34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Mining Pattern Sequences in Respiratory Tumor Motion Data

Arvind Balasubramanian, Balakrishnan Prabhakaran, *The University of Texas at Dallas*
Amit Sawant, *UT Southwestern Medical Center*

Abstract— **Management of respiration induced tumor motion during radiation therapy is crucial to effective treatment. Pattern sequences in the tumor motion signals can be valuable features in the analysis and prediction of irregular tumor motion. In this study, we put forward an approach towards mining pattern sequences in respiratory tumor motion data. We discuss the use of pattern sequence distributions as effective representations of motion characteristics, and find similarities between individual tumor motion instances. We also explore grouping of patients based on similarities in pattern sequence distributions exhibited by their respiratory motion traces.**

I. Introduction

Respiratory motion causes significant errors in dose delivery to tumors in the thorax and abdomen. Such errors reduce the efficacy of radiation therapy due to the fact that tumors may receive less than prescribed dose (thereby not achieving the desired cell-kill) or normal tissue and critical organs may receive more than intended dose (thereby causing excessive radiation-related toxicity). Effective management of respiratory motion is key to achieving the clinical goals of thoracic and abdominal radiotherapy.

While a large body of research exists on respiratory tumor motion prediction and management [1][5], one aspect that has not received as much attention is the personalization of respiratory management strategies based on rigorous analysis and profiling of respiratory patterns. The intrinsic characteristics in respiratory motion offer a lot of information regarding the nature of motion in terms of regularity, motion range etc. Since the motion of thoracic and abdominal tumors is heavily influenced by respiration, it is fairly obvious that a person's breathing patterns would be reflected in the tumor motion traces. If we assume that patients with similar breathing styles and patterns may be addressed as a type or profile, and would exhibit similar irregularities in tumor motion, then physicians may develop customized motion-management and treatment plans to address each such patient profile.

In this work, we study intra- and inter-patient respiration-induced tumor motion patterns in order to (a) identify pattern sequences in the motion traces and (b) explore grouping of patients based on similarities of pattern sequences exhibited by their respiratory motion traces. We propose an approach to depict a motion signal in terms of the pattern sequences which constitute the signal. We also discuss the use of pattern sequence distributions in identifying similarities between motion signals. Grouping of patients based on similarities between motion pattern sequence histograms and its validation is also explored.

Arvind Balasubramanian and Balakrishnan Prabhakaran are with the University of Texas at Dallas, Richardson, TX 75080, USA (e-mail: {arvind, praba}@ utdallas.edu).

Amit Sawant is with UT Southwestern Medical Center, Dallas, TX 75390, USA (e-mail: amit.sawant@utsouthwestern.edu).

II. Method

A. Dataset

The respiratory tumor motion data used in this study is modeled by the CyberKnife Synchrony system [3] and is taken from the dataset created and documented by Suh et al [2] for 143 treatment fractions in 42 patients. The tumor motion is estimated from correlations between external on-body markers and internal fiducial markers implanted around, or sometimes within, the tumor mass, and monitored using periodic stereoscopic x-ray imaging. The data provides modeled 3D coordinate location of the tumor in time, documenting the tumor motion (in millimeters) along three dimensions – Superior-Inferior (SI), Anterior-Posterior (AP) and Left-Right (LR) – as a function of time, at a sampling rate of 25 Hz. The validity of the estimated data has been discussed in previous works, with Seppenwoolde et al [4] finding the systematic error of position estimation to be less than 1 mm for all patients and mean 3D error to be less than 2 mm for 80% of the time. The mean and standard deviation of the 3D position estimation root mean square error documented in the dataset is 1.5 ± 0.8 mm.

According to the AAPM Task Group 76 [1], respiratory management techniques are required when the range of respiratory tumor motion is greater than 0.5 cm in any direction. In the available dataset, the resultant motion calculated from individual coordinate data had a mean greater than 0.5 cm in 56 treatment fractions. These constitute the data used in this analysis. Basic motion signal statistics and ranges for these 56 instances are documented in Table I.

TABLE I. Tumor motion data (Statistical values and ranges)

Number of motion instances	Peak-to-trough distance (cm) (Amplitude)		Peak-to-peak time (sec) (Respiratory period)	
	Mean	SD	Mean	SD
56	0.79 (0.51 - 1.44)	0.25 (0.06 - 0.73)	3.91 (2.52 - 6.37)	0.79 (0.22 - 1.73)

Two of the prominent types of irregular or anomalous motion for which respiratory motion management is necessary are (i) motion outliers – these can be described as periods of irregular motion involving substantial to extreme temporary displacement of the tumor about its mean position. A special case of this can be an occasional lapse in inhalation or hypoxia leading to deep inhalation, and a possible irregular movement of the tumor; (ii) baseline shifts – these can be described as periods of irregular motion involving substantial displacement (temporary or permanent) in the mean position of the tumor. Examples of both types of anomalies are illustrated in Figure 1.

B. Segmentation

This analysis borrows from the previous study based on this dataset by Suh et al [2], and builds on the segmentation proposed in it. Individual respiratory cycles in a motion trace hold promise as primary units of analysis. Respiratory cycles not only enable pattern comparisons within a single patient's tumor motion data, but also across multiple patients. Therefore, motion information can be studied in terms of the features extracted from each of those cycles. The overall mean respiratory period is 3.8 seconds (calculated over 143 treatment fractions in 42 patients). The sampling rate for the respiratory motion data is 25 samples per second. Therefore, windows having a width of 100 samples (~ 4 seconds) were chosen as appropriate units for feature extraction (See Figure 2). Windows were segmented such that consecutive windows have an overlap of 50%. Using an overlapping window segmentation approach ensures that any sequence of patterns present in the original signal is preserved by emphasizing the continuity information between two consecutively segmented windows.

C. Feature Extraction

As is the case with any other time series data, representation and comparison of the patterns in the motion traces is heavily influenced by parameterization and feature selection. In this study, a number of features have been experimented with. In addition to statistical quantities such as mean and variance, features associated with the spatial and temporal displacement of the tumor, such as the respiratory period and the amplitude of motion were considered. Respiratory motion is primarily periodic in nature, and the periodicity in the tumor motion is a direct measure of the regularity in the patient's breathing. Frequency domain features have been included in the study to reflect the periodic characteristic in signals. These features are derived by transforming the pattern window into the frequency domain using a fast Fourier transform (FFT). Features such as spectral energy (sum of the squared FFT component magnitudes) and frequency-domain entropy (normalized information entropy of discrete FFT component magnitudes) were computed.

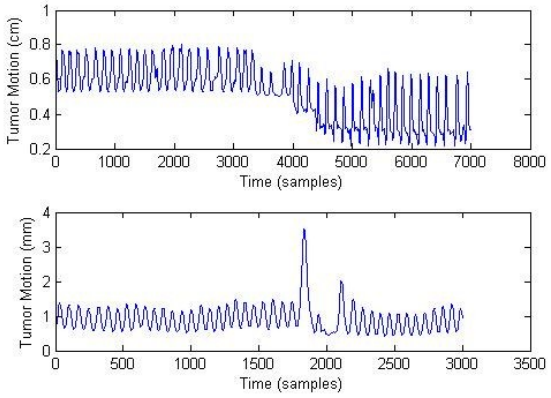

Figure 1. Tumor Motion anomalies – Baseline Shifts (top) and Motion Outliers (bottom).

D. String encoding of motion signals

The windows are then subjected to a clustering routine resulting in groupings of windows based on similarities between the extracted features. Since the features from each window represent the "pattern" in that segment of the signal that the feature window represents, each group would intuitively bring together signal segments exhibiting a similar pattern. By the same argument, windows containing similar motion outliers can be expected to be clustered together. The cluster assignment is an indicator of the primary pattern characteristic exhibited by the member segments in a particular cluster. So, the cluster assignments can be mapped back to segments in the motion signal which belong to that cluster. In other words, every motion signal can be encoded into a string of cluster assignments ordered in the same sequence as the windows they represent (See Figure 2). *k*-means clustering [6] was employed to cluster the feature windows extracted from the motion signals. The process would result in each feature window receiving a cluster assignment from 1, 2, 3… to *k*. The motion signals subsequently would be encoded into strings made up of these *k* characters. We conducted the clustering using a *k*-value of 5 clusters. While neighboring *k*-values did not yield any substantial differences, higher values were not used due to computational complexity.

Each motion signal represented as a string is segmented into *n*-grams (all possible sets of *n* consecutive feature windows in terms of n consecutive cluster assignments). These *n*-grams represent a progression or sequence of individual patterns. It is noteworthy that within each *n*-gram, every two consecutive windows represented by their cluster assignments would preserve continuity of the pattern sequence due to the overlap between the original segmented windows. The significance of a pattern sequence is that it captures not only a pattern corresponding to a possible motion outlier, but also the patterns corresponding to the

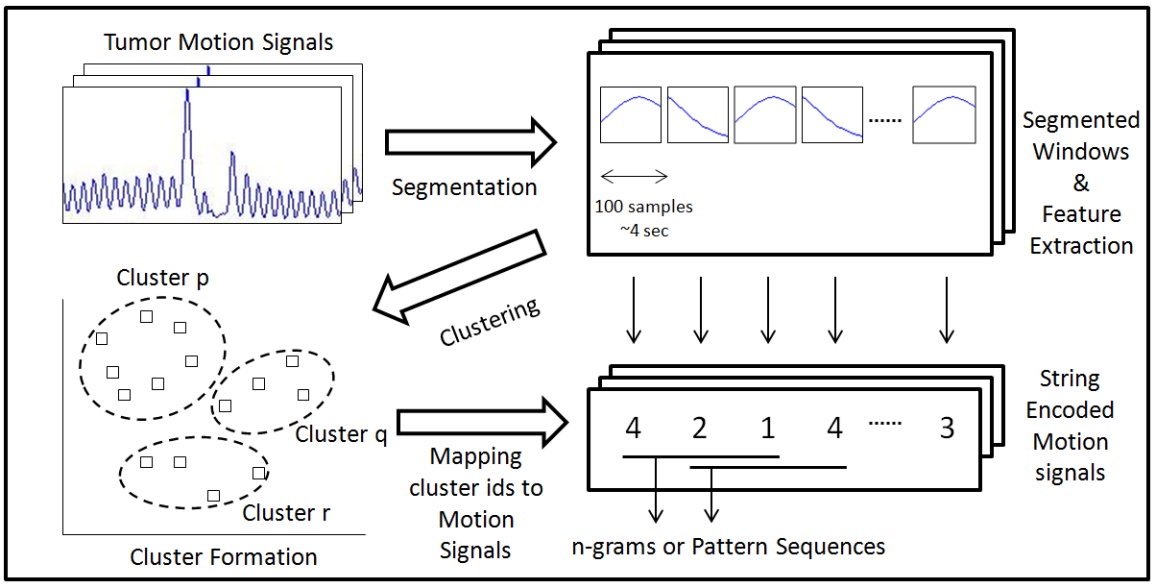

Figure 2. Illustration of the steps in the proposed method -- Segmentation, Feature Extraction, Clustering and String encoding of tumor motion signals.

segments preceding and following a motion outlier. This would help in preserving patterns sequences involving motion outliers that are prominent over a single motion trace or over multiple instances of tumor motion data. In our experiments, we employed 3-grams or trigrams to segment pattern sequences.

E. Pattern Sequence Histograms

In order to study the distribution of different patterns sequences in a motion signal, pattern sequence histograms are generated. These histograms can be used to examine the prominence of observed pattern sequences in the signal. Each bin in the histogram of a motion instance corresponds to a pattern sequence occurring in the motion signals, and its value is determined by its proportional contribution in the motion instance. A pattern sequence histogram can be considered as a signature of the original motion signal itself, since the histogram represents the signal in terms of the patterns that constitute it, and each of the pattern sequences retain the temporal ordering of such patterns in the signal. Figure 3 presents the pattern sequence histograms corresponding to a few motion traces.

F. Uses of Pattern Sequence Histograms

The similarities between the generated pattern sequence histograms can be an effective indicator of a possible grouping among patients whose respiratory tumor motion traces might show similar characteristics. We employ hierarchical agglomerative clustering to discover natural groups in the current dataset. This clustering approach begins with the individual motion instances and finishes with a single cluster, merging two similar groups in each iteration until all groups finally merge into one single group. Using the Euclidean distance as the similarity measure, a hierarchical cluster tree or dendrogram is obtained (see Figure 4a) that illustrates the grouping or "linking" performed in every iteration, from bottom to top. The distance between two linked groups is computed as the distance between the centroids of the two groups, and is represented by the the height of the link that joins the two groups in the dendrogram.

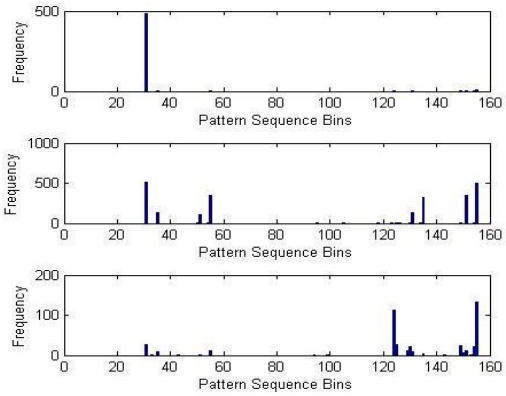

Figure 3. Sample Motion pattern sequence histograms in case of a 5-means clustering and trigram sequences.

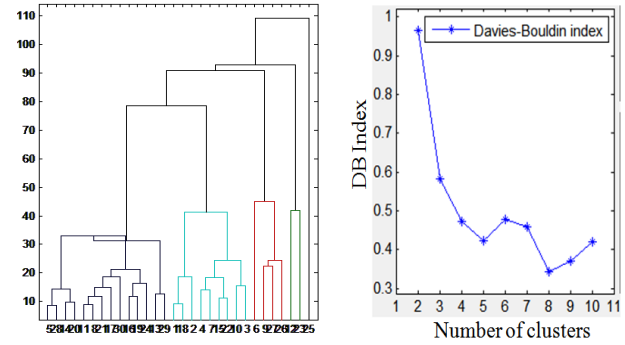

Figure 4. (a) Hierarchical cluster Tree for grouping pattern sequence histograms;(b) Variation of DB Index with number of clusters.

TABLE II. STATISTICAL VALUES FOR EACH CLUSTER OF TUMOR MOTION INSTANCES FORMED BY HIERARCHICALCLUSTERING OF PATTERN SEQUENCE BINS

Cluster Number	Number of motion instances	Peak-to-trough distance (cm) (Amplitude)		Peak-to-peak time (sec) (Respiratory period)	
		Mean	SD	Mean	SD
1	20	0.68	0.15	3.81	1.07
2	9	0.77	0.18	4.41	0.89
3	1	0.72	0	3.8	0
4	1	0.87	0	3.4	0
5	3	0.79	0.45	4.07	0.99
6	3	0.79	0.09	4.1	0.98
7	18	0.93	0.22	3.75	0.78
8	1	0.53	0	3.5	0

The groups formed are validated using the Davies-Boulding (DB) cluster validity index [7], which is based on the compactness and well-separation of the groups. An optimal clustering configuration is one that minimizes the DB Index value. We study the validity for different cluster configurations, by observing the variation of the DB Index with the increase in number of resultant clusters from 2 to 10 (Figure 4b). The DB Index is found to be the most optimal for a configuration of 8 clusters, the statistics for which are presented in Table II. Since the original dataset did not have any annotations for patient profile or type, validating the clusters based on specificity and sensitivity is not applicable here. It would however be desirable to identify features in the biometrics and breathing characteristics that maximize correlations among motion instances belonging to the individual natural groups produced by the hierarchical clustering. Such features can indicate the nature of possible classification or stratification of patients for personalization of respiratory management techniques.

The pattern sequence histogram over all the motion instances gives the overall prominence of pattern sequences. However, the discrimination between locally and globally prominent pattern sequences can be achieved by observing the bin-wise variance in the histogram values. The globally prominent pattern sequences would occur in more number of motion instances and therefore would typically have lower variance values in their respective histogram bins. This technique can be used to identify dominant pattern sequences consisting of motion outliers. A pattern sequence consisting of one or more motion outliers can be found to be prominent over a group of motion instances. Motion outlier pattern sequences that are common among a group of patients could be identified and categorized. Patients exhibiting a certain set of motion outlier pattern sequences in their respiratory motion data could also be studied for correlations in motion features.

Further studies can focus on the attributes addressed by pattern sequence histograms that can have application in patient profiling such as the prominence of pattern sequences in intra-patient and inter-patient data, the frequency of occurrence as well as the temporal distribution of the sequences along motion signals etc. These characteristics of patterns sequences can help identify their relevance to different groups of patients for whom personalized motion management and prediction techniques may prove beneficial.

CONCLUSION

Identifying dominant pattern sequences in tumor motion data is valuable in studying and potentially predicting the occurrences of irregular motion of abdominal and thoracic tumor. To serve this purpose, we have presented motion pattern sequence histograms as an effective representation of pattern progressions in motion signals. Further research into the correlations between the respiratory motion features of patients exhibiting similar anomalous motion events would be helpful in recognizing groups among patients for whom treatment can be specialized. In addition to existing methods of respiratory tumor motion management, it is necessary to explore strategies that would enable physicians to stratify patients based on similarities in the respiratory motion features for personalized treatment and therapy.

ACKNOWLEDGMENT

The authors are thankful to Dr. Yelin Suh for sharing the MATLAB program routines for preprocessing and segmentation of the respiration-induced tumor motion data.

REFERENCES

[1] P.J. Keall, et al., "The Management of Respiratory Motion in Radiation Oncology Report of AAPM Task Group 76", Medical Physics, vol. 33, no. 10, 2006, pp. 3874–3900.W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[2] Y. Suh, S. Dieterich, B. Cho, and P.J. Keall, "An analysis of thoracic and abdominal tumour motion for stereotactic body radiotherapy patients", Physics in medicine and biology, vol. 53, Jul. 2008, pp. 3623-40.

[3] J.R. Adler, S.D. Chang, M.J. Murphy, "The Cyberknife: A frameless robotic system for radiosurgery", Stereo Funct Neurosurg, 69 (1997), pp. 124–128E.

[4] Y. Seppenwoolde, R.I. Berbeco, S. Nishioka et al., "Accuracy of tumor motion compensation algorithm from a robotic respiratory tracking system: A simulation study", Med Phys, 34 (2007), pp. 2774–2784C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[5] P. Verma, H. Wu, M. Langer, I. Das, and G. Sandison, "Review of Real-time tumor motion prediction for Image Guided Radiation Treatment", Computing in Science and Engineering, vol. 99, 2010.

[6] MacQueen, J. B. (1967). "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297

[7] D. Davies and D. Bouldin, "A cluster separation measure", IEEE Trans. Pattern Anal. Machine Intell., 1:224–227, 1979.