Subsequence Matching on Structured Time Series Data

Huanmei Wu
Northeastern University
maggiewu@ccs.neu.edu

Betty Salzberg
Northeastern University
salzberg@ccs.neu.edu

Gregory C Sharp
Harvard Medical School
gcsharp@partners.org

Steve B Jiang
Harvard Medical School
Jiang.steve@mgh.harvard.edu

Hiroki Shirato
Hokkaido University
hshirato@radi.med.hokudai.ac.jp

David Kaeli
Northeastern University
kaeli@ece.neu.edu

ABSTRACT

Subsequence matching in time series databases is a useful technique, with applications in pattern matching, prediction, and rule discovery. Internal structure within the time series data can be used to improve these tasks, and provide important insight into the problem domain. This paper introduces our research effort in using the internal structure of a time series directly in the matching process. This idea is applied to the problem domain of respiratory motion data in cancer radiation treatment. We propose a comprehensive solution for analysis, clustering, and online prediction of respiratory motion using subsequence similarity matching. In this system, a motion signal is captured in real time as a data stream, and is analyzed immediately for treatment and also saved in a database for future study. A piecewise linear representation of the signal is generated from a finite state model, and is used as a query for subsequence matching. To ensure that the query subsequence is representative, we introduce the concept of *subsequence stability*, which can be used to dynamically adjust the query subsequence length. To satisfy the special needs of similarity matching over breathing patterns, a new subsequence similarity measure is introduced. This new measure uses a weighted L_1 distance function to capture the relative importance of each source stream, amplitude, frequency, and proximity in time. From the subsequence similarity measure, stream and patient similarity can be defined, which are then used for offline and online applications. The matching results are analyzed and applied for motion prediction and correlation discovery. While our system has been customized for use in radiation therapy, our approach to time series modeling is general enough for application domains with structured time series data.

1. INTRODUCTION

Modeling and analysis of time series stream data is a rich and rapidly growing research field. Time series stream data often arise when following industrial processes, monitoring patient treatments, or tracking corporate business metrics. Analysis of time series stream data is widely used for many applications such as economic forecasting, stock market analysis, process and quality control, budgetary analysis, and workload projections. In database research, there has been an explosion of interest on time series databases. Many high level representations of time series [4, 5, 7, 11, 14, 17, 22, 27], and distance functions for subsequence matching [1, 7, 23, 25] have been proposed. *Subsequence matching* methods that try to find subsequences similar to a query sequence within a large time series databases have attracted recent interest and many solutions have been proposed [1, 4, 7, 15, 16, 19, 23, 25].

However, far less attention has been paid to the internal structure within the data. Many time series display periodic fluctuations. For example, temperature of a region tends to peak in the summer and then declines in the fall. It reaches the lowest in the winter and then climbs up in the spring. So time series of temperature of a region will typically show the periodical seasonal changes. Periodicity is also quite common in economic time series and some medical time series. In addition to periodicity, there are other non-periodic structures in time series, such as trends, correlation and autocorrelation among time series data streams. An example is a plateau with exponential growth.

The internal structure gives meaning to a time series, which allows for a more accurate forecast. However, structured data generally contains random noise which makes it difficult to analyze a time series. Randomness cannot be predicted and it limits the certainty of future prediction.

Previous research on subsequence matching over structured time series data has not explicitly considered the meaning of the time series data and its influence over similarity matching. The internal structure and its meaning should be modeled and analyzed for two reasons. First, a model can be used in forecasting and monitoring a time series. Second, it helps to obtain an understanding of the underlying forces and structure that produced the corresponding time series.

This paper introduces our research effort in modeling, analysis, clustering and prediction of motion with structured time series data using subsequence similarity matching. In our research we consider the internal structure of a time series, the underlying meaning of the structure and the influence of the structure in subsequence matching. The data in our research is tumor respiratory motion stream data from image guided radiation therapy, which is one example of structured time series data. Although the solution is customized for tumor motion, our approach can be applied to other application domains with structured time series patterns. Following is a brief background introduction of this problem domain including the motivation and challenges for respiratory motion analysis.

SIGMOD 2005 June 14-16, 2005 Baltimore, Maryland, USA.

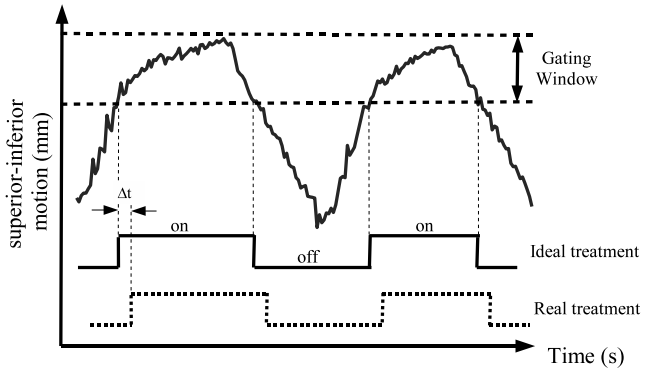

Figure 1: Respiration gating and latency.

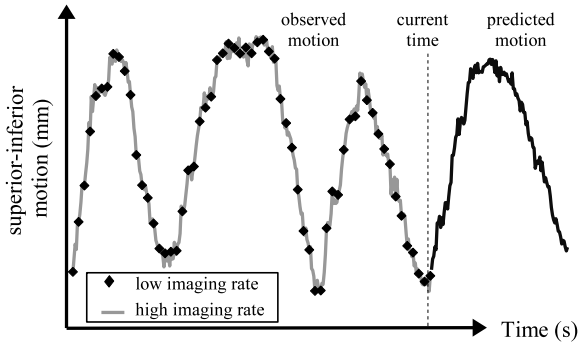

Figure 2: Prediction based on limited data.

Radiation therapy is a common treatment for cancers in the thoracic and abdominal regions. The goal of radiation therapy is to ensure precise radiation delivery to kill tumor cells. To avoid side effects, radiation to surrounding healthy tissues and critical structures must be minimized. However, the quality of radiation treatment is complicated by respiratory motion. As a patient inhales and exhales, tumors in the thoracic and abdominal regions move according to the breathing patterns. Therefore, effective radiation treatment of a moving tumor requires an adequate understanding of the motion characteristics.

In the medical literature, several strategies to compensate respiratory motion have been proposed, including *respiration gating* and *beam tracking* [12]. *Respiration gating* delivers radiation doses only when the tumor is in a predetermined location. It has a *gating window* for the radiation treatment beam, as shown in Figure 1. The tumor may move in or out of the gating window, and treatment is delivered when the tumor is in the gating window. *Beam Tracking* is another alternative method for precise dose delivery, in which the radiation beam follows the tumor dynamically. Both require online prediction of tumor position, primarily due to *system latency* and *imaging rate*.

System latency is the time interval between the time when the tumor is at a certain location and the time when the radiation beam can be turned on for treatment at that location. System latency is always an issue because of the time needed for image acquisition, image processing, and radiation system processing. Precise dose delivery is heavily impacted by system latencies. If treatment is based on the *last observed* position rather than the *current* position, this latency will reduce the effectiveness and efficiency of treating a moving tumor. Figure 1 illustrates the effects of system latency on gated treatment. The ideal treatment is the radiation beam on/off signal without any system latency. The real treatment is the treatment signal, without consideration of any system latency (total system latency is Δt), which treats the tumor at the last observed position.

The *imaging rate* is the number of images taken in a given time period. It is governed by radiation safety limits. Only a limited number of diagnostic images are allowed to be collected, because excess imaging is toxic to healthy tissue. However, knowledge of the true tumor positions is crucial for effective real time radiation therapy. Thus tumor position between two adjacent measurements needs to be predicted, especially when the sampling rate is low.

The first goal of this research is to predict tumor motion in real-time. There are a number of technical hurdles that must be overcome to be successful. First, respiratory motion is patient specific. Simply using another patient's breathing pattern to predict the current patient's tumor motion is not satisfactory. On the other hand, there is only limited data that can be used if prediction is based solely on the historical data of the same patient. Second, respiratory motion can be very complicated to predict, even for the same patient. The motion is non-uniform between two adjacent measurements, which can be observed in Figure 2. Furthermore, tumor motion varies from one breathing period to another, and can include frequency changes (duration changes for different breathing cycles), amplitude changes (spatial position changes of a tumor during breathing), base line shifting (tumor position changes at the end of exhale), or combinations of these effects. This is illustrated in Figure 3a and b. Third, the raw tracking signal is very noisy, as manifested in Figure 3c and d, which includes two kinds of noise: one is cardiac motion (tumor motion due to heart beat) and the other is spike noise. Cardiac motion is a major contributor to noise by adding short-term oscillations to long term breathing signals. Spike noise is an artifact of the data acquisition process and exists in both regular and non-regular breathing.

The second goal of this research is to find a correlation between respiratory motion and patient physiological conditions. If such a correlation can be reliably established, it will provide multiple benefits. On one hand, tumor motion prediction can then be guided by a patient's physiological condition and changes to that condition. Alternatively, motion patterns can help to identify changes in a patient's physiological condition. One example is the correlation between respiratory motion and common pulmonary (lung) pathology.

However, finding potential correlations between patient information and respiratory motion is a very difficult task. A patient's physiological information includes patient characteristics (e.g., sex and age of the patient), a patient's physical condition (e.g., cough or fever), past historical treatment data (e.g., medications), tumor characteristics (e.g., tumor size and location, and whether the tumor is a primary occurrence, recurrence or metastasis tumor), and treatment conditions (e.g., the marker position and marker size). Also, it is required that the correlation must be flexible enough to compensate for changes in amplitude and frequency during the treatment, which makes correlation discovery more challenging.

Although recently there are research efforts in characterization and prediction [3, 20, 24, 26], the best parameterization for respiratory

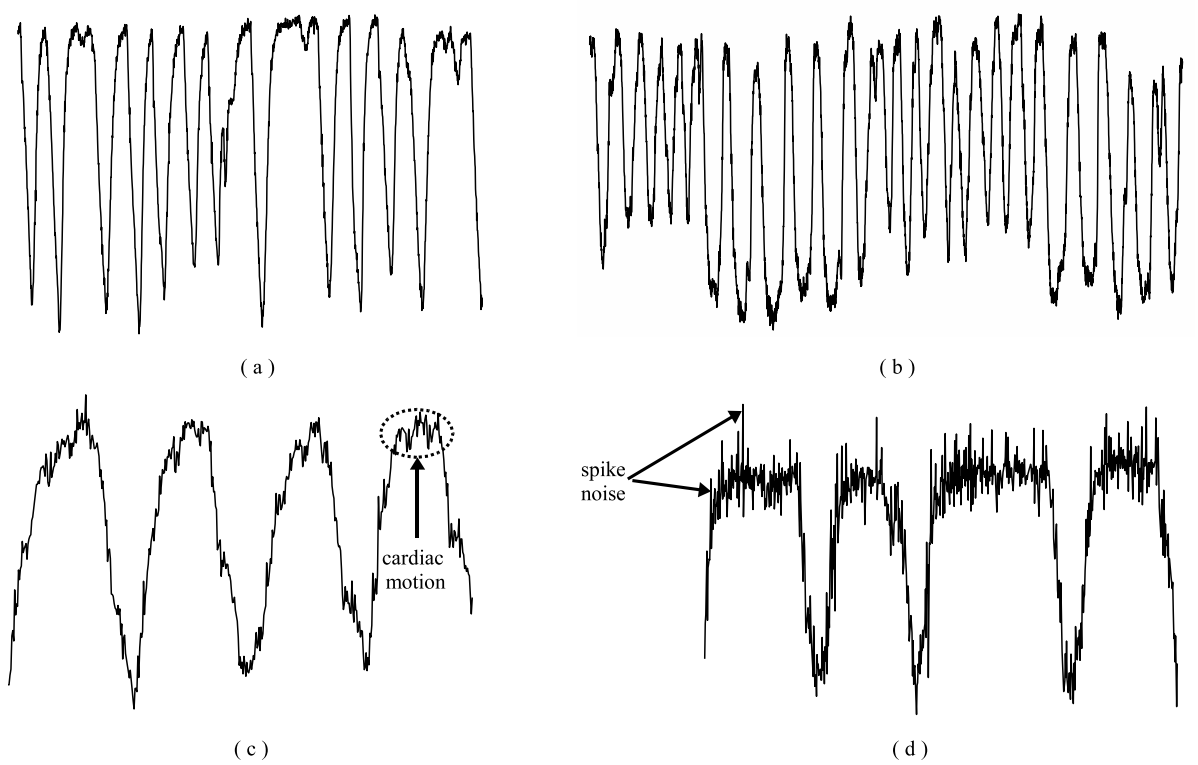

Figure 3: Complex tumor motion. (a) with amplitude and frequency changes, (b) with base line shifts, amplitude and frequency changes, (c) with cardiac motion, (d) with cardiac motion and spike noise.

motion analysis and prediction remains an open question. Subsequence similarity is an ideal technique because it has advantages for pattern matching, future movement prediction, rule discovery and computer-aided diagnosis. So we propose a solution using subsequence similarity matching for tumor respiratory motion analysis, correlation discovery and motion prediction. Our solution has addressed the special concerns for medical problems, and is suitable for both online and offline applications. Our main contributions are the following:

- We introduce a new concept, *subsequence stability*, to evaluate the qualitative representativeness of a given subsequence. We adopt a flexible criteria to generate query subsequences online. The length of a query subsequence is dynamically adjusted based on the subsequence's stability of the most recent motion.
- We propose a model-based, multi-layer, weighted, and parametric subsequence similarity measure, which takes into account specific concerns for both online and offline respiratory motion analysis. This subsequence similarity measure can be generalized to other applications by adjusting parameter values.
- We define distance functions between whole streams and patients for motion analysis. They are defined based on subsequence similarity and they provide a convenient way to correlate tumor motion with other patient information.
- We analyze similarity matching results for tumor motion prediction used during image-guided dynamic radiation treatment and correlation discovery. The statistical results are valuable both in treatment planning and physiological diagnosis, which will result in better care for cancer patients.
- We generalize our solution for tumor motion analysis to other application domains. The generalized framework can be applied to any motion with structured time series data, which can be described by a finite set of linear states.

This paper is organized as follows: Section 2 reviews related work. Section 3 will briefly introduce a finite state model for tumor motion and our hierarchical data structure. Our online subsequence similarity matching algorithm is discussed in section 4. In this section we describe query subsequence generation, a new subsequence similarity measure, and a new online motion prediction algorithm. An offline data analysis approach is discussed in Section 5. In this section we define stream similarity and patient similarity based on subsequence similarity. In Section 6, we generalize the similarity matching method to other domains. Section 7 presents performance results for each application. The last section concludes the paper and discusses directions for future work.

2. RELATED WORK

In this section, we discuss relevant work on similarity matching in the database community and work on respiratory motion analysis in the medical community. Similarity searching in time series data is used in many data mining applications. Agrawal et al. [1] introduced whole sequence similarity matching. Faloutsos et al. [7] performed subsequence similarity matching using a Discrete

Fourier Transformation (DFT). Moon et al. used generalized windows to reduce false negatives [19]. Other feature extraction functions, such as the Discrete Wavelet Transformation (DWT) [4, 11], Adaptive Piecewise Constant Approximation (APAC) [14], Piecewise Aggregate Approximation (PAA), and Single Value Decomposition (SVD) [17] have been proposed to reduce the dimensionality of time series data. *Dimensionality* of time series refers to sequence length. New distance functions such as Dynamic Time Warping [22, 27] and Longest Common Subsequences [5] have been explored to overcome the brittleness of the Euclidean distance measure or its variations [1, 7, 23].

There is extensive research activity in the database community on data streams. Some recent papers include [2, 6, 8, 9, 10, 13, 18, 21, 28]. But most research on streams focuses on basic statistics and on how to define and evaluate continuous queries, which is different from the focus of our work. Wu *et al.* [25] combines subsequence similarity matching with data streams, but on financial data, which has very different characteristics from tumor motion.

In the medical community, tumor respiratory motion has recently been studied in image-guided radiotherapy. An integrated radiotherapy imaging system has been designed for patient setup and for tumor motion localization [3]. A waveform model using a concept called the *average tumor trajectory* can synchronize the moving radiation beam [20]. A finite state automaton and a piecewise linear representation (PLR) of tumor motion has been proposed to capture the natural breathing actions [26]. Commonly used predictive methods to compensate respiratory motion have been evaluated [24]. The best parameterization for respiratory motion analysis and prediction remains an open question.

In this paper, we present our approach to online subsequence matching for tumor motion data. We process the data stream online to produce a piecewise linear Representation (PLR) of raw streams based on a finite state model. The length of a query subsequence is dynamically adjusted based on a new concept, *subsequence stability*, rather than using a fixed length. Our subsequence similarity measure has addressed the specific requirements for respiratory motion analysis by a model-based, multi-layer, weighted, and parametric distance function rather than using other partially weighted distance functions [16, 25]. In addition, we have developed new definitions for whole stream and patient similarity based on subsequence similarity, which is a departure from previous schemes that used whole sequence similarity measures.

3. MOTION MODEL AND DATA MODEL

In this section we will first briefly introduce the tumor respiration motion model used in our work. Then we will describe in detail our new data model that can be applied to a stream database, and which is suitable for both online and offline subsequence similarity matching for tumor motion analysis.

3.1 Motion Model

Due to the sheer volume and the noisy signal nature of raw data, it is impractical to determine subsequence similarity on raw data. A good data representation is needed to reduce the dimensionality (sequence length) of the raw data and to smooth noisy signals. We have adopted a finite state model [26] as the base for our data model and subsequence similarity matching. The rationale for adopting this model will be discussed after a brief introduction of the model.

The finite state model is illustrated in Figure 4. As a patient inhales or exhales, a tumor moves in a periodic pattern. These motion patterns are modeled using three regular breathing states: exhale (EX), end-of-exhale (EOE) and inhale (IN), and one irregular breathing state (IRR). The transition from one state to another is guided by the *finite state automaton (FSA)*, as illustrated in Figure 4b. In a regular breathing, motion proceeds from state to state in a fixed order:

$$\ldots \Rightarrow EX \Rightarrow EOE \Rightarrow IN \Rightarrow EX \Rightarrow \ldots$$

The IRR state is entered during irregular breathing, and is left when regular breathing resumes. Examples of identifying these states are illustrated in Figure 4c and d.

We adopt this finite state model because the model has several advantages in addressing the special concerns for subsequence similarity matching over respiratory motion. First, the model is based on the natural understanding of breathing motion and the requirements of motion compensated treatment. Each of these states corresponds to a natural action: EX is the motion due to lung deflation, EOE is the motion for rest after lung deflation, and IN is the motion due to lung expansion. Therefore subsequence matching based on this representation matches the problem domain.

Second, this model produces a piecewise linear representation (PLR) of the raw data. The PLR reduces the size of the raw data, lowers the dimensionality of a subsequence, and filters out noise.

Third, this model also provides an online algorithm [26] to generate the PLR segments in a streaming way, which serves a good starting point for real-time tumor motion prediction. The online algorithm can detect the current state and line segment in real time.

Finally, PLR sequences allow one to define different application-specific distance measures between two time series. Also we can place different relative importance on subsequence sources, amplitudes, frequency changes, and proximity in time, as required by respiratory motion analysis.

3.2 Data Model

A patient has multiple treatment sessions, and each session can yield multiple motion streams. Thus a patient has multiple streams. We propose a hierarchical data model for our stream database based on the PLR segments of the finite state model. This data model includes the corresponding stream relationship of the clinical procedure and satisfies the special requirements for subsequence matching.

The database is composed of a set of *patient records*. Each patient record has a set of data streams. Each stream has an ordered list of connected line segments, which is represented by an ordered list of vertices (a *vertex* is the intersection of two adjacent line segments).

Each vertex v_i is represented by three elements:

$$(t_i,\ x_i,\ s_i)$$

The vertex time t_i both denotes the start time of a line segment (beginning with the vertex) and the end time of the previous line segment (terminated with the vertex). The space position of v_i is denoted x_i. Since measurements of tumor motion have different spatial dimensionalities, we have proposed an approach that can work for any n-dimensional space. Thus x_i can be an n dimensional point. Last, s_i is the breathing state (i.e., EX, EOE, IN or IRR) of the line segment beginning with vertex v_i. This motion

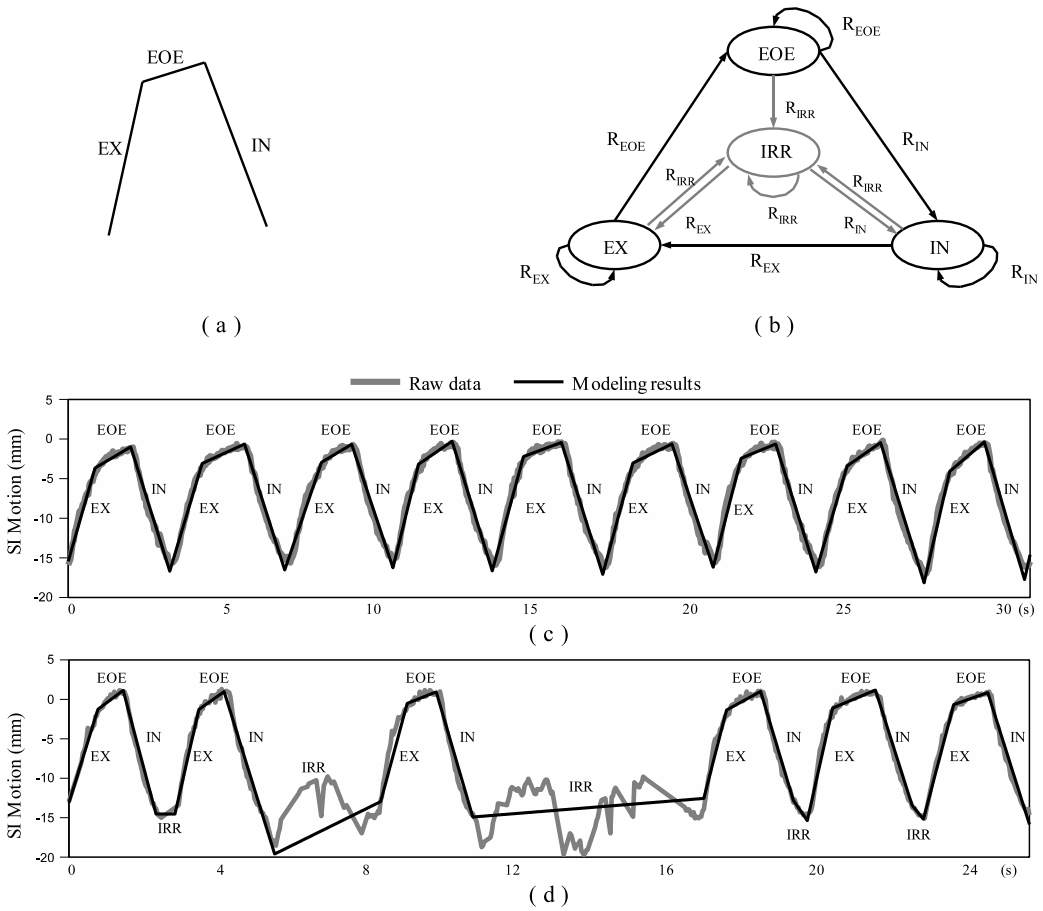

Figure 4: Tumor motion modeling (the motion is shown in one dimension, but the model can be used for multi-dimensional motion). (a) Three states of a regular breathing cycle, (b) Finite state automaton for respiratory motion, (c) Regular motion and the corresponding PLR segments, (d) Irregular motion and the corresponding PLR segments.

dimensionality refers to spatial motion, i.e., 1-D, 2-D and 3-D respiratory motion. Only one spatial dimension is shown in Figure 4. (*Spatial dimensionality* of tumor motion is orthogonal to and should not be compared with sequence dimensionality.)

With this data model, a breathing state corresponds to a single line segment. The state is stored in its beginning vertex. This data model is effective in capturing the features we need for tumor motion subsequence matching.

4. ONLINE SUBSEQUENCE MATCHING

In this section, we will discuss the issues for online subsequence similarity matching, including dynamic generation of a query subsequence, special concerns for subsequence similarity measures, and medical applications which can benefit from our method.

4.1 Dynamic query subsequence generation

For real-time applications, query subsequences must be generated in an online fashion and must provide an accurate representation of the target's current moving condition. Thus, a *query subsequence* must be the most recent part of a motion stream.

But how do we determine the appropriate length of the query subsequence? Shorter query subsequences lower the quality of the representation of the current motion characteristics. Longer query subsequences require additional computation and introduce longer delay. There is a tradeoff between response time and the length of a query subsequence. To address this problem, we propose a flexible scheme to adjust the query subsequence length dynamically based on *subsequence stability*, which is defined as:

DEFINITION 1. **(Subsequence Stability)** *Given a subsequence* $S = [(t_1, x_1, s_1), ..., (t_n, x_n, s_n)]$, *S is stable if* $\rho(S) < \rho_c$, *where* ρ_c *is a predefined parameter and* $\rho(S)$ *is computed by the following formula:*

$$\rho(S) = \frac{1}{n}\sum_{k=0}^{3}\sum_{i=1, s_i=k}^{n}(\alpha \cdot |\,|x_{i+1} - x_i| - \Delta\overline{x}_k\,| + \beta \cdot |\,(t_{i+1} - t_i - \Delta\overline{t}_k)\,|)$$

where k=0, 1, 2, 3 for each state EX, EOR, IN or IRR, and the inner sum is computed over line segments from vertex v_i *to* v_{i+1}, *where* s_i *= k as indicated.* $\Delta\overline{t}_k$ *is the average time interval in S for state k, and* $\Delta\overline{x}_k$ *is the average amplitude of S for state k.* α *and* β *are different weights for amplitude and frequency changes.* $\rho(S)$

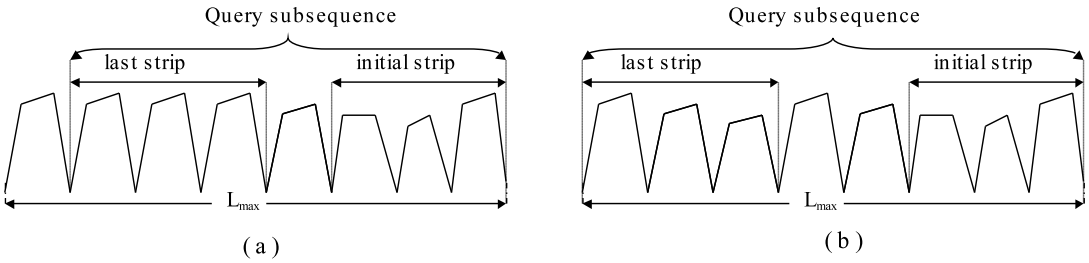

Figure 5: Dynamical adjusted query subsequence. (a) query length $< L_{max}$, (b) query length $= L_{max}$.

is called the stability of S. The smaller $\rho(S)$ is, the more stability S has.

The more stable the most recent subsequence is, the shorter the query subsequence will be. The length of the query subsequence is between the user specified minimum length L_{min} and the maximum length L_{max}. For example, in Figure 5a, L_{min} = 3 and L_{max} = 8 breathing cycles. We use a *stability checking strip* to determine the length of the query subsequence. A stability checking strip is a window of fixed size L_{min}, moving from the most recent portion back to historical data. The subsequence stability of the strip is checked after each move to determine the start point of a query subsequence. As illustrated in Figure 5, in the beginning, the strip covers the subsequence of the most recent L_{min} vertices. The stability of that subsequence is checked. If the subsequence is stable, the strip halts. If not, the strip will move one vertex back to history data, and stability on the new strip is checked again. This will go on until one of the two conditions is met: a stable subsequence is found (Figure 5a), or there are L_{max} vertices for the query subsequence (Figure 5b). The query subsequence is from the beginning vertex of the last strip to the most recent vertex. As a consequence, breathing with high regularity will have shorter query sequences, while breathing with low regularity tends to have longer query subsequences.

4.2 Online subsequence similarity

Subsequence matching over respiratory motion is very challenging because there are some special concerns for motion similarity comparison. The first special concern is the consideration of the meaning of a subsequence. Similar subsequences must have the same meaning. For tumor respiratory motion, the states of a subsequence defines its meaning. A pair of similarity subsequences must correspond to the same natural actions, such as exhaling and inhaling. Intuitively, a sequence that starts with an inhale cannot be compared with one that starts with an exhale.

The second special concern is the consideration of the source of the comparison subsequences. A patient has multiple treatment sessions. Each session produces a data stream. Subsequences from the same session are the most important. Subsequences from different sessions of the same patient are less important than those from the same session, though they are more important than those from a different patient.

Third, online subsequence similarity should have weighted importance over past data according to the proximity in time. In other words, recent data are given relatively more weight in defining similarity than the older data.

Finally, two similar subsequences should have limited flexibility for amplitude rescaling and frequency elasticity. Breathing motion is complicated. Two patients rarely breathe with the same amplitude and same frequency. Even for the same patient, there are variations from one breathing cycle to another. Both amplitude and frequency differences should be limited by certain thresholds.

There have been many research efforts for efficient similarity search based on Euclidean distance or its variations [4, 7, 15, 19, 23], Dynamic Time Warping distance [22, 27], or Longest Common Subsequence distance [5]. However, to the best of the authors' knowledge, existing subsequence similarity measures either are too general to address the above special requirements, or only meet part of them. For example, Euclidean distance is sensitive to conditions such as offset translation and amplitude scaling. The weighted distance used in [16] assigns a relative importance to each individual linear segment. But they did not consider frequency changes. The weighted distance used in [25] considered different weights over amplitude and frequency changes, but it does not assign different weights to different line segments based on proximity to an event. Furthermore, none of these studies have used weights to address the significance for the source streams of the corresponding subsequences and none have discerned the meaning of a sequence.

So a new similarity measure is required to address all these special concerns. We hereby propose such a model-based, multi-layer, weighted, and parametric subsequence similarity measure based on the PLR stream, which addresses the special concerns of respiratory motion stream similarity comparison, yet is still applicable for subsequence similarity in other domains.

DEFINITION 2. **(Online Subsequence Similarity)** *Two subsequences with the same length n:*

$$S_1 = [(t_{11}, x_{11}, s_{11}), ..., (t_{1n}, x_{1n}, s_{1n})]$$

$$S_2 = [(t_{21}, x_{21}, s_{21}), ..., (t_{2n}, x_{2n}, s_{2n})]$$

are similar if they satisfy the following conditions:

1. *The states of S_1 and S_2 are the same, i.e., $s_{1i} = s_{2i}$, for $i = 1, 2, ..., n$.*
2. *$d_s(S_1, S_2) < \delta$, where $d_s(S_1, S_2)$ is the weighted online subsequence distance function defined as:*

$$d_s(S_1, S_2) = \frac{1}{(n-1)w_s}\left[\alpha \sum_{i=1}^{n-1}(w_i \cdot \Delta x_i) + \beta \sum_{i=1}^{n-1}(w_i \cdot \Delta t_i)\right]$$

where

$$\Delta x_i = \mid \; |x_{1(i+1)} - x_{1i}| - |x_{2(i+1)} - x_{2i}| \; \mid$$

$$\Delta t_i = \mid \; (t_{1(i+1)} - t_{1i}) - (t_{2(i+1)} - t_{2i}) \; \mid$$

and α and β are the different weights for the amplitude and frequency, w_i is the weight for vertex i, and w_s is the weight between the two subsequences.

This similarity definition has addressed the special concerns in respiratory motion matching. Condition 1 ensures that the corresponding subsequences have the same natural physiological actions. That is one big advantage of using the piecewise linear representation based on the finite state model.

In condition 2, w_s gives different weights for subsequences of different patients. For example, similar subsequences from the same session of the same patient are the most valuable, and thus has the largest w_s. And w_s is the smallest when the two compared subsequences come from different patients.

Different values of α and β provide a tradeoff in the relative importance of amplitude and frequency. In our applications, we always have $\alpha > \beta$ to ensure that the amplitude has more significance than the frequency.

The weight w_i assigns different levels of importance to different portions of the compared subsequences. The more recent fragments more closely represent the current pattern, and have more influence on future move prediction with a larger w_i. In our research, w_i is a function of the query subsequence. For a given query subsequence:

$$Q \; = \; [(t_1, x_1, s_1), ..., (t_n, x_n, s_n)]$$

The weight for vertex v_i is computed as follows:

$$w_i = \frac{t_{n-1} - t_1}{\lambda(t_{n-1} - t_1) + (t_{n-1} - t_i)}, \;\; for \; i = 1, ..., n-1$$

so w_i is between $\frac{1}{\lambda+1}$ ($\lambda > 0$) and 1. The nearer the vertex is to the end of the subsequence, the higher weight it has.

So our distance function is sensitive to similar subsequences from the same patient, flexible with respect to amplitude and frequency changes, insensitive to offset translation, and suitable for online application with different weights for different parts of the subsequences. One salient feature of our similarity measure is its parametric. It can be applied in other application domains by adjusting the parameters of w_s, w_i, α and β. For instance, the weighted distance functions in [16, 25] are special cases of our distance measure.

Currently, the settings if different parameters are based on prior experience and experimental results. We are working on new strategies for automatic parameter tuning, so that they can learn the proper settings from training data and dynamically adjust their values during online procedures.

4.3 Motion prediction

The immediate future of a *historical* subsequence is known. By matching a *current* query subsequence with a *similar historical* subsequence, one can predict that the future of the query subsequence will be similar to that of the historical subsequence. Future frequency, amplitude or position can be predicted. This section describes position prediction. Prediction of the other future characteristics is analogous.

A typical prediction task is to locate the tumor after Δt from the current time. Statistical analysis over the retrieved subsequences can predict future tumor position through the following steps. To simplify the process, we assume that the current time is the time of the last vertex of the query subsequence.

For each retrieved similar subsequence S_i, its position after Δt from the last vertex is retrieved, denoted as $X_{s_i}(\Delta t)$. Then, the future position of the query subsequence after Δt is predicted using the following formula:

$$x_q(\Delta t) = X_{q1} + \frac{1}{\sum_{i=1}^{m} w_{s_i}} \sum_{i=1}^{m} (X_{s_i}(\Delta t) - X_{s_i 1}) \cdot w_{s_i}$$

where m is the total number of retrieved similar subsequences, X_{q1} is the first vertex position of the query subsequence, $X_{S_i 1}$ is the first vertex position of S_i, and w_{s_i} is the subsequence weight of S_i.

5. OFFLINE CLUSTERING

Subsequence matching in the previous section is introduced for online applications, but it can be used for offline data analysis too. In this section, we define whole stream similarity and patient similarity based on subsequence similarity. Stream and patient similarity are important for tumor motion characterization, computer aided diagnosis, and discovery of patient information correlations.

The objectives for online and offline analysis are distinguished from each other, which leads to different solutions. For online analysis, the focus is on how a subsequence represents the current motion status and how to predict future motion, so we propose a distance function with different weights for vertices. For offline analysis, the goal is to discover correlations between moving patterns and patient information. Since there is no 'current time', in offline analysis, all vertices from the same subsequence have the same weights. But the weights over amplitude (α) and frequency (β) are still necessary, so is the weight for a source stream (w_s). Thus, the offline subsequence distance can be obtained from Definition 2 by setting all w_i to 1. We denote the offline subsequence distance as $d'_s(S_1, S_2)$.

$$d'_s(S_1, S_2) = \frac{1}{(n-1) \cdot w_s} \left(\alpha \sum_{i=1}^{n-1} \Delta x_i + \beta \sum_{i=1}^{n-1} \Delta t_i \right)$$

$d'_s(S_1, S_2)$ is used below to define the distance functions between a pair of streams or patients, and clustering over stream and patients.

5.1 Stream similarity

The whole stream similarity between two PLR streams R_1 and R_2 is defined based on the offline subsequence similarity between two sets of subsequences. One set is from R_1 and another set is from R_2. For a given stream R with N vertices, there are $N - n + 1$ possible different subsequences with subsequence length n. According to our tumor motion model, there are only three regular states and one irregular state. If a patient breathes regularly, the occurrence of the irregular state is rare. Suppose there are N_1 and N_2 subsequences with length n from stream R_1 and R_2. So for each subsequence of R_1, there are about $\frac{N_2}{3}$ subsequences from R_2 with the same order of subsequences, and vice versa.

Only a small set of the most similar subsequences is used for stream similarity. So for each query subsequence from R_1, the most similar $\gamma \cdot N_2$ retrieved subsequences from R_2 will be used to define

the distance between R_1 and R_2, where γ is a user specified parameter. For example, γ can be 10%. If a query cannot find at least $\gamma \cdot N_2$ subsequences with the same state order from R_2, that query subsequence is an outlier and will be removed from the query subsequences of R_1.

In the end, there are m_1 ($m_1 \leq N_1 - n - 1$) query subsequences from R_1, which will retrieve $\gamma \cdot N_2$ similar subsequences from R_2. Similarly, there are m_2 ($m_2 \leq N_2 - n - 1$) query subsequences from R_2, which will retrieve $\gamma \cdot N_1$ similar subsequences from R_1. Based on these subsequences and their subsequence distance, the stream distance of R_1 and R_2 is defined as:

DEFINITION 3. ***(Stream Distance)*** *Given two raw streams R_1 and R_2, the stream distance function is:*

$$d_r(R_1, R_2) = \frac{1}{m_1 \cdot \gamma \cdot N_2} \sum_{i=1}^{m_1} \sum_{k=1}^{N_2 \cdot \gamma} d'_s(S_{1i}, S_{2k}) + \frac{1}{m_2 \cdot \gamma \cdot N_1} \sum_{i=1}^{m_2} \sum_{k=1}^{N_1 \cdot \gamma} d'_s(S_{2i}, S_{1k})$$

where S_{1i} is one query subsequence from R_1 and S_{1k} is a retrieved subsequence from R_1. Similar is for S_{2i} and S_{2k}. d'_s is the offline subsequence distance.

The stream distance function is a good indicator of whole stream similarity. The smaller the distance, the more similar the two streams are. The distance between two stream is symmetric, $d_r(R_1, R_2) = d_r(R_2, R_1)$, which makes it convenient to define patient similarity in the next section.

5.2 Patient similarity

Due to the hierarchical structure of patient information, a patient has several streams. Patient similarity is defined based on the stream similarity. Stream similarity is based on subsequence similarity, thus patient similarity is also based on subsequence similarity. The distance between two patients is the average distance between two streams, one from the first patient and the other from the second patient.

DEFINITION 4. **(Patient Distance)** *Given two patients P_a and P_b, the patient distance function is:*

$$d_p(P_a, P_b) = \frac{1}{N_a \cdot N_b} \sum_{m=1}^{N_a} \sum_{n=1}^{N_b} d_r(R_{am}, R_{bn})$$

where N_a and N_b are stream numbers of patient P_a and P_b. R_{am} is the m^{th} session stream of P_a and R_{bn} is the n^{th} session stream of P_b. $d_r(R_{am}, R_{bn})$ is the stream distance between R_{am} and R_{bn}.

Patient similarity provides a convenient way to clustering patients based on their similarity distances. After clustering of patients, one may then identify patient features (e.g., age, tumor position, historical treatments), which are correlated with tumor movement. Next we will discuss some sample applications by the means of patient clustering.

5.3 Clustering applications

The stream and patient similarity defined in the previous section is important for characterization of tumor motion. And they provide a convenient way to correlate a tumor moving pattern with patient physiological information. We present three applications based on stream and patient similarity. These are traditional clustering problems.

First, it can be used for correlation discovery between tumor motion and a tumor's geometric location. It is known that a tumor's moving pattern is affected greatly by the tumor's location in an organ. But there is no consensus about the effect yet. We are working on a solution to partition the organ based on motion similarity. The basic idea is clustering patients based on patient similarity. Then the correlation can be discovered based on the clustered information.

Second, stream similarity and patient similarity can be used to discover physiological correlations with tumor motion. For instance, stream similarity among different treatment sessions of the same patient can be used to correlate a patient's physiological changes with moving pattern changes. Patient similarity is used to decide if there is a correlation between genetic diseases with certain tumor moving patterns. Conversely, knowing the correlations between tumor moving patterns and a patient's physiological conditions will help in predicting tumor motion with fewer diagnostic samplings.

Third, patient distance can also be used for prediction with clustering. After clustering, if a patient has a new treatment session, then the subsequence similarity matching will only retrieve subsequences from the same cluster, not from patients in other clusters.

6. GENERALIZATION OF THE METHOD

Although we have been focusing on tumor respiratory motion analysis, our method is generally suitable for any motion with structured time series data, which can be described by a finite set of linear states. The general picture of the framework using our method can be sketched in the following four steps:

(1) Motion Modeling: Build a finite state model to simulate the motion with line segments. Each line segment corresponds to one of a finite set of states.

(2) Segmentation: Develop a segmentation algorithm to produce the piecewise linear representation of the raw data. The line segment and the state of the line segment must be decided in an online fashion for real-time applications.

(3) Subsequence Similarity: Define a subsequence similarity measure for subsequence matching. The similarity measure can be either general or application specific.

(4) Result analysis: Propose some application specific statistical methods to analyze the retrieved similar subsequences.

The four steps are independent from each other. An existing solution for one application can be adopted by another application, with modification of one or more steps.

In addition to respiratory motion, there are many other applications which can be simulated and analyzed using the above framework. A few examples are briefly listed here.

Table 1: Settings of Parameters.

Parameters	*Symbols*	*Values*
Weight for amplitude	α	1.0
Weight for frequency	β	0.25
Weight for vertexes	λ	0.25
Weight for source streams	w_s	1.0, if from the same session 0.9, if from different session of the same patient 0.3, if from different patients
Subsequence distance threshold	δ	8.0
Stability threshold	ρ_c	6.0

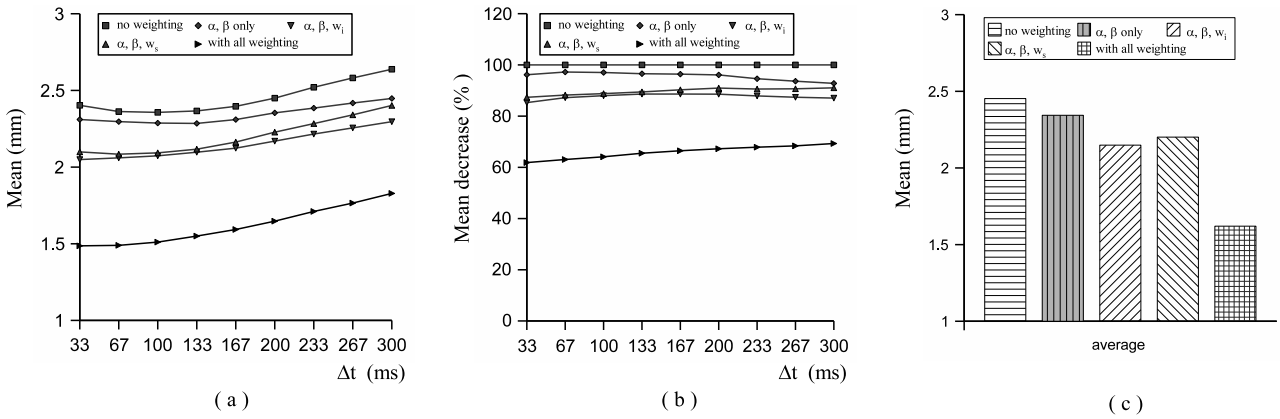

Figure 6: Prediction results using different weighted factors for subsequence similarity. (a) Mean prediction error for different time intervals, (b) Error reduction by different weighting factors, (c) Averaged prediction results over all the time intervals.

- A very similar application is patient heartbeat analysis and characterization. The regularity of a heartbeat may be affected by fever, blood pressure, medication, or other physiological conditions. Heart beat analysis is a major part in patient diagnosis and treatment procedure.

- Quite a few mechanical instruments move in a predictive patterns with linear motion of a finite set of states. The variations of the mechanical movement may be affected by different factors, such as temperature, wind, or electronic power. Predicting future motion or detecting abnormal moving patterns is of great importance for some applications.

- In an assembly line, the motion of a robot arm may be limited to a finite set of predefined states. We can pursue dynamic robot control and automatic robot manipulation through motion prediction and corresponding response actions.

- The tide's rhythmic rise and fall is in a predictive pattern, mostly following the moon's motion and position. The tidal motion is affected by weather conditions, such as the force of wind. By learning more about tidal motion, we can discover how the phases of the moon or the moon's distance from Earth affects the tidal range. We can also correlate tides with coastal catastrophes.

7. PERFORMANCE

We have evaluated our subsequence matching approach and its applications by a series of experiments. We test our similarity measures using online prediction results as an accuracy indicator. We have experiments to evaluated different weighting factors, as well as to compare our new weighted L_1 distance functions to the corresponding weighted Euclidean distance. The second set of experiments evaluates our online query subsequence generation mechanism by comparing with fixed length query subsequences. Another set of experiments shows the results for offline stream and patient similarity, and how the results of offline analysis can help for online tumor motion prediction.

7.1 Experimental setup

Real patient data is used in our experiments. More than 2,000,000 raw data points from 42 patients with about 1200 treatment sessions, were used in experiments. Data is imaged at 30Hz. All our experiments were conducted on a DELL OPTIPLEX GX 260 with Pentium(R) 4 processor, 2.66GHz CPU, 1GB RAM.

We have done a series of experiments to set up the parameter values for subsequence matching. To determine the values for one parameter (for instance β), we first fixed all the other parameters (such as α, λ, w_i, δ and ρ_c). Then we run experiments with different β values. Finally, β is fixed to the value with the best prediction results. Later, the fixed β is used to determine the values of other parameters. In the following experiments, we use the settings shown in Table 1, unless otherwise specified.

7.2 Subsequence similarity measures

The subsequence similarity defined in Definition 2 uses a weighted L_1 distance function. The similarity measure is evaluated by prediction quality of future tumor positions. The mean difference be-

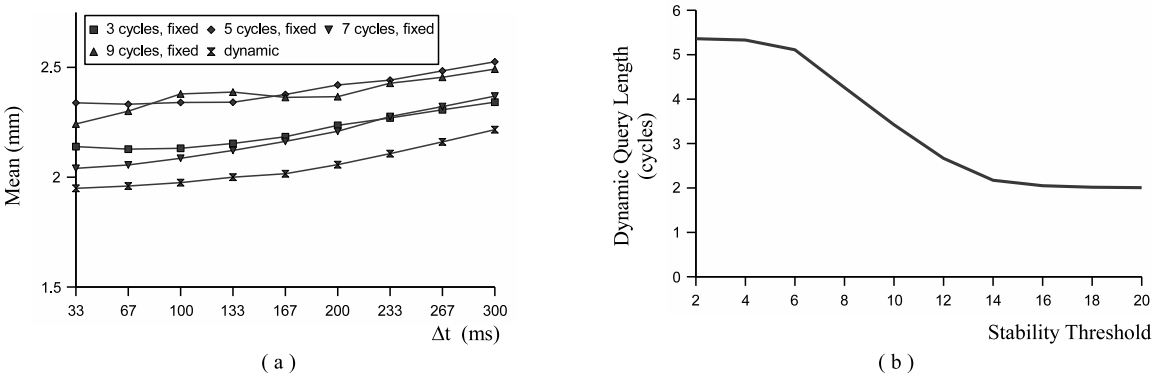

Figure 7: Dynamic and fixed query subsequences. (a) Prediction results for fixed and dynamic lengths, (b) Relationship between dynamic length and subsequence stability threshold.

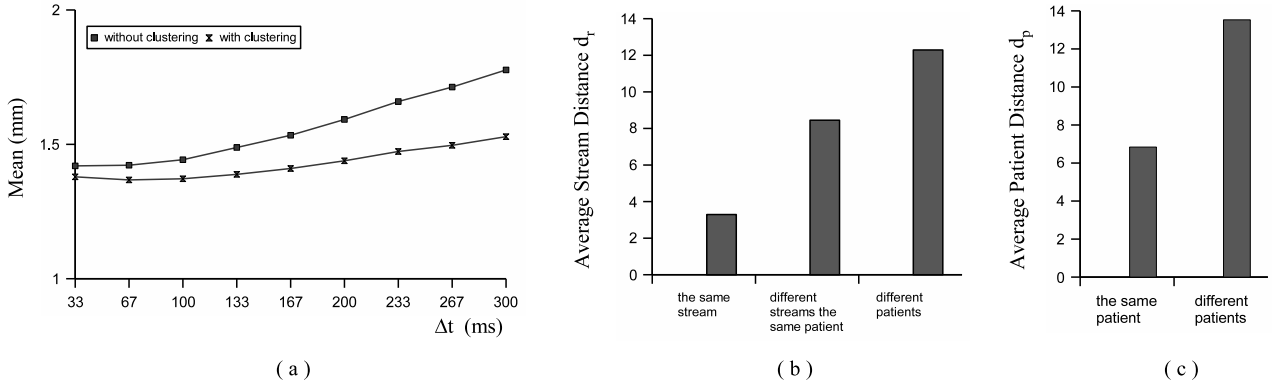

Figure 8: Clustering and Stream and patient similarity. (a) Prediction with or without clustering, (b) Stream distances, (c) Patient distances.

tween the predicted positions and PLR values is used to measure the quality of prediction. Smaller mean difference means better prediction results.

One thing to be noted here is that tumor motion is limited in space, averaged about 15mm for superior-inferior motion. Even a slightly better prediction, on the order of 0.5mm more accurate, is of great significance, since it will reduce patient radiation toxicity to healthy tissues or structures.

Figure 6 shows the prediction results using different weighting factors (α, β, w_s and w_i) for subsequence distance functions. Figure 6a is the prediction results after different time periods. Δt is between 0 and 300 milliseconds. 300ms is a reasonable upper bound approximation for system delay of average imaging rate. Figure 6b illustrates how much prediction error has been reduced by the weighting factors, comparing to the one with no weighting factors at all. Figure 6c is the average prediction results for all the time intervals.

The results show that prediction without any weighted information (*no weighting*) gives the worst prediction results. Prediction using the distance function with different weights for amplitude and frequency, but with neither weighted stream information nor weighted line segments ($\alpha, \beta\ only$) gives a slightly better prediction results. Prediction with one additional weighting factor, either weighted streams ($\alpha, \beta,\ w_s$) or weighted line segments ($\alpha, \beta,\ w_i$) is also slightly better. Prediction with all the weighting factors (*with all weighting*) yields the best results.

The threshold for the distance function (δ) also affects prediction results. With a smaller threshold, the prediction results are better, which is illustrated in Figure 9. The drawback is that there will be fewer similar subsequences with a smaller δ. We predict only if there are a certain number of retrieved subsequences. A smaller δ will result in fewer predictions. There is a tradeoff between the number of predictions and the prediction accuracy.

There are several existing distance functions, such as Euclidean distance, longest common subsequence (LCS) [5], and dynamic time warping (DTW) [22, 27]. LCS is proposed for string matching. It is not applicable for tumor motion analysis because tumor position is continuous. We did not run experiments to compare our weighted L_1 distance function with DTW because, first, DTW has no weighted information, and second, the running time of DTW is very computationally expensive, which makes it not suitable for real-time prediction. Third, DTW does not create any meaningful description of the data.

7.3 Query subsequence generation

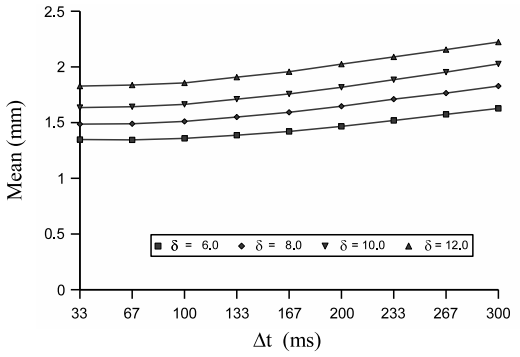

Figure 9: Effects of distance threshold δ.

Experiments are carried out to test the query subsequence generation mechanism, which dynamically adjusts the length of a query subsequence based on subsequence stability. The results are summarized in Figure 7.

The performance results show that the dynamic method has overall better performance than fixed length query subsequences. Figure 7a compares the prediction quality for query subsequences with different fixed lengths along with the dynamic method. The dynamic method has better performance.

In the dynamic method, the length of a query subsequence is determined by the subsequence stability threshold. The relationship between the average dynamic query subsequences and stability thresholds (ρ_c) is presented in Figure 7b, with L_{min} = 2 and L_{max} = 9 breathing cycles (defined in Section 4.1). The lengths of query subsequences increase with a smaller stability threshold. Subsequences have a length ranging from 3 to 5 breathing cycles.

7.4 Clustering

Clustering of streams and patients based on subsequence similarity is an effective tool for tumor motion analysis, prediction and correlation discovery. One application based on stream and patient similarity is illustrated in Figure 8a. It displays prediction results with or without clustering based on patient distances. With clustering, subsequence similarity matching will only search for similar subsequences from patients in the same cluster. Without clustering, similar subsequences from all patients will be retrieved and used in prediction. The results show that prediction with clustering gives better accuracy.

Next we want to show the stream and patient similarity based on subsequence similarity. Intuitively, a stream should be the most similar to itself, less similar to other streams from the same patient, and the least similar to streams from other patients. Similarly, patients' data is more similar to other data from him/herself, than to data from other patients. These ideas are supported by our performance results shown in Figure 8b and c.

7.5 Efficiency

We have tested the efficiency for subsequence similarity matching. Since online prediction is real-time application, response time is very important. All the data can fit in memory, no disk I/O is needed. Our online segmentation runs with constant space and in linear time with respect to raw data points. So for each new incoming data point, the segmentation runs in constant time. Each subsequence similarity matching runs in linear time with respect to segmented line segments. The average time of one prediction is less than 30 millisecond in our data sets, including the time for segmentation and similarity matching. So the computation time is short enough for image guided dynamic targeting radiation treatment.

8. SUMMARY AND FUTURE WORK

In this paper, we introduced a solution for tumor respiratory motion analysis, clustering and online prediction by subsequence similarity matching over tumor motion data. To address the special requirements for tumor motion analysis, we defined a new subsequence similarity measure based on a finite state model and piecewise linear representation (PLR) of raw data. The new weighted L_1 distance function has different weights on amplitude, frequency, source stream, and proximity to current time. To guard against query sequences which are abnormal (not representative), we proposed a new concept: subsequence stability. A flexible scheme was adopted to dynamically adjust the length of a query subsequence based on subsequence stability. Experimental results proved that the dynamically generated query subsequences have overall better performance than queries with fixed lengths. And the weighted L_1 distance function outperforms the corresponding weighted Euclidean distance function.

Based on subsequence similarity, whole stream and patient similarity are defined. Stream and patient similarity are important for motion characterization, and for the discovery of correlations between motion patterns and other patient information. Patient similarity also provides a convenient way to cluster patients, which can be used for both online and offline applications.

Our approach has taken into account the internal structure of a time series, the underlying meaning of the structure, and the influence on subsequence matching. The solution can be generalized into a framework, which is applicable to a wide class of problems involving motion analysis and prediction with structured time series data.

Future research can proceed in several directions. Our immediate plans include applying our approach for image guided dynamic radiation treatment. One ongoing project is automatic dynamic parameter tuning, in which the system will learn the proper parameter settings from training data and adapt them during online operation. Another research area is to improve noise detection strategies and to find better cardiac motion modeling to obtain more precise motion prediction. Still another problem is incorporating indexing in the search algorithm to more rapidly retrieve similar subsequences.

9. ACKNOWLEDGEMENT

This work was supported in part by NSF grant IIS-0073063, the Whitaker Foundation Grant RG-01-0175, and CenSSIS, the Center of Subsurface Sensing and Imaging Systems, under the Engineering Research Centers Program of the NSF (Award Number EEC-9986821).

10. REFERENCES

[1] R. Agrawal, C. Faloutsos, and R. Swami, A. Efficient Similarity Search in Sequence Database. *FODO*, pages 69–84, 1993.

[2] S. Babu and J. Widom. Continuous Queries Over Data Stream. *SIGMOD Record*, 30(3):109–120, 2001.

[3] R. Berbeco, S. Jiang, G. Sharp, G. Chen, H. Mostafavi, and H. Shirato. Integrated Radiotherapy imaging system (IRIS): Design Considerations of Tumour Tracking with Linac Gantry-mounted Diagnostic X-ray System with Flat-panel Detectors. *Phys. Med. Biol.*, 49(2):243–255, 2004.

[4] K.-P. Chan and A.-C. Fu. Efficient Time Series Matching by Wavelets. *ICDE*, pages 126–133, 1999.

[5] G. Das, D. Gunopulos, and H. Mannila. Finding Similar Time Series. *PKDD*, pages 88–100, 1997.

[6] M. Datar, A. Gionis, P. Indyk, and R. Motwani. Maintaining Stream Statistics over Sliding Windows. *SODA*, pages 635–644, 2002.

[7] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast Subsequence Matching in Time-Series Database. In *SIGMOD*, pages 419–429, 1994.

[8] L. Gao and X. S. Wang. Continually Evaluating Similarity-Based Pattern Queries on a Streaming Time Series. In *SIGMOD*, pages 370–381, 2002.

[9] L. Gao, Z. Yao, and X. S. Wang. Evaluating Continuous Nearest Neighbor Queries for Streaming Time Series via Pre-fetching. In *CIKM*, pages 485–492, 2002.

[10] J. Gehrke, F. Korn, and D. Srivastava. On Computing Correlated Aggregates over Continual Data Streams. *SIGMOD*, pages 126–133, 2001.

[11] A. C. Gilbert, Y. Kotidis, and S. Muthukrishnan. Surfing Wavelets on Streams: One-pass Summaries for Approximate Aggregate Queries. *VLDB*, pages 79–88, 2001.

[12] M. Goitein. "Organ and Tumor Motion: An Overview". *Semin. Radiat. Oncol.*, 14(1):2–9, Jan 2004.

[13] L. Golab and M. T. Ozsu. Issues in Data Stream Management. *SIGMOD Record*, 32(2):5–14, 2003.

[14] E. J. Keogh, K. Chakrabarti, S. Mehrotra, and M. J. Pazzani. Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. In *SIGMOD*, pages 151–162, 2001.

[15] E. J. Keogh, K. Chakrabarti, M. J. Pazzani, and S. Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, 3(3):263–286, 2001.

[16] E. J. Keogh and M. J. Pazzani. An Enhanced Representation of Time Series Which Allows Fast and Accurate Classification, Clustering and Relevance Feedback . In *KDD*, pages 239–243, 1998.

[17] F. Korn, H. V. Jagadish, and C. Faloutsos. Efficiently Supporting ad hoc Queries in Large Datasets of Time Sequences. In *SIGMOD*, pages 289–300, 1997.

[18] X. Liu and H. Ferhatosmanoglu. Efficient k-NN Search on Streaming Data Series. In *SSTD*, pages 83–101, 2003.

[19] Y.-S. Moon, K.-Y. Whang, and W.-S. Han. General Match: a Subsequence Matching Method in Time-series Databased Based on Generalized Windows. In *SIGMOD*, pages 382–393, 2002.

[20] T. Neicu, H. Shirato, Y. Seppenwoolde, and S. Jiang. Synchronized Moving Aperture Radiation Therapy (SMART): Average Tumor Trajectory for Lung Patients. *Phys. Med. Biol.*, 48(5):587–598, 2003.

[21] L. O'Callaghan, A. Meyerson, R. Motwani, N. Mishra, and S. Guha. Streaming-Data Algorithms for High-Quality Clustering. *ICDE*, pages 685–, 2002.

[22] S. Park, S.-W. Kim, and W. W. Chu. Segment-Based Approach for Subsequence Searches in Sequence Databases. *SAC*, pages 248–252, 2001.

[23] D. Rafiei and A. Mendelzon. Similariy-Based Queries for Time-series data. *SIGMOD*, pages 13–24, 1997.

[24] G. Sharp, S. Jiang, S. Shimizu, and H. Shirato. Prediction of Respiratory Tumour Motion for Real-time Image-guided Radiotherapy. *Phys. Med. Biol.*, 49(3):425–440, 2004.

[25] H. Wu, B. Salzberg, and D. Zhang. Online Event-driven Subsequence Matching over Financial Data Streams. *SIGMOD*, pages 23–34, 2004.

[26] H. Wu, G. C. Sharp, B. Salzberg, D. Kaeli, and S. Jiang. A Finite State Model for Respiratory Motion Analysis in Image Guided Radiation Therapy. *Phys. Med. Biol.*, 2004.

[27] B.-K. Yi, H. V. Jagadish, and C. Faloutsos. Efficient Retrieval of Similar Time Serquences under Time Warping. In *ICDE*, pages 201–208, 1998.

[28] Y. Zhu and D. Shasha. StatStream: Statistical Monitoring of Thousands of Data Streams in Real Time. *VLDB*, pages 358–369, 2002.