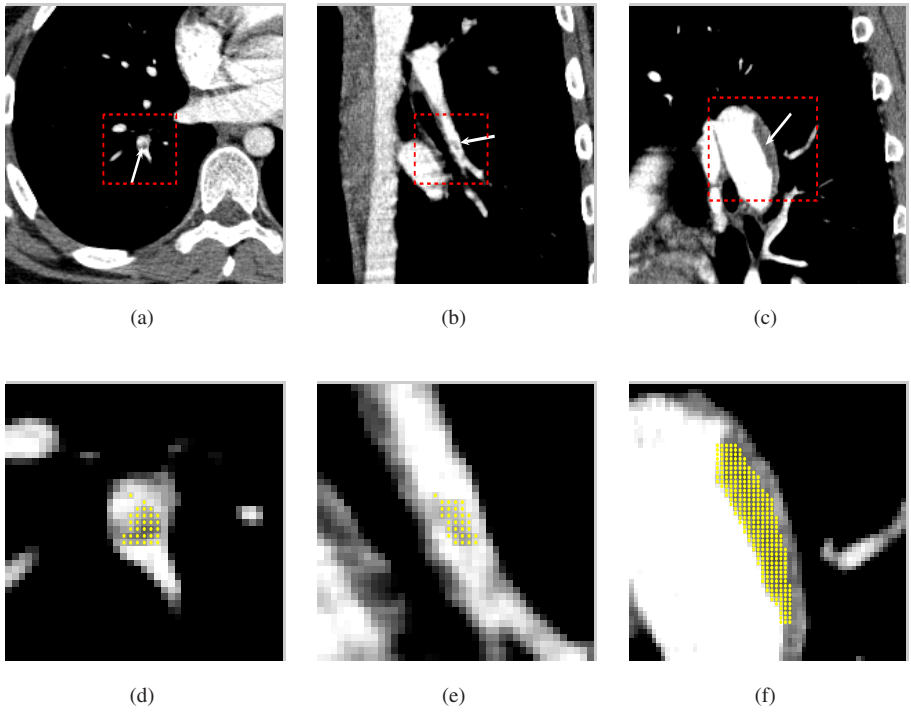


Figure 1. Our toboggan-based approach is capable to detect both acute (a, b) and chronic (c) pulmonary emboli, offering simultaneous detection and segmentation as in (d, e, f) which are zoomed images of (a, b, c) respectively.

dark regions (see Figure 1) with Hounsfield Units (HU) between -50 HU and 100 HU. Therefore, we can effectively exclude all the voxels outside the range [-50 HU, 100 HU] and only slide the voxels inside this HU range to accelerate the tobogganing process. Although PEs only exist in pulmonary arteries, which makes artery segmentation necessary, the Hounsfield density criterion ([-50 HU, 100 HU]) excludes areas above 100 HU in the arteries, therefore, large part of segmented arteries would be excluded in any case.

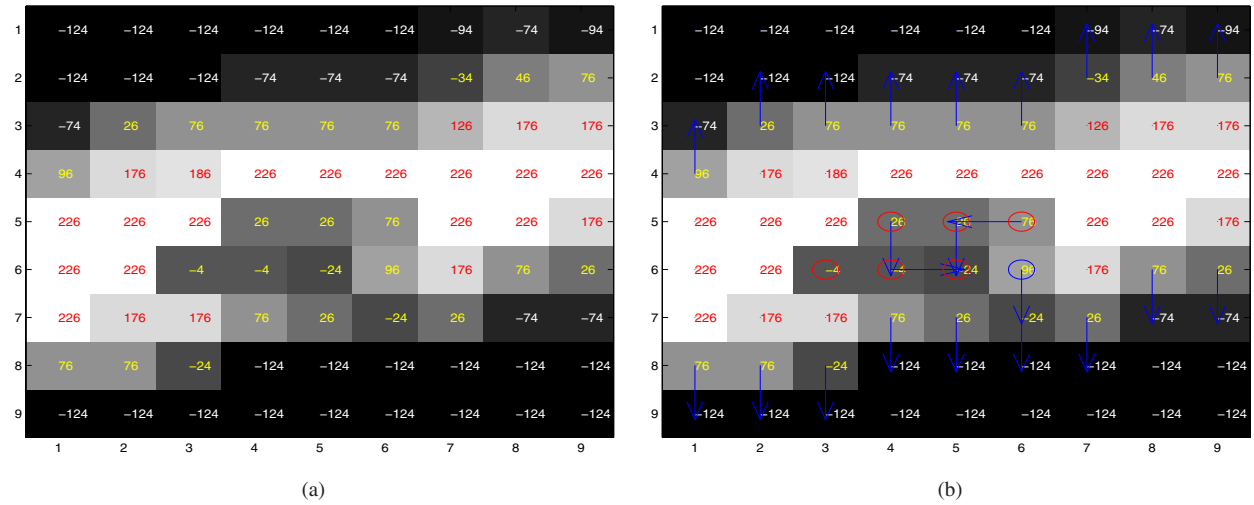
Our technique permits the use of a fast and coarse overall lung segmentation as PE search region, bounded by the outer pleural surface, instead of an arterial segmentation. Moreover, when only a coarse lung mask is used, it is crucial to properly handle partial volume artifacts around the vessel wall and around the airway wall, because all the voxels in those areas have the similar original CT values to those in the PE regions. In the tobogganing process, almost all those voxels automatically merge into regions outside of the PE search region (below -50 HU), and are therefore automatically removed when collecting toboggan clusters. An example is illustrated in Figure 2 to show how our candidate generator works and readers can consult with [7] for more complete algorithmic description.

For each candidate, a set of 120 image-based features are computed and are normalized to a unit range with a feature specific mean. The features can be categorized into those that are indicative of voxel intensity distributions within the candidate, those summarizing distributions in neighborhood of the candidate, and those that describe the 3-D shape of the candidate and enclosing structures. When combined, these features can capture candidate properties that can disambiguate true emboli from typical false positives such as dark areas that result from poor mixing of bright contrast agents with blood in veins, and dark connective tissues between vessels.

4. Classification Formulations

In this section, we propose a classification framework that is general enough to be used in conjunction with any existing classification formulation, such as Fisher linear discriminant analysis and classic support vector machines (SVM), to exploit the spatial distance information and to learn with multiple positive instances for PE detection. Particularly, we use the 1-norm SVM as an example to illustrate our approach since the 1-norm SVM often achieves sparse classifiers [15, 1] which are desired in our system to reduce the number of features for model capacity control and time efficiency.

Assume that totally ℓ candidates $(\mathbf{x}_i, y_i),\ i = 1, \cdots, \ell$ are extracted from the n training volumetric images and d



(a) (b)

Figure 2. This 2D artificial image was created to resemble a small artery with pulmonary embolism. Pixels with CT values below -50 HU in white, pixels with CT value above 100 HU in red and all other pixels in yellow. Naturally, all the PE pixels are in yellow. However, due to partial volume effects, the pixels around the artery boundaries are also in yellow. It is also true for the airway boundaries (not shown here). Therefore, for automatic PE detection, it is crucial to remove the pixels around the artery boundaries and those around the airway boundaries, while preserving the PE pixels. This can be done with tobogganing by sliding all the yellow pixels to its neighbor with minimal CT value (2D four-connected neighborhood used in this example), and collecting all pixels that don't slide into regions with Hounsfield Unit below -50 HU, forming PE detections (*i.e.*, PE regions or PE clusters). As a by-product from the tobogganing process, the PE detected regions along with the pixels with CT value above 100 HU (in red) form a rough vessel segmentation. This paper utilizes this rough vessel segmentation in computing the geodesic distance between each pair of PE candidates.

image-based features are computed after the candidate generation and feature calculation steps. Let I^+ and I^- be two index sets containing indices of candidates that intersect with PEs and do not intersect with PEs, respectively. Let m be the total number of PEs marked by expert radiologists for the n images. Denote I_j as the index set of the candidates that intersect with the j-th PE, $j = 1, \cdots, m$. Notice that $\cup_{j=1,\cdots,m} I_j = I^+$ but any two index sets I_j's are not necessarily disjoint since there may exist a candidate cluster that intersects with more than one segmented PE. The vector $\mathbf{x}_i$ comprises all the feature values computed for the i-th candidate, and y_i is the label obtained by consulting with the markers given by expert radiologists.

The regular 1-norm SVM constructs a hyperplane classifier as $\text{sign}(\mathbf{w}^T\mathbf{x}+b)$ by optimizing the following problem:

$$\begin{array}{ll} \min_{\mathbf{w},\xi} & \gamma||\mathbf{w}||_1 + \sum_{i\in I^+}\xi_i + \sum_{i\in I^-}\xi_i \\ \text{s.t.} & \mathbf{w}^T\mathbf{x}_i + b \geq 1-\xi_i,\ i\in I^+, \\ & \mathbf{w}^T\mathbf{x}_i + b \leq -1+\xi_i,\ i\in I^-, \\ & \xi_i \geq 0,\ \ i = 1,\cdots,\ell. \end{array} \tag{1}$$

where γ is the regularization parameter that controls the trade-off between the regularization factor $\|\mathbf{w}\|_1$ and the error term. A candidate (represented by a vector $\mathbf{x}$) is classified correctly when $\xi = 0$, or otherwise a positive ξ defines the hinge loss of the candidate.

4.1. Learning with multiple positive instances

The classification approach can be more efficient by exploring the observation that once a candidate in I_j is classified as a positive, then the j-th PE is considered being identified. This consideration motivates the resulting classifier to focus on different PEs instead of multiple candidates within a single PE. Especially it helps reduce false positives by possibly ignoring extremely noisy candidates that intersect with some PEs where, for the same PE, other associated candidates can be easily classified correctly.

Mathematically, distinguishing at least one candidate for each PE from the negative class is equivalent to requiring the minimum of the errors (ξ) that occur on the candidates associated with a PE to be 0. For example, if a PE is associated with 3 candidates, and a classifier generates $\xi_1 = 0$ for the first candidate, $\xi_2 = 5,\ \xi_3 = 100$ for the other two candidates, the classifier detects the PE. Correspondingly, this implies to construct the classifier by solving the following optimization problem:

$$\begin{array}{ll} \min_{\mathbf{w},\xi} & \gamma||\mathbf{w}||_1 + \sum_{j=1}^{m}\min\{\xi_i,\ i\in I_j\} + \sum_{i\in I^-}\xi_i \\ \text{s.t.} & \mathbf{w}^T\mathbf{x}_i + b \geq 1-\xi_i,\ i\in I^+, \\ & \mathbf{w}^T\mathbf{x}_i + b \leq -1+\xi_i,\ i\in I^-, \\ & \xi_i \geq 0,\ i = 1,\cdots,\ell. \end{array} \tag{2}$$

This optimization problem is computationally difficult to solve since it involves a minimization of the to-be-determined variables ξ in the evaluation of the objective function, and is neither differentiable nor convex. Hence an equivalent conversion of the problem to a tractable optimization problem is necessary. We prove that the following quadratic program is equivalent to problem (2). The following theorem characterizes our result.

$$\begin{array}{ll} \min_{\mathbf{w},\xi,\boldsymbol{\lambda}} & \gamma||\mathbf{w}||_1 + \sum_{j=1}^{m}(\sum_{i\in I_j}\lambda_i\xi_i) + \sum_{i\in I^-}\xi_i \\ \text{s.t.} & \mathbf{w}^T\mathbf{x}_i + b \geq 1-\xi_i,\ i\in I^+, \\ & \mathbf{w}^T\mathbf{x}_i + b \leq -1+\xi_i,\ i\in I^-, \\ & \xi_i \geq 0,\ i=1,\cdots,\ell, \\ & \sum_{i\in I_j}\lambda_i = 1, \lambda_i \geq 0,\ i\in I_j,\ j=1,\cdots,m. \end{array} \tag{3}$$

Theorem 4.1 *Any optimal solution $\hat{\mathbf{w}}$ of Problem (2) is optimal to Problem (3) with properly chosen λ, and vice versa.*

Proof. First of all, we prove that an optimal solution of Problem (3) has nonzero λ's only on the candidates for which the classifier achieves $\min\{\xi_i,\ i\in I_j\},\ \forall\, j$.

Let $(\hat{\mathbf{w}},\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\lambda}})$ be the optimal solution of Problem (3). For notational convenience, denote the objective of Problem (3) as $\mathcal{J}(\mathbf{w},\boldsymbol{\xi},\boldsymbol{\lambda}) = \gamma||\mathbf{w}||_1+\sum_{j=1}^{m}(\sum_{i\in I_j}\lambda_i\xi_i)+\sum_{i\in I^-}\xi_i$. Then let $\hat{\mathcal{J}}$ be the objective value attained at the optimal solution. Notice that the hinge loss $\hat{\boldsymbol{\xi}}$ is uniquely determined by $\hat{\mathbf{w}}$ as $\hat{\xi}_i = \max\{0, 1-y_i(\mathbf{w}^T\mathbf{x}_i+b)\}$ for each candidate $\mathbf{x}_i$.

If $\exists\, j\in\{1,\cdots,m\}$, and $\exists i_0\in I_j$, such that $\lambda_{i_0}>0$ but $\hat{\xi}_{i_0} \neq \min\{\xi_i, i\in I_j\}$. Then let $\xi_{I_j} = \min\{\xi_i, i\in I_j\} < \hat{\xi}_{i_0}$. Then $\tilde{\mathcal{J}} = \hat{\mathcal{J}} - \lambda_{i_0}\xi_{i_0} + \lambda_{i_0}\xi_{I_j} < \hat{\mathcal{J}}$. This contradicts to the optimality of $(\hat{\mathbf{w}},\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\lambda}})$.

By this contradiction, $\forall i,j$, such as $\lambda_i > 0$, the corresponding ξ_i has to be the minimum loss that the classifier achieves on the j-th PE. This implies that at the optimality of Problem (3), $\mathcal{J} = \gamma||\mathbf{w}||_1 + \sum_{j=1}^{m}\min\{\xi_i,\ i\in I_j\} + \sum_{i\in I^-}\xi_i$ which is the same as the objective of Problem (2). ∎

In comparison with a previous work [6] that finds the best convex combination of candidate feature vectors ($\mathbf{x}$'s) within each PE and constructs a classifier to classify correctly as many these convex combinations as possible, our approach gains computational efficiency since no convex hull needs to be formed, instead, only a specific candidate (corresponding to the minimum ξ) needs to be determined properly for each PE. A fundamental assumption of [6] is that any convex combination of instances of a PE is also associated with the PE, which may not be true in practice. Moreover, our formulation can easily incorporate spatial information into the classifier construction as shown in the next section.

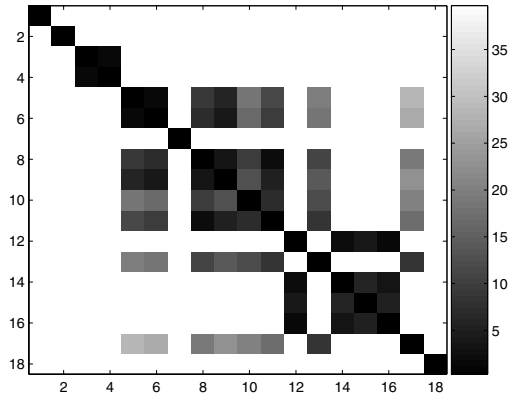


Figure 3. A geodesic distance map illustrated in gray-scale for a patent with 18 candidates. Naturally, it is always completely dark along the major diagonal, as the geodesic distance from a candidate to itself is zero. The geodesic distance map is always symmetric, as the geodesic distance from Candidate A to Candidate B is the same as from Candidate B to Candidate A. It is completely white if there is no shortest path between a pair of candidates (*e.g.*, from a candidate in the pulmonary artery to a candidate in the pulmonary vein, or from a candidate located in the left lung to a candidate in the right lung). The geodesic distance map is sparse in most cases. For this case, the max geodesic distance is found between candidate 5 and candidate 17 and it is 28.2 mm.

4.2. Incorporating geodesic distance maps

The geodesic distance from one PE candidate to another PE candidate along the vessel is approximated with a modified Dijkstra's algorithm. Dijkstra's algorithm is a greedy algorithm that solves the single-source shortest path problem for a graph with nonnegative edge weights. It is common to regard an image as a graph, but we only limit to those pixels (voxels) in the vessel as a node and create an edge between each neighboring pixels with its Euclidean distance (taking pixel/voxel size into consideration) as the weight. Since a PE candidate is a cluster of voxels, the geodesic distance we are computing is from a group of voxels to groups of voxels. To this end, we modify the graph by assigning a zero weight to the edges created between any pair of neighboring voxels within the same cluster. With this modified graph, Dijkstra's algorithm computes the shortest distance from the boundary of one PE candidate to the boundary of another PE candidate. For each patient, a geodesic distance map matrix G is thus created among the candidates. A geodesic distance map is shown as an image in Figure 3.

There may not exist a path of reasonable length between a pair of candidates, for instance, a candidate in the pulmonary artery and a candidate in the pulmonary vein, or a

candidate located in the left lung and a candidate in the right lung. Consequently, the resulting similarity matrix of candidates will be sparse. The similarity matrix for each patient is calculated in this paper as $S = \exp(-\alpha G)$ where $\alpha > 0$ is a proper scaling parameter.

Further investigation is needed to incorporate the similarity matrices (geodesic distance maps) into the learning formulation (3). Let us introduce more notation here. We use p to index the patients. Assume that candidates generated from the p-th patient's image are $\mathbf{x}_{i_1}, \cdots, \mathbf{x}_{i_p}$. We form a matrix $X_p = [\mathbf{x}_{i_1} \cdots \mathbf{x}_{i_p}]^T$ to represent the candidate set for the patient p where each row represents one candidate from the p-th patient. Then problem (3) can be rewritten as follows:

$$\begin{array}{ll} \min_{\mathbf{w},\xi,\boldsymbol{\lambda}} & \gamma||\mathbf{w}||_1 + \sum_{j=1}^{m}(\sum_{i\in I_j}\lambda_i\xi_i) + \sum_{i\in I^-}\xi_i \\ \text{s.t.} & Y_p\left(X_p\mathbf{w}+b\right) \geq \mathbf{1} - \boldsymbol{\xi}_p, \\ & \quad \boldsymbol{\xi}_p \geq 0,\ p = 1,\cdots,n, \\ & \sum_{i\in I_j}\lambda_i = 1, \lambda_i \geq 0,\ i \in I_j,\ j = 1,\cdots,m. \end{array} \tag{4}$$

where Y_p is a diagonal matrix with diagonal element equal to the labels of the candidates, that is, $y = 1$ if the corresponding candidate intersects with a PE, or $y = -1$ otherwise. The variable $\boldsymbol{\xi}_p$ is a vector containing all the hinge losses ξ_i occurred on the candidates $\mathbf{x}_i$ from patient p.

Some previous work has explored the spatial information such as in [13, 12]. The portion of our approach for utilizing the spatial information is reminiscent of the method in [13]. However, for most of the previous work, only Euclidean distance is considered and implemented. The basic rational to use spatial relationship of candidates is that candidates spatially closely distributed within the vessels share similar characteristics. Hence, to classify an individual candidate in an image, the classifier needs to examine the class membership of close-by candidates in the same patient's image. For instance, the candidate generator produces 3 candidates from a patient's image, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, and we want to predict the label of the first candidate. The prediction $y_1 = z_1 + s_{12}z_2 + s_{13}z_3$ where $z_i = \mathbf{x}_i^T\mathbf{w} + b$, s_{12} and s_{13} are the similarity measures based on the geodesic distances, and they determine how similar the first candidate is to the second and the third candidates, respectively.

More specifically, for patient p, the spatially-related classifier can be essentially written as $\mathbf{y} = (\theta S_p + E)\left(X_p\mathbf{w} + b\right)$ where S_p is the similarity matrix of candidates which is often very sparse as shown in Figure 3, and E is the identity matrix, and θ is a tuning parameter that defines how much impact the neighboring candidates will have on a specific candidate. This model generates prediction results simultaneously for all candidates by mutually connecting each other with the correlation matrix $\theta S_p + E$.

In particular, for images in the training set, we classify candidates with hinge losses, $Y_p(\theta S_p + E)(X_p\mathbf{w} + b) \geq \mathbf{1} - \boldsymbol{\xi}_p$. This derives the following optimization problem

$$\begin{array}{ll} \min_{\mathbf{w},\xi,\boldsymbol{\lambda}} & \gamma||\mathbf{w}||_1 + \sum_{j=1}^{m}(\sum_{i\in I_j}\lambda_i\xi_i) + \sum_{i\in I^-}\xi_i \\ \text{s.t.} & Y_p(\theta S_p + E)(X_p\mathbf{w}+b) \geq \mathbf{1} - \boldsymbol{\xi}_p, \\ & \quad \boldsymbol{\xi}_p \geq 0,\ p = 1,\cdots,n, \\ & \sum_{i\in I_j}\lambda_i = 1, \lambda_i \geq 0,\ i \in I_j,\ j = 1,\cdots,m. \end{array} \tag{5}$$

Problem (5) is the final learning formulation we shall solve to construct classifiers. Although the indexing is a bit complicated with notation for index sets (I_j) of candidates corresponding to different PEs and index sets (p) of candidates from different patients, the resulting optimization problem is simply a quadratic program, i.e., quadratic objective with linear constraints. The problem is not necessarily convex due to the cross term $\lambda_i\xi_i$ in the objective. It can be optimized using some conjugate gradient approaches or alternating optimization approaches. Note that if $\theta = 0$, problem (5) is the same as problem (3) where spatial distance does not take an effect.

5. Experiments

We validate the proposed approach on a set of clinically-collected CTA images. We name our algorithm the Spatial Multiple-Instance-Learning (MIL) approach, and we compare our approach with the standard 1-norm SVM.

5.1. Clinical cases

We have collected 177 clinical cases from four different institutions (two North American sites and two European sites). Our expert chest radiologists reviewed each case and marked all the PEs. The cases were randomly divided into training and test sets. The training set includes 45 cases, while the test set contains the remaining 132 cases.

5.2. Candidate generation and labeling

All the 177 cases were processed with our candidate generator, which generated a total of 8806 candidates: 2431 candidates appear in the training set and 6375 candidates in the test set. Each candidate is a cluster of voxels (the 3-D analog of pixels), and represented by a representative point with a 3-D coordinate derived from the cluster of voxels.

Each candidate was then labeled as a PE or not based on 3-D landmark ground truth provided by the experts. In order to automatically label each candidate, each PE pointed out by an expert landmark is semi-automatically extracted and segmented. Therefore, the ground truth for each PE is also a cluster of voxels (*i.e.*, the segmented PE). Any candidate that was found to be intersected with any of the segmented PEs in the ground truth was labeled as a PE. Multiple candidates may intersect with the same segmented PE, that is, multiple candidates may correspond to a single PE. Each PE is assigned with a unique identifier, therefore, multiple candidates may be labeled with the same PE identifier.

The sensitivity achieved by the candidate generator was 90.4% and 91%, respectively, on the training and test sets. The related information has been summarized in Table 1.

5.3. Feature selection and classification

We compare the results of the proposed algorithm to those obtained by standard 1-norm SVM in terms of accuracy and dimension reduction. An efficient algorithm based on the commercial software CPLEX [4] was implemented to optimize Problem (5). Notice that Problem (5) is not in the canonical form of a mathematical program. Using a traditional technique as in [1], we can re-cast it into the following optimization problem where $\mathbf{w} = \mathbf{u} - \mathbf{v}$.

$$\begin{aligned} \min_{\mathbf{u},\mathbf{v},\boldsymbol{\xi},\boldsymbol{\lambda}} \quad & \gamma \sum_{k=1}^{d} (u_k + v_k) + \sum_{j=1}^{m} (\textstyle\sum_{i \in I_j} \lambda_i \xi_i) + \sum_{i \in I^-} \xi_i \\ \text{s.t.} \quad & Y_p(\theta S_p + E)(X_p(\mathbf{u} - \mathbf{v}) + b) \geq \mathbf{1} - \boldsymbol{\xi}_p, \\ & \boldsymbol{\xi}_p \geq 0,\ p = 1, \cdots, n, \\ & u_k \geq 0,\ \ v_k \geq 0,\ \ k = 1, \cdots, d, \\ & \textstyle\sum_{i \in I_j} \lambda_i = 1, \lambda_i \geq 0,\ i \in I_j,\ j = 1, \cdots, m. \end{aligned} \tag{6}$$

When λ is fixed, the above problem is simply a linear program. In our implementation, we initialize the algorithm with equal values of λ's for each PE. We then solve the problem (6) for the best $\mathbf{w}$, and then fix $\mathbf{w}$, solve (6) for the best λ.

The model parameters γ and θ were tuned through a cross validation procedure where γ was chosen from the choices of 0.01, 0.1, 0.25, 0.5, 1, 2, 5, and θ was chosen from 0.1, 1, 2, 5, 10. Cross validation results showed that $(\gamma = 0.25, \theta = 1)$ was the best choice for our approach, and $\gamma = 1.1$ was the best choice for the standard 1-norm SVM (1).

We then fixed these parameters to the chosen values, and performed a 10-fold cross validation by drawing bootstrap samples from the training data. The resulting receiver operation characteristic (ROC) curves are provided in Figure 4 together with the confidence intervals (error bars). Clearly, our approach dominates the curve obtained by the 1-norm SVM. Especially, the variance of our model is relatively smaller than that of the 1-norm SVM classifier as shown in the error bars.

In the 10-fold cross validation process, 10 classifiers were constructed. The final classifier is a bagging classifier [2] which is formed by averaging the weight vectors $\mathbf{w}$ of these 10 linear models. It is well-known that bagging often reduces the model uncertainty and stabilizes the classifiers. The bagging classifier is applied to the entire training and independent test sets. The resulting ROC curves are depicted in Figure 5 to demonstrate the final performance of our detection system. Bear in mind that some true PEs have been missed at the candidate generation step which shall not be recovered by the classifier. The system performance figure reflects the entire system performance by factorizing the classification performance with the CG sensitivity.

Table 1. Datasets Summary of the data and feature selection results.

Dataset	Training	Testing
No. Candidates	2431	6375
No. PEs in ground truth	156	716
No. PEs detected in CG	141	645
No. PEs detected by Spatial MIL	125	580
No. FP rate by Spatial MIL	4	4
No. PEs detected by 1-norm SVM	117	511
No. FP rate by 1-norm SVM	4	4
No. features used by Spatial MIL	28	28
No. features used by 1-norm SVM	56	56
Run time by Spatial MIL	26 sec	0.15 sec
Run time by 1-norm SVM	20 sec	1 sec

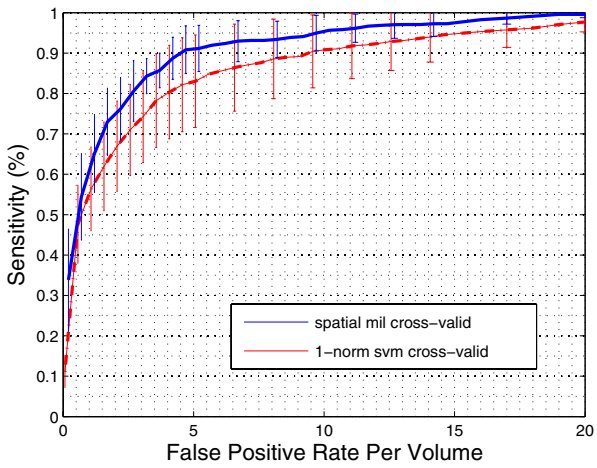


Figure 4. The 10-fold cross validation ROC plot of our proposed approach and the standard 1-norm SVM. The plot is drawn in terms of sensitivity versus the average false positive per volume.

The numbers of selected features by our approach and the 1-norm SVM are given in Table 1. There were 28 features selected in the final classifier obtained by the spatial MIL approach, and 56 features were used in the 1-norm SVM classifier. Due to the more sophisticate formulation of the spatial MIL approach, the training time is slightly longer than that needed for 1-norm SVM classifier, but the execution time is much shorter since less features were selected in the spatial MIL model.

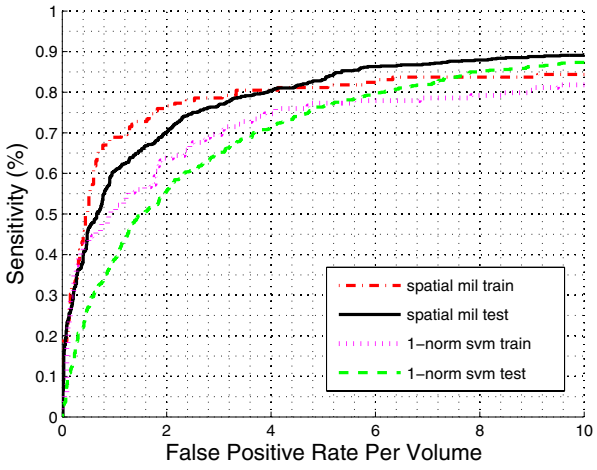


Figure 5. The final system performance on the training and test patient sets acquired by our approach and the standard 1-norm SVM

6. Conclusions

We have discussed the challenges of automatic pulmonary embolism detection and have proposed our solution which is a unified framework to deal both with the exploration of side information provided by the vascular structure and the fact that multiple candidates can refer to the same PE in the ground truth.

Our approach differs from previous methods in two aspects. First, all previous methods require significant efforts to segment the vascular structure and use the segmented vascular tree in advance to constrain the search space of PE candidates, whereas our approach uses the segmented vascular tree as a post-candidate-generator step to refine the detection results of the CG by constructing efficient classifiers. We claim that limiting the search space for PE detection is a overwhelming constraint which will potentially impair the PE detection accuracy, and the useful spatial information can be more effectively exploited in the classification stage after all possible candidates are identified in the full space.

Second, another characteristic of our approach is a novel learning formulation designed to build classifiers when multiple positive instances are present for one positive example. The goal of our multiple-positive-instance learning formulation is to classify at least one instance from each positive example correctly. By successfully converting the optimization problem into a quadratic program, we are able to solve the learning formulation using standard optimization tools such as CPLEX. The resulting algorithm is easy to implement. Computational results validate our algorithm by showing significantly improved detection accuracy and the reduced number of features needed in the detection process.

References

[1] J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3:1229–1243, 2003. 3, 7

[2] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. 7

[3] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997. 2

[4] I. C. Division. *ILOG CPLEX 6.5 Reference Manual*. Incline Village, NV, 1999. 7

[5] J. Fairfield. Toboggan contrast enhancement for contrast segmentation. In *IEEE Proceedings of 10th International Conference on Pattern Recognition*, volume 1, pages 712–716, Atlantic City, June 1990. 2

[6] G. Fung, M. Dundar, B. Krishnapuram, and B. R. Rao. Multiple instance algorithms for computer aided diagnosis. In *Advances in Neural Information Processing Systems*, 2006. 5

[7] J. Liang and J. Bi. Computer aided detection of pulmonary embolism with Tobogganing in CT angiography. In *Proceedings of the 20th International Conference on Information Processing in Medical Imaging*, 2007. 3

[8] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. *Advances in Neural Information Processing Systems*, 10:570–576, 1998. 2

[9] Y. Masutani, H. MacMahon, and K. Doi. Computerized detection of pulmonary embolism in spiral CT angiography based on volumetric image analysis. *IEEE Transactions on Medical Imaging*, 21(12):1517–1523, 2002. 1

[10] S. Park, C. Bajaj, G. Gladish, and D. Cody. Automatic pulmonary embolos detection and visualization. poster at http://www.ices.utexas.edu/∼smpark, 2004. 1, 2

[11] M. Quist, H. Bouma, C. van Kuijk, O. van Delden, and F. Gerritsen. Computer aided detection of pulmonary embolism on multi-detector CT. In *RSNA*, page (Conference Poster), Chicago, USA, Nov 2004. 1, 2

[12] J. Stoeckel and G. Fung. A mathematical programming approach for automatic classification of spect images of alzheimer's disease. In *Proceedings of the 5th IEEE International Conference on Data Mining*, 2005. 6

[13] V. Vural, G. Fung, and B. R. B. Krishnapuram, J. Dy. Batch-wise classification with applications to computer aided diagnosis. In *Proceedings of the 17th European Conference on Machine Learning*, 2006. 6

[14] C. Zhou, L. M. Hadjiiski, S. Patel, H.-P. Chan, and B. Chan. Computerized detection of pulmonary embolism in 3D computed tomographic (CT) images. In *RSNA*, page (abstract), Chicago, USA, Nov 2004. 1, 2

[15] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004. 3