IOP PUBLISHING PHYSICS IN MEDICINE AND BIOLOGY

Phys. Med. Biol. **54** (2009) 1555–1563 doi:10.1088/0031-9155/54/6/010

Markerless gating for lung cancer radiotherapy based on machine learning techniques

Tong Lin[1,2], Ruijiang Li[1], Xiaoli Tang[1], Jennifer G Dy[3] and Steve B Jiang[1]

[1] Department of Radiation Oncology, University of California San Diego, La Jolla, CA 92093, USA
[2] Key Laboratory of Machine Perception (Ministry of Education), School of EECS, Peking University, Beijing 100871, People's Republic of China
[3] Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, USA

E-mail: sbjiang@ucsd.edu

Received 24 July 2008, in final form 09 January 2009
Published 19 February 2009
Online at stacks.iop.org/PMB/54/1555

Abstract

In lung cancer radiotherapy, radiation to a mobile target can be delivered by respiratory gating, for which we need to know whether the target is inside or outside a predefined gating window at any time point during the treatment. This can be achieved by tracking one or more fiducial markers implanted inside or near the target, either fluoroscopically or electromagnetically. However, the clinical implementation of marker tracking is limited for lung cancer radiotherapy mainly due to the risk of pneumothorax. Therefore, gating without implanted fiducial markers is a promising clinical direction. We have developed several template-matching methods for fluoroscopic marker-less gating. Recently, we have modeled the gating problem as a binary pattern classification problem, in which principal component analysis (PCA) and support vector machine (SVM) are combined to perform the classification task. Following the same framework, we investigated different combinations of dimensionality reduction techniques (PCA and four nonlinear manifold learning methods) and two machine learning classification methods (artificial neural networks—ANN and SVM). Performance was evaluated on ten fluoroscopic image sequences of nine lung cancer patients. We found that among all combinations of dimensionality reduction techniques and classification methods, PCA combined with either ANN or SVM achieved a better performance than the other nonlinear manifold learning methods. ANN when combined with PCA achieves a better performance than SVM in terms of classification accuracy and recall rate, although the target coverage is similar for the two classification methods. Furthermore, the running time for both ANN and SVM with PCA is within tolerance for real-time applications. Overall,

ANN combined with PCA is a better candidate than other combinations we investigated in this work for real-time gated radiotherapy.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

A major uncertainty in treating lung cancer with radiation is the respiratory lung tumor motion, which can be clinically significant for some patients. Respiratory gated lung cancer radiotherapy holds promise to precisely deliver prescribed radiation dose to the tumor, while minimizing the incidence and severity of normal tissue complications (Jiang 2006a). Respiratory gating limits radiation exposure to a portion of the breathing cycle when the tumor is in a predefined gating window. Due to a reduced planning target volume, precise target localization in real time is particularly important for gated radiotherapy (Jiang 2006b).

We have developed several template-matching methods which have been proved feasible for solving the fluoroscopic markerless gating problem (Berbeco *et al* 2005, Cui *et al* 2007). However, template matching does not utilize the information outside the gating window in building templates, which could be important for improving the accuracy and robustness of the algorithm. Recently, we have modeled the gating problem as a binary classification problem and solved it by means of a support vector machine (SVM) classifier (Vapnik and Cortes 1995) combined with a dimensionality reduction technique called principal components analysis (PCA) (Alpaydin 2004). This method achieved slightly higher accuracy compared to template matching at the price of significant amount of training time to search the optimal parameter set (Cui *et al* 2008).

Dimensionality reduction and classification are two essential parts in machine learning algorithms (Alpaydin 2004). Selecting the right dimensionality reduction technique and classification algorithm can yield an optimal final result. In this work, we adopt the same binary classification framework as proposed previously (Cui *et al* 2008). Within this framework, we investigate four other dimensionality reduction techniques besides PCA, namely locally linear embedding (LLE), local tangent space alignment (LTSA), Laplacian eigenmap (LAP) and diffusion maps (DMAP) (Lin and Zhang 2008). PCA is one of the most widely used dimensionality reduction techniques. It finds the best linear representation of the data in the mean-square sense. Unlike PCA, these four algorithms belong to manifold learning algorithms, which represent the latest nonlinear dimensionality reduction techniques. In recent research literatures, these algorithms have demonstrated better performance than classical dimensionality reduction techniques such as PCA, especially on curved and nonlinear data sets (Roweis and Saul 2000). Specifically, the benefit of any manifold learning method over classical methods was shown in the famous Swiss roll data, which are curved and nonlinear. Linear methods such as PCA often fail while most manifold learning methods can successfully model the Swiss roll data. In exploratory data analysis, the linearity assumption is not warranted. It is therefore also important to investigate nonlinear dimensionality reduction techniques despite their higher computational cost. For classification, in addition to SVM, we combine the dimensionality reduction techniques with a three-layer artificial neural network (ANN) for gated lung cancer radiotherapy. The performance of the algorithms is evaluated in a retrospective fashion on ten fluoroscopic video sequences of nine patients. We will compare the performance of ANN with SVM when combined with the aforementioned five dimensionality reduction techniques.

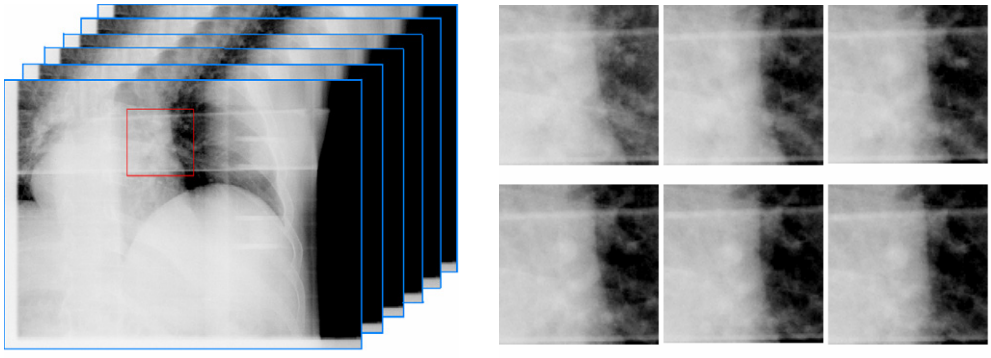

Figure 1. A set of training fluoroscopic image sequence (left) and the selected ROI images (right).

This paper is organized as follows. In section 2, we present the proposed algorithms in detail. Section 3 describes the experiment set-up and evaluation metrics used to test the proposed algorithm. Section 4 reports the experimental results. Finally, section 5 presents our conclusions.

2. Methods and materials

The goal of gated radiotherapy is to decide when to turn the beam on or off. As such, the gating problem can be reformulated as a binary classification problem. The goal now becomes finding a decision boundary that will separate future fluoroscopic images into beam ON or OFF class. The purpose of the classification algorithms is to learn an optimal decision boundary in a well-defined sense based on a subset of the entire data set called the training set.

2.1. Training fluoroscopic images

Prior to treatment, a sequence of fluoroscopic images are acquired for training purposes. A region-of-interest (ROI) that includes the tumor motion is selected on the training fluoroscopic images. Each ROI is manually labeled by human expert observers with either class beam ON or beam OFF based on the gating window size determined during treatment planning. In patients where the tumor was hard to identify visually in the fluoroscopic images, an anatomical structure nearby the tumor was used.

Figure 1 shows an example of a set of training fluoroscopic images and the selected ROIs. The first two ROIs are within the gating window and labeled as beam ON class, and the next four are outside the gating window and thus labeled as beam OFF class.

2.2. Dimensionality reduction techniques

The purpose of dimensionality reduction is twofold. First, it is to reduce the training sample size and therefore decrease the computational cost. A typical ROI size could be 100×100 pixels. This means that the dimensionality of a training sample would be $100 \times 100 = 10\,000$ (assuming each pixel in the ROI is independent). Significant computational time and resources would be needed with these high dimensional samples. This is simply not practical for real-time gated radiotherapy. With dimensionality reduction, the dimensionalities of the training samples can be significantly reduced, and consequently, much less computational time and resource would be needed. Secondly, it is to extract significant features of the ROI

automatically. A dimensionality reduction technique will automatically sort the information based on its importance and retain the most important components. For instance, PCA involves a mathematical procedure that transforms the original correlated variables into a small number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

Besides PCA, four other dimensionality reduction techniques are investigated in this work, including LLE, LTSA, LAP and DMAP (Lin and Zhang 2008). These four manifold learning methods find a nonlinear transformation by preserving neighborhood distances. Most manifold learning methods need two key parameters, one to describe the neighborhood size and one to describe the intrinsic dimension or output dimension. The results might be very different if these two parameters are varied. A detailed comparison of different dimensionality reduction techniques can be found in Lin and Zhang (2008).

We first map each training ROI image into a 30-dimensional linear space using PCA. Then, we further reduce the dimensionality of the training ROI image to 10 using the five methods described above. These features lie in a significantly lower dimensional space compared with the original ROI image and will be fed later into the classification algorithms as the input.

2.3. Artificial neural network

An ANN is an effective computational model for pattern classification and function approximation (or regression analysis). It is inspired by the way the biological nervous system processes information. ANN includes massively parallel systems with large numbers of interconnected simple processors, and it can solve many challenging computational problems. An ANN can learn any arbitrarily complex function by adding neurons and layers to the network. A function is learned by adjusting the weights of the network. These weights are tuned by minimizing a least-squares error optimization function through a back-propagation algorithm. For this gating problem, we employ a standard three-layer neural network with an error back-propagation algorithm. Figure 2 illustrates a three-layer neural network. In our gating system, the input layer has ten neurons to match the ten-dimensional input data after dimensionality reduction. There are five neurons in the hidden layer and only one neuron in the output layer.

A trained ANN can be thought of as an 'expert' in the category of information it has been given to analyze. In our gated lung radiotherapy application, a trained ANN automatically processes the fluoroscopic images acquired during the simulated treatment and classify them into the beam ON or OFF class and then generate the corresponding gating signals.

2.4. Simulated treatment delivery

During simulated treatment in this retrospective study, ROI at the same location as in the training fluoroscopic images is automatically selected on each new fluoroscopic image acquired during simulated treatment. The same dimensionality reduction technique is applied on the new ROIs. The trained ANN is then applied to automatically process each ROI with the reduced dimensionality and classifies it into either the beam ON or OFF class. Gating signals are then generated accordingly.

3. Experimental setup and evaluation metrics

For this study, ten fluoroscopic image sequences of nine lung cancer patients have been acquired at University of California San Diego (UCSD) using a Varian on-board imaging (OBI) system

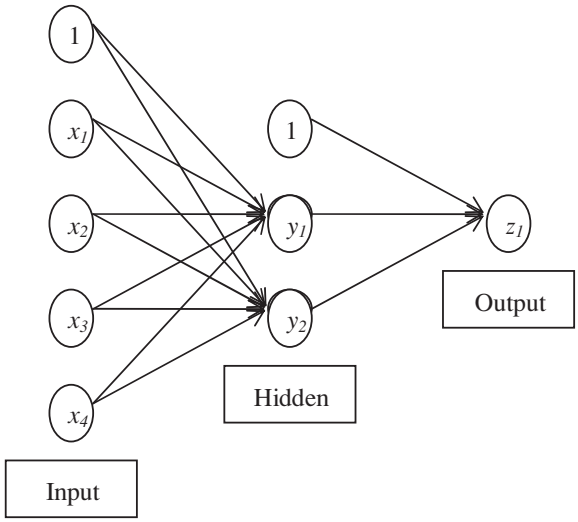

Figure 2. A simple neural network with three layers neurons. In this figure, there are four neurons in the input layer, two neurons in the hidden layer and one neuron in the output layer.

Table 1. Categories of classification results.

		Benchmark	
		Beam ON	Beam OFF
Classification results	Beam ON	True positive	False positive
	Beam OFF	False negative	True negative

(Varian Medical Systems, Palo Alto, CA, USA). The fluoroscopic image frequency is 15 Hz, and each image size is 1024 × 768 pixels. The average image sequence length is about 40 s (i.e. 600 frames). For each patient, 15 s of fluoroscopic images (225 frames) at the beginning of the sequence are used as training data. The rest of the frames are used as testing data for validation purpose. We used CMU's ANN implementation in C (CMU 1995) and Libsvm's SVM implementation (Chang and Lin 2007) in C++. Both implementations were wrapped into DLLs that can be called from MatLab (MathWorks, Natick, MA, USA). All other proposed computational frameworks were implemented on MatLab 7.3.

We define the true positive (*tp*), false positive (*fp*), true negative (*tn*) and false negative (*fn*) in table 1. Figure 3 shows an example of *tp*, *fp*, *tn* and *fn*. A sequence of fluoroscopic images are mapped into a three-dimensional space as shown in figure 3(a), where the blue points belong to the true beam ON class and the red points are from the true beam OFF class. Figure 3(b) shows the classification results in terms of *tp*, *fp*, *tn* and *fn*, where different colors represent different categories. Figure 4 shows an example of the gating signal generated by using the classification results. The corresponding *tp*, *fp*, *tn* and *fn* are displayed in the figure. The classification results are measured in terms of the classification accuracy (CA), the recall rate (RR) and the target coverage (TC), where $\mathrm{CA} = \frac{tp+tn}{\mathrm{all}}$, $\mathrm{RR} = \frac{tp}{tp+fn}$ and $\mathrm{TC} = \frac{tp}{tp+fp}$. Among these three metrics, CA and RR are common in machine learning field, while TC is more clinically relevant. TC is a first-order measure of the delivered target dose relative to the prescribed target dose. For instance, a TC of 80% means that 80% of the prescribed dose

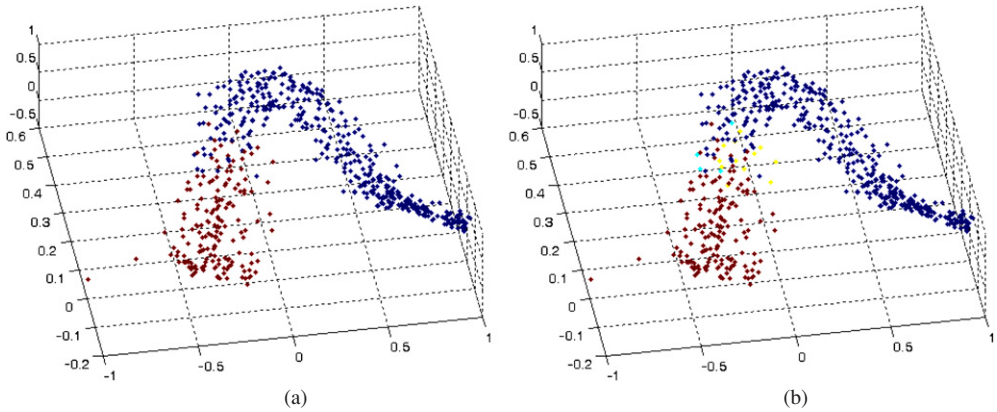

Figure 3. (a) 3D representation of the original fluoroscopic images using PCA projection, where red and blue dots indicate the true beam ON and beam OFF classes, respectively. (b) Classification results: red (*tp*), blue (*tn*), cyan (*fp*) and yellow (*fn*).

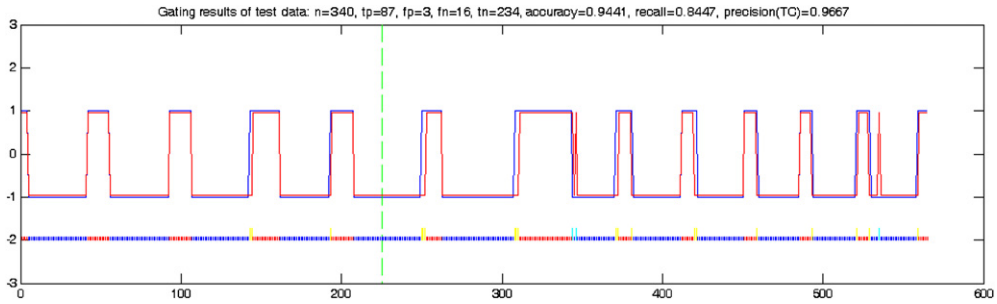

Figure 4. Top: gating signals with ground truth in blue and the result of the algorithm (PCA and combined with ANN) in red for fluoroscopic sequence 4; Bottom: four kinds of classification results: red (*tp*), blue (*tn*), cyan (*fp*) and yellow (*fn*). CA = 0.9441, RR = 0.8447 and TC = 0.9667. This differs slightly from table 2 since the result in table 2 for each sequence is an average of ten such runs from different initial weights of the ANN. The vertical green dashed line divides the images into training and test sets.

is delivered to the target and 20% delivered to the surrounding normal tissues. In Cui *et al* (2008), besides TC, duty cycle (DC) was used as the evaluation metric, which is defined as $\mathrm{DC} = \frac{tp}{\mathrm{all}}$. It is easy to see that given a particular data set with a ground truth gating signal, RR and DC are always proportional to each other independent of what algorithms are used. Thus, they can be considered as equivalent metrics for evaluation purpose.

4. Results and discussions

The experimental results are reported in tables 2 and 3. From table 2, we can see that the RR for fluoroscopic sequences 1, 2, 3, 4 and 9 is markedly lower than the other sequences for most dimensionality reduction techniques (except LTSA, whose average RR is much lower than others). A similar pattern may also be observed for SVM-based results in table 3. A low RR indicates a large number of false negatives compared with true positives, meaning

Table 2. Performance of various dimensionality reduction techniques combined with ANN. Three numbers in each cell represent the classification accuracy, the recall rate and the target coverage in percentage.

Fluoroscopic sequences	PCA	LLE	LTSA	LAP	DMAP
1	94.5, 79.1, 100	92.0, 80.6, 89.3	88.7, 65.7, 88.9	94.2, 81.3, 96.9	92.6, 81.7, 91.3
2	95.4, 82.7, 98.2	90.6, 82.3, 82.0	89.3, 65.0, 88.8	94.0, 80.9, 94.3	93.4, 84.2, 90.1
3	95.3, 82.6, 97.9	92.5, 80.8, 88.8	88.8, 65.1, 86.5	93.3, 81.9, 91.5	94.4, 84.0, 93.2
4	95.5, 83.3, 97.8	92.9, 83.6, 87.8	89.4, 65.1, 88.6	93.8, 82.1, 92.5	93.4, 85.6, 89.2
5	97.9, 95.8, 98.3	93.1, 95.9, 86.7	73.8, 72.1, 62.3	91.3, 96.5, 82.9	92.4, 97.8, 84.6
6	97.3, 97.4, 97.4	98.1, 98.4, 97.9	62.6, 50.0, 68.3	97.8, 96.8, 98.9	97.8, 98.9, 96.9
7	98.0, 95.4, 98.3	96.9, 98.0, 93.3	77.1, 73.3, 63.1	93.6, 98.5, 85.4	93.8, 99.0, 85.4
8	98.9, 98.4, 99.2	98.6, 99.2, 97.7	55.5, 29.1, 50.7	98.9, 97.6, 100	98.6, 96.9, 100
9	95.8, 87.0, 97.7	96.0, 88.4, 96.7	83.5, 44.2, 93.3	96.0, 88.0, 97.4	96.0, 87.8, 97.7
10	94.6, 97.5, 93.5	92.9, 95.7, 92.4	70.0, 94.8, 67.5	93.6, 97.5, 92.2	96.8, 99.0, 95.7
Mean	96.3, 89.9, 97.8	94.4, 90.3, 91.3	77.9, 62.4, 75.8	94.7, 90.1, 93.2	94.9, 91.5, 92.4
±SD	±1.6, ±7.6, ±1.7	±2.8, ±7.9, ±5.3	±12, ±18, ±15	±2.3, ±7.9, ±5.6	±2.2, ±7.4, ±5.2

Table 3. Performance of various dimensionality reduction techniques combined with SVM. Three numbers in each cell represent the classification accuracy, the recall rate and the target coverage in percentage.

Fluoroscopic sequences	PCA	LLE	LTSA	LAP	DMAP
1	92.6, 74.6, 97.9	92.0, 71.9, 97.3	87.9, 56.8, 97.0	87.6, 64.7, 89.2	90.9, 76.9, 91.5
2	94.5, 78.0, 99.1	91.4, 68.1, 95.8	89.0, 60.3, 94.2	87.3, 62.5, 86.3	93.7, 76.5, 97.5
3	94.4, 78.9, 97.8	91.9, 71.1, 95.1	88.7, 57.2, 96.1	89.0, 60.0, 94.2	94.0, 81.6, 94.6
4	94.0, 77.0, 98.3	91.5, 67.9, 96.1	88.2, 54.8, 96.6	88.5, 61.8, 92.4	93.1, 75.0, 96.9
5	97.0, 94.4, 97.3	94.5, 93.4, 92.6	88.1, 80.1, 88.0	89.8, 95.6, 82.4	94.3, 94.6, 91.6
6	96.7, 92.4, 99.1	86.9, 66.8, 98.7	76.8, 41.6, 96.6	96.6, 92.1, 99.0	93.1, 82.9, 99.0
7	92.2, 77.2, 98.4	95.0, 89.5, 94.9	90.9, 89.4, 84.5	93.7, 94.4, 89.1	96.4, 90.8, 98.2
8	96.8, 93.9, 97.0	95.6, 90.4, 97.0	78.0, 40.9, 93.7	96.7, 93.2, 97.3	96.3, 92.2, 97.0
9	94.7, 84.2, 96.4	95.6, 89.0, 95.1	89.9, 69.1, 92.8	95.0, 83.6, 98.3	92.6, 75.5, 97.0
10	96.5, 98.4, 95.8	95.2, 97.8, 94.3	78.1, 90.3, 76.6	92.8, 96.8, 91.4	96.4, 99.2, 95.0
Mean	94.9, 84.9, 97.7	93.0, 80.6, 95.7	85.6, 64.1, 91.6	91.7, 80.5, 92.0	94.1, 84.5, 95.8
±SD	±1.7, ±9.0, ±1.1	±2.8, ±12.4, ±1.7	±5.6, ±18, ±6.7	±3.7, ±16, ±5.4	±1.8, ±9.0, ±2.6

that the radiation beam will be turned off more frequently when it should be on, leading to a prolonged treatment. At this point, we wish to point out that the three evaluation metrics used here are not equally important in clinical sense. A false positive gating signal is much worse clinically than a false negative one. Therefore, when comparing with two results, TC should be given more weight than RR. It is comforting to see that in most cases, TC is in the upper 90% although RR seldom reaches 90%.

The reason for a lower RR is very complicated. It may be because the tumor cannot be seen clearly in the fluoroscopic images and the 'ground truth' itself is not accurate. There may be shape and size changes in the tumor as well as interfering movements of other organs (particularly heart) during simulated treatment. These variations will be captured by PCA even if the tumor is in the same location. Another potential issue is the increase of breathing

amplitude during simulated treatment beyond that of the training set. Since these new images are not represented in the training set, they may be mapped incorrectly in the PCA subspace.

From table 2, the average *CA and TC* values of the ANN and PCA combination are greater than other combinations with the smallest standard deviation although the RR for PCA with ANN is slightly lower than several other manifold learning methods with ANN. Similarly, we can see from table 3 that the SVM and PCA combination performs better than other combinations too. This suggests that for the gating problem, a linear representation of the data using PCA might suffice for a good classification result with ANN or SVM.

Comparing table 2 with table 3, we can see that when PCA is used for dimensionality reduction, there is a slight advantage of using ANN over SVM in all the three evaluation metrics. In order to quantitatively assess the performance gain (if any) of ANN over SVM when used with PCA, we performed a one-side paired *t*-test on three evaluation metrics for the ten fluoroscopic sequences. The reason why we used this particular *t*-test is that we are suspecting that ANN is better than SVM (hence one-side) and there is a one-to-one correspondence between each entry in tables 2 and 3 since they came from the same sequence in the same patient (hence paired). We rejected the null hypothesis for CA and RR under the 0.05 significance level (*p*-value: 0.024 and 0.006, respectively). This means that the gain of ANN over SVM in terms of CA and RR is statistically significant. However, we cannot reject the null hypothesis for TC (*p*-value: 0.405). Combining these results with those of Cui *et al* (2008), we can say that ANN with PCA is better than the template-based approach in terms of TC and RR (or DC).

Another important issue for real-time applications is the computational cost (or latency) in the system. Training an ANN is more efficient than training SVM since SVM needs a brute force search for the best combination of two parameters. The running time of training ANN ten times is about 0.1521 s on an Intel Core 2 Duo 2.66 GHz Machine, while it takes 2.6332 s on training SVM to search the best parameters and learn the SVM decision boundary. However, since the training process is done before treatment and only needs to be done once, the training time may be of secondary concern. What is more relevant here is the running time in the testing phase. For this particular data set, we estimated that the average running time for ANN in combination with PCA is about 6.7 ms for each frame. For SVM combined with PCA, it takes about 11 ms to process each frame. This is consistent with the previous literature in that SVM in general is slower in terms of running time than ANN for a similar generalization performance (Haykin 1994). One of the reasons is that there is no control over the number of data samples selected by SVM for use as support vectors (during testing, the running time of SVM grows linearly with respect to the number of support vectors). This is in contrast to ANN, which has fixed computational complexity once a network architecture is selected. In this respect, ANN is more appealing than SVM. At this stage, it seems that the system latency for both ANN and SVM when combined with PCA is acceptable for real-time gated radiotherapy.

5. Conclusions and future work

In this work, the gating problem was reformulated as a binary classification problem. Five dimensionality reduction techniques and two machine learning classification approaches were investigated. We found that PCA is in general superior compared to the other four nonlinear manifold learning methods for our application, suggesting that our data might lie in a (roughly) linear space. PCA combined with ANN leads to more accurate results in terms of CA, RR than SVM although TC is similar for both methods. In particular, for the clinically meaningful performance measure, we can achieve an upper 90% TC for most sequences in our data

set. Both ANN and SVM when combined with PCA give a system latency within tolerance for real-time applications. Therefore, for our future real-time gated lung radiotherapy, we recommend PCA combined with ANN.

To overcome the potential problems mentioned in section 4, we may need to resort to fluoroscopic images with implanted markers to get more accurate ground truth in the future. In some cases, using a smaller ROI may reduce the amount of interfering movements of other organs and increase classification accuracy. Some preprocessing such as low pass filtering may help reduce the effects of background noise. In general, the issue of shape and size changes in the tumor as well as irregular breathing patterns is more difficult to deal with. It is obvious that PCA is sensitive to these changes since it is an area-based approach and works directly with image intensity values. We plan to investigate other feature extraction techniques or special-purpose image processing and computer vision algorithms in the future to overcome this problem.

Acknowledgments

The project is partially supported by an NCI grant (1 R21 CA110177 A 01A1), a National Science Foundation of China (NSFC) Grant 60775006, a National Key Basic Research Program of China (NKBRP) Grant 2004CB318005.

References

Alpaydin E 2004 *Introduction to Machine Learning* (Cambridge, MA: MIT Press)

Berbeco R I, Mostafavi H, Sharp G C and Jiang S B 2005 Towards fluoroscopic respiratory gating for lung tumours without radiopaque markers *Phys. Med. Biol.* **50** 4481–90

Chang C C and Lin C J 2007 LIBSVM: a library for support vector machines (http://www.csie.ntu.edu.tw/∼cjlin/papers/libsvm.pdf)

CMU 1995 Neural networks benchmarks (http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/neural/bench/0.html)

Cui Y, Dy J G, Alexander B and Jiang S B 2008 Fluoroscopic gating without implanted fiducial markers for lung cancer radiotherapy based on support vector machines (SVM) *Phys. Med. Biol.* **53** N315–27

Cui Y, Dy J G, Sharp G C, Alexander B and Jiang S B 2007 Robust fluoroscopic respiratory gating for lung cancer radiotherapy without implanted fiducial markers *Phys. Med. Biol.* **52** 741–55

Haykin S 1994 *Neural Networks: A Comprehensive Foundation* (Englewood Cliffs, NJ: Prentice-Hall)

Jiang S B 2006a Radiotherapy of mobile tumors *Semin. Radiat. Oncol.* **16** 239–48

Jiang S B 2006b Technical aspects of image-guided respiration-gated radiation therapy *Med. Dosim.* **31** 141–51

Lin T and Zhang H 2008 Riemannian manifold learning *IEEE Trans. Pattern. Anal. Mach. Intell.* **30** 796–809

Roweis S T and Saul L K 2000 Nonlinear dimensionality reduction by locally linear embedding *Science* **290** 2323–26

Vapnik V and Cortes C 1995 Support vector networks *Mach. Learn.* **20** 273–97