Prospective detection of large prediction errors: a hypothesis testing approach

2010 Phys. Med. Biol. 55 3885

(http://iopscience.iop.org/0031-9155/55/13/021)

IOP PUBLISHING PHYSICS IN MEDICINE AND BIOLOGY

Phys. Med. Biol. **55** (2010) 3885–3904 doi:10.1088/0031-9155/55/13/021

Prospective detection of large prediction errors: a hypothesis testing approach

Dan Ruan

Department of Radiation Oncology, Stanford University, Stanford, CA, USA

E-mail: druan@stanford.edu

Received 26 March 2010, in final form 11 May 2010
Published 22 June 2010
Online at stacks.iop.org/PMB/55/3885

Abstract

Real-time motion management is important in radiotherapy. In addition to effective monitoring schemes, prediction is required to compensate for system latency, so that treatment can be synchronized with tumor motion. However, it is difficult to predict tumor motion at all times, and it is critical to determine when large prediction errors may occur. Such information can be used to pause the treatment beam or adjust monitoring/prediction schemes. In this study, we propose a hypothesis testing approach for detecting instants corresponding to potentially large prediction errors in real time. We treat the future tumor location as a random variable, and obtain its empirical probability distribution with the kernel density estimation-based method. Under the null hypothesis, the model probability is assumed to be a concentrated Gaussian centered at the prediction output. Under the alternative hypothesis, the model distribution is assumed to be non-informative uniform, which reflects the situation that the future position cannot be inferred reliably. We derive the likelihood ratio test (LRT) for this hypothesis testing problem and show that with the method of moments for estimating the null hypothesis Gaussian parameters, the LRT reduces to a simple test on the empirical variance of the predictive random variable. This conforms to the intuition to expect a (potentially) large prediction error when the estimate is associated with high uncertainty, and to expect an accurate prediction when the uncertainty level is low. We tested the proposed method on patient-derived respiratory traces. The 'ground-truth' prediction error was evaluated by comparing the prediction values with retrospective observations, and the large prediction regions were subsequently delineated by thresholding the prediction errors. The receiver operating characteristic curve was used to describe the performance of the proposed hypothesis testing method. Clinical implication was represented by miss detection rate and delivery efficiency. Both characterizations demonstrated the promising

results and provided insight into the tradeoffs in the detection task. This study opens the discussion on real-time analysis of prediction accuracy and promises important information in automatically adjusting treatment and/or target monitoring schemes.

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

1. Introduction

The goal of radiotherapy is to deliver ablative radiation to tumors and to best spare surrounding normal structures. For highly mobile tumors, motion management is crucial and usually follows the streamline of (1) accurately monitoring the tumor target; (2) predicting the tumor location at the instant of delivery, to account for system latency; (3) managing the treatment to maintain the alignment between the external treatment beam and the internal tumor target, by either passively gating or adaptive tracking (Keall *et al* 2002, Nuyttens *et al* 2006). A powerful treatment management mechanism requires not only the prediction of the tumor location at beam delivery but also the uncertainty associated with such prediction: a prediction with low uncertainty may be used to move the beam to track the tumor, while a prediction with high uncertainty suggests a more cautious motion management scheme and may lead to a beam pause decision when necessary.

The location-prediction problem has received much attention and yielded a large body of prediction algorithms, including but not limited to regression (Vedam *et al* 2004, Ruan *et al* 2007, McCall and Jeraj 2007, Ernst *et al* 2007), Kalman filter and its variations (Putra *et al* 2008), neural network and fuzzy logic inference (Kakar *et al* 2005, Isaksson *et al* 2005, Murphy and Dieterich 2006). These studies rely on the assumption that a consistent and deterministic inference structure exists between the observable historical segment and the unknown tumor location to be predicted. The basic setting is deterministic: a 'true' prediction value is sought after. There notion of uncertainty needs to be substituted with its counterpart of 'prediction error' in this context. Even though it may be possible to 'learn' the behavior of the prediction error, the deterministic assumptions make such task quite difficult. In fact, there is not yet any study on the *prospective* characterization of prediction errors.

More recently, studies have emerged where the future tumor location is treated as a random variable rather than a deterministic quantity (Ruan 2010). The rationale for such statistical setting is to admit the intrinsic stochasticity of the tumor motion and to allow identical historical segments to evolve into different future states. An important outcome is that the prediction algorithm outputs a pdf for the future tumor location, and a numeric estimate is only obtained via a post-processing step, by taking the mean, median or mode estimate of the underlying random variable.

The intermediate result of the pdf provides rich distributional information of the random future tumor location. In particular, the first moment yields an efficient location predictor (Ruan 2010). The second moment, on the other hand, captures the variability of the potential future tumor location and describes the uncertainty associated with its estimate. We review the KDE-based prediction method and present the derivation of the corresponding uncertainty estimate in section 2. Section 3 describes the test data, the implementation procedure and reports the experimental results. Section 4 provides some structural discussions and summarizes this study.

2. Methods

2.1. Basic inference and hypothesis testing setup[1]

For simplicity, we focus our analysis on a 1D discrete-time trajectory s. At time t_k, the goal of prediction is to estimate the value of s_{k+L} given samples s_i for $i = 1, 2, \ldots, k$, where L is the prediction length corresponding to system latency. Let p be the number of augmented states for sufficient description of the system dynamics and Δ be a 'lag length' chosen to balance the effect of dynamics and observation noise (Ruan *et al* 2007); then we can redefine the covariate $\boldsymbol{x}_i = [s_{i-(p-1)\Delta}, s_{i-(p-2)\Delta}, \ldots, s_i]$ and the response $\boldsymbol{y}_i = s_{i+L}$. The prediction goal can be rephrased as estimating the test response $\boldsymbol{y}_k$ from the observed test covariate $\boldsymbol{x}_k$, given known covariate/response pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ for $i = 1, 2, \ldots, M = k - L$.

Obviously, the magnitude of prediction error depends on the choice of prediction algorithms, and we adopt a KDE-based method with demonstrated efficacy (Ruan 2010).

2.2. KDE-based estimation of test response distribution

We regard the observed covariate-response pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ as realizations of random vector/variables X and Y and define a random vector $Z = (X, Y) \in \Re^{p+1}$. With independent samples $\boldsymbol{z}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i)$, $i = 1, 2 \ldots, M$, the distribution of Z can be obtained via KDE (Duda *et al* 2001):

$$p(\boldsymbol{z}_k) = \frac{1}{M} \sum_{i=1}^{M} \ker(\boldsymbol{z}_k | \boldsymbol{z}_i). \tag{1}$$

where the local density kernel $\ker(\boldsymbol{z}_k|\boldsymbol{z}_i)$ is assumed to be a spatial invariant Gaussian pdf with the block diagonal covariance $\Sigma_z = \mathrm{diag}\{\Sigma_{\mathrm{ker},x}, \sigma^2_{\mathrm{ker},y}\}$, where $\Sigma_{\mathrm{ker},x}$ and $\sigma^2_{\mathrm{ker},y}$ are the covariance for the covariate and response variables, respectively. The conditional distribution of the test response $p(\boldsymbol{y}_k|\boldsymbol{x}_k)$ turns out to be a Gaussian mixture,

$$\begin{aligned} p(\boldsymbol{y}_k|\boldsymbol{x}_k) &= p(\boldsymbol{z}_k)/p(\boldsymbol{x}_k) \\ &= \frac{1}{M} \sum_i \ker((\boldsymbol{x}_k, \boldsymbol{y}_k)|\boldsymbol{z}_i) \\ &= \frac{1}{C} \sum_i w_i \exp\left[-\|\boldsymbol{y}_k - \boldsymbol{y}_i\|^2 / 2\sigma^2_{\mathrm{ker},y}\right], \end{aligned} \tag{2}$$

where the wight w_i determines the contribution of the sample response $\boldsymbol{y}_i$ and is given by the relative closeness of the sample covariate $\boldsymbol{x}_i$ to the test covariate $\boldsymbol{x}_k$:

$$w_i = (2\pi)^{-p/2} \det(\Sigma_{\mathrm{ker},x})^{-1/2} \exp[-(\boldsymbol{x}_k - \boldsymbol{x}_i)^T \Sigma^{-1}_{\mathrm{ker},x} (\boldsymbol{x}_k - \boldsymbol{x}_i)]. \tag{3}$$

The normalization parameter C is independent of both i and $\boldsymbol{y}_k$.

2.3. The first and second moments

Given the approximate distribution of the response random variable in (2), it is natural to use the first moment as the prediction output. We also derive its second moment that will be used in section 2.4.

[1] Material in section 2.1 on the KDE setup is partially adopted from Ruan (2010); please refer to the original text for technical details.

2.3.1. The first moment: prediction with the mean estimate. Given the conditional distribution (2), the mean estimate reads

$$\begin{aligned}\hat{\boldsymbol{y}} = E[Y_k|\boldsymbol{x}_k] &= \frac{1}{\tilde{C}}\int \sum_i \boldsymbol{y} w_i \exp\left[-\|\boldsymbol{y}-\boldsymbol{y}_i\|^2/2\sigma^2_{\text{ker},y}\right]\mathrm{d}\boldsymbol{y} \\ &= \frac{1}{\sum w_i}\sum_i w_i \boldsymbol{y}_i. \end{aligned} \tag{4}$$

It is the weighted sum of the training response values and can be computed efficiently. Furthermore, the normalization parameter $\tilde{C}$ is eliminated by cancellation.

2.3.2. The second moment: the variance estimate. The variance of a random variable measures its uncertainty/variability:

$$\begin{aligned}\hat{\sigma}^2_{\boldsymbol{y}} &= E[(Y_k - E(Y_k|\boldsymbol{x}_k))^2|\boldsymbol{x}_k] \\ &= \frac{1}{\tilde{C}}\int \sum_i (\boldsymbol{y}-\hat{\boldsymbol{y}}_k)^2 w_i \exp\left[-\|\boldsymbol{y}-\boldsymbol{y}_i\|^2/2\sigma^2_{\text{ker},y}\right]\mathrm{d}\boldsymbol{y} \\ &= \sigma^2_{\text{ker},y} + \sum_i \frac{w_i}{\sum_i w_i}(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_k)^2, \end{aligned} \tag{5}$$

where $\sigma^2_{\text{ker},y}$ is the y-direction variance in KDE, reflecting the intrinsic uncertainty in the response variable due to variability of breathing. The second term of the weighted sum reflects the distance between the training responses and the prediction estimate. This conforms to our intuition that the further way the prediction is from the 'known' training set, the less confidence we have (higher uncertainty) on the estimate. The derivation of (5) takes advantage of the Gaussian mixture structure and yields a computationally efficient form. Derivation details are provided in section 2.1.

2.4. Hypothesis testing

For each time k, the following null and alternative hypotheses are to be tested:

H_0: the prediction $\hat{\boldsymbol{y}}_k$ is accurate, i.e., close to the realization of $\boldsymbol{y}_k$;
H_1: the prediction $\hat{\boldsymbol{y}}_k$ is highly uncertain and may be far from the realization of $\boldsymbol{y}_k$.

To perform the hypothesis testing, we need to make statistical assumptions. Section 2.2 estimates the empirical distribution of $\boldsymbol{y}_k$, and the corresponding mean is obtained in section 2.3 as the prediction $\hat{\boldsymbol{y}}_k$. When the prediction value is strongly supported by the training samples, the conditional pdf of $\boldsymbol{y}_k$ should be concentrated, and we can assume $p(\boldsymbol{y}_k|\boldsymbol{x}_k)$ to be a Gaussian distribution $\mathcal{N}(\mu_k, \sigma_k^2)$ for the null hypothesis. On the other hand, high prediction uncertainty arises when the empirical distribution is not informative enough. It is typical in statistics to assume uniform distribution in this case. In short, the statistical formulation of the hypothesis testing is given by

H_0: $(\boldsymbol{y}_k|\boldsymbol{x}_k) \sim \mathcal{N}(\mu_k, \sigma_k^2)$;
H_1: $(\boldsymbol{y}_k|\boldsymbol{x}_k) \sim \mathcal{U}(\delta_L, \delta_U)$,

where δ_L and δ_U are the lower and upper boundary for the uniform distribution and are assumed to be constants across all k.

In order to perform the hypothesis test, we need to determine the parameters in the null hypothesis first. With access to the empirical distribution $p(\boldsymbol{y}_k|\boldsymbol{x}_k)$, we adopt the method of moments (MOM) (Hansen 1982) to estimate the parameters for the assumed Gaussian distribution, by matching the first and second moments (4), (5). This results in

$$\begin{aligned} \mu_k &= \hat{\boldsymbol{y}}_k = \frac{1}{\sum w_i} \sum_i w_i \boldsymbol{y}_i; \\ \sigma_k^2 &= \hat{\sigma}_{\boldsymbol{y}}^2 = \sigma_{\text{ker},y}^2 + \sum_i \frac{w_i}{\sum_i w_i} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_k)^2. \end{aligned} \tag{6}$$

From the detection perspective, the prediction value $\hat{\boldsymbol{y}}_k$ can be treated as a realization of the future location random variable, distributed according to $p(\boldsymbol{y}_k|\boldsymbol{x}_k)$. Therefore, the likelihood ratio test (LRT) (Mood *et al* 1974) reads

$$\begin{aligned} \Lambda(\hat{\boldsymbol{y}}_k) &= \frac{\mathcal{N}(\hat{\boldsymbol{y}}_k; \mu_k, \sigma_k)}{\mathcal{U}(\hat{\boldsymbol{y}}_k; \delta_L, \delta_U)} \\ &\propto \frac{1}{\sigma_k} \exp\left[-\left(\hat{y}_k - \mu_k\right)^2 / \sigma_k^2\right] \\ &= \frac{1}{\sigma_k}. \end{aligned} \tag{7}$$

The second line is obtained by dropping the constants from the uniform distribution as well as the $\frac{1}{\sqrt{2\pi}}$ factor in the Gaussian density. The third line is obtained by identifying the prediction value $\hat{y}_k$ as the mean of the random variable $\boldsymbol{y}_k$ under the conditional distribution (4).

A typical decision rule for the LRT reads

$$\begin{cases} \text{if } \Lambda > c, & \text{do not reject } H_0; \\ \text{if } \Lambda < c, & \text{reject } H_0; \\ \text{reject with probability } q & \text{if } \Lambda = \eta. \end{cases} \tag{8}$$

One could choose the values of c and q to satisfy a preset significance level α, so that the probability of mistakenly rejecting the null hypothesis is kept under the stated probability α. In this work, we adopt the Neyman–Pearson (Neyman and Pearson 1933) frequentist perspective and examine the probability of both type I and type II errors as the value of c varies. Furthermore, since the variance σ_k is a continuous variable, $\Lambda = \frac{1}{\sigma_k}$ has zero measure at the point c, indicating that a 'tie' rarely occurs, and we ignore this event hereafter.

We rewrite the LRT so that the decision rule is on σ_k rather than on its reciprocal:

$$\begin{cases} \text{if } \sigma_k < \eta, & \text{do not reject } H_0, \text{ 'likely good prediction'}; \\ \text{if } \sigma_k > \eta, & \text{reject } H_0, \text{ claim } H_1, \text{ 'potentially large prediction error'}. \end{cases} \tag{9}$$

This decision rule suggests to reject the null hypothesis when the estimated variance is high. In other words, a large variance estimate at t_k indicates that the prediction $\hat{\boldsymbol{y}}_k$ may be subject to a large prediction error. This conforms to our intuition that if the underlying variable is highly volatile, then its specific realization, which can only be observed retrospectively, cannot be predicted precisely.

2.4.1. Algorithmic description for location prediction and risk detection. Algorithm 1 summarizes the operation flow for simultaneously predicting the future target location and detecting potentially large prediction errors.

Algorithmic 1 real time location prediction and detection of a potentially large prediction error.

1: Determine covariance $\Sigma_{\text{ker},x}$ and $\sigma_{\text{ker},y}$ for covariate and response variables.
2: Preset the discrimination threshold η.
3: **for** each k **do**
4: Compute the weighting according to (3).
5: Compute the mean estimate $\hat{y}_k = \frac{w_i y_i}{\sum_i w_i}$, equivalent to (4).
6: Compute the variance estimate $\hat{\sigma}_k^2 = \sigma_{\text{ker},y}^2 + \frac{1}{\sum_i w_i} \sum_i w_i (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_k)^2$.
7: If $\hat{\sigma}_k > \eta$, claim a potentially large prediction error, and trigger risk management mechanisms.
8: **end for**

Although the derivations in section 2.2 and section 2.3 involve characterizing various probability density functions and integrations with respect to them, algorithm 1 is very simple with no explicit computation of the pdfs or integrations at all. This property leads to efficient implementations critical for real-time executions.

3. Experimental evaluation and results

3.1. Experiment setup

To evaluate the performance of the proposed large-prediction-error detector, we used patient-derived traces acquired with the Cyberknife Synchrony system at Georgetown University (Suh *et al* 2008). The KDE-based method (Ruan 2010) was used to predict tumor locations 160 ms and 570 ms ahead, spanning the range of typical system latencies (Keall *et al* 2006, Poulsen *et al* 2010). The specific implementation in this paper used three-dimensional augmented covariate variables so that $\boldsymbol{x}_i = [s_{i-2\Delta}, s_{i-\Delta}, s]$, with Δ corresponding to 0.5 s delay between the consecutive coordinates of the variate $\boldsymbol{x}$. A moving window of 20 s was used to dynamically update the training set $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, and the covariance matrices $(\Sigma_{\text{ker},x}, \sigma_{\text{ker},y})$ were estimated from their corresponding empirical values from the training set. In general, the performance of the KDE-based prediction algorithm is quite insensitive to the numerical values of these setup parameters (Ruan 2010).

The ground-truth[2] prediction errors were obtained by computing the discrepancy between the tumor locations predicted in real time and their corresponding retrospective observations:

$$e_k = \hat{\boldsymbol{y}}_k - \boldsymbol{y}_k. \tag{10}$$

Subsequently, the ground-truth large-prediction-error regions were identified by comparing such point-wise discrepancy (10) with a preset prediction error tolerance ε. In

[2] Strictly speaking, ground truth is unaccessible, due to the presence of measurement noise. In addition, the traces from Cyberknife data are partially inferred from a correlation model. However, for the purpose of this study, we treat the monitoring mechanism as a black-box module and focus on prediction.

general, this tolerance reflects the accuracy requirement of the treatment system and motion management mechanism. In our experiment, we used ε values of 1, 2, 3 and 4 mm:

$$\begin{cases} |e_k| < \varepsilon \Rightarrow \text{point } k \text{ is predicted with satisfactory accuracy, } H_0 \text{ true;} \\ |e_k| \geqslant \varepsilon \Rightarrow \text{point } k \text{ is predicted with prediction error larger than tolerance, } H_1 \text{ true.} \end{cases}$$

The online detection results were obtained according to (9). The performance of the proposed detection method was quantified with the receiver operating characteristic (ROC) curve, by plotting the true positive rate (sensitivity) versus the false positive rate (1-specificity), as the discrimination threshold η varies. More specifically,

$$\begin{aligned} &\text{true positive rate (TPR)} = \#\{\text{claim } H_1\}/\#\{H_1 \text{ true}\}; \\ &\text{false positive rate (FPR)} = \#\{\text{claim } H_1\}/\#\{H_0 \text{ true}\}. \end{aligned} \tag{11}$$

It is also of clinical interest to understand the relationship between the likelihood of failing to detect large errors (miss) versus delivery efficiency. Requiring high delivery efficiency means being oblivious to relatively weak indications of prediction errors (setting η high), resulting in higher risk of missing. We measured the delivery efficiency as the ratio of the beam-on time to the overall time duration, assuming detection of large errors triggers beam pause. To characterize the clinical tradeoff, we examined the relationship between the following two quantities:

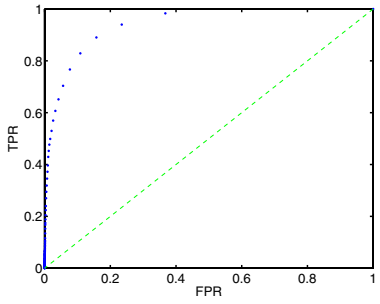
$$\begin{aligned} &\text{miss/false negative rate (FNR)} = \#\{\text{claim } H_0\}/\#\{H_1 \text{ true}\}; \\ &\text{delivery efficiency} = \#\{\text{claim } H_0\}/\{\text{total number of tests performed}\}. \end{aligned} \tag{12}$$

3.2. Results

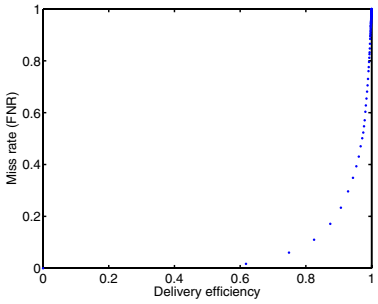
The KDE-based prediction method has been previously demonstrated as efficient and accurate, especially for approximately self-reproducible trajectories (Ruan 2010). For 159 Cyberknife synchrony traces, the collective performance of the proposed large error detection scheme for 160 ms and 570 ms prediction is depicted in figures 1 and 2, respectively. Large errors are defined as prediction errors exceeding 1, 2, 3 or 4 mm. We see that the proposed method performs well, as reflected by the ROC and clinical tradeoff curves. The ROC curves are close to the upper-left corner, indicating high sensitivity as well as high specificity. From the clinical tradeoff curve, we see that high delivery efficiency can be achieved with slight sacrifice of the detection sensitivity. Comparing across different definitions of a 'large prediction error', one observes incremental improvement of detection performance when the threshold for a 'large error' increases (e.g. 4 mm versus 1 mm), as random noise becomes less likely to corrupt the true H_0/H_1 affiliation.

For a closer performance examination, we have purposely selected two challenging cases for prediction.

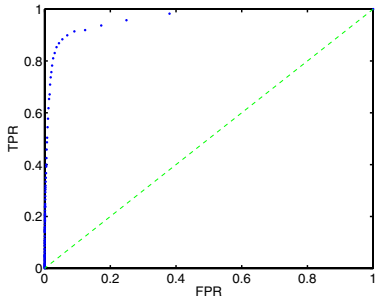
The first trajectory exhibits significant drifting, which makes it difficult for the KDE to learn the proper conditional pdf from training covariate/response pairs. Even so, the KDE method performs reasonably well, yielding a root mean squared error (RMSE) of 0.50 mm for 160 ms prediction and a RMSE of 1.19 mm for 570 ms prediction, as seen in figures 3 and 6. Furthermore, the 160 ms prediction error is uniformly below 2 mm, so the only applicable detection task is to identify the prediction errors exceeding 1 mm (figure 4). Figure 5(a) illustrates the ROC curve, the clinical miss rate versus efficiency tradeoff



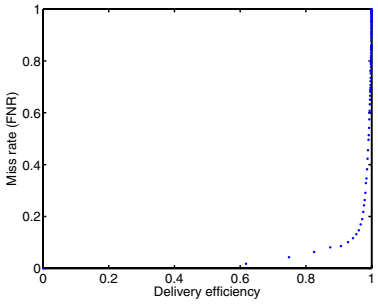
ROC for detecting $>$ 1mm error



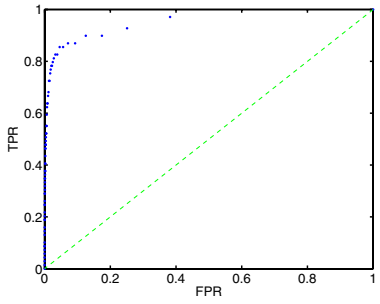
miss-detection rate vs. delivery efficiency for $>$ 1mm error



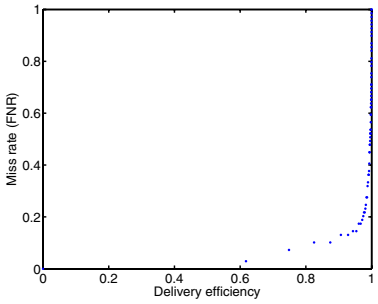
ROC for detecting $>$ 2mm error



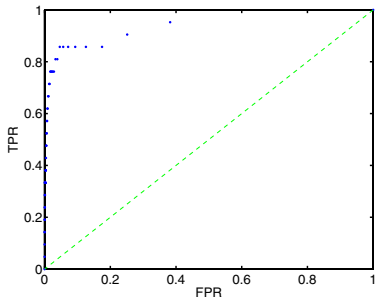
miss-detection rate vs. delivery efficiency for $>$ 2mm error



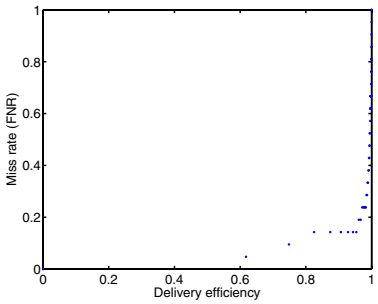
ROC for detecting $>$ 3mm error



miss-detection rate vs. delivery efficiency for $>$ 3mm error

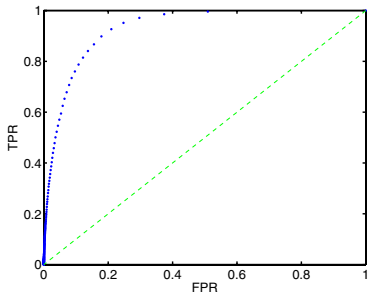


ROC for detecting $>$ 4mm error

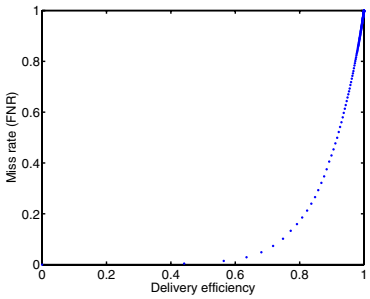


miss-detection rate vs. delivery efficiency for $>$ 4mm error

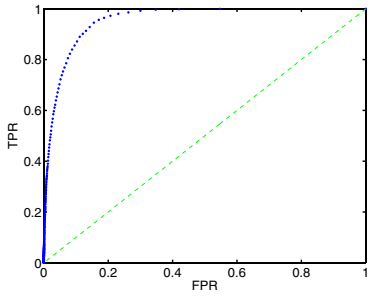
Figure 1. Detection performance for 160 ms prediction length. Left column: ROC curve for detecting the prediction error exceeding 1 mm. Blue dots represent the performance in terms of FPR and TPR at various discrimination thresholds η, and the diagonal dashed green line illustrates the detection performance with a random decision rule. Right column: miss rate versus delivery efficiency.



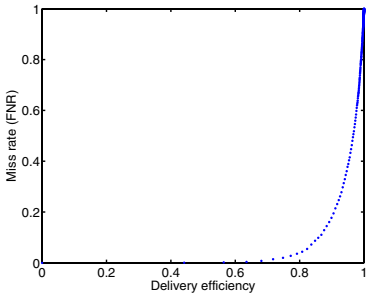
ROC for detecting $>$ 1mm error



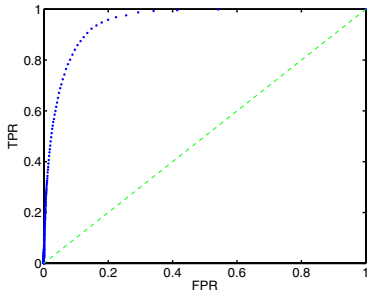
miss-detection rate vs. delivery efficiency for $>$ 1mm error



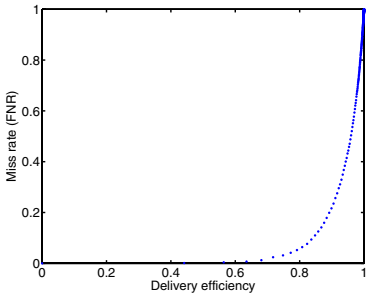
ROC for detecting $>$ 2mm error



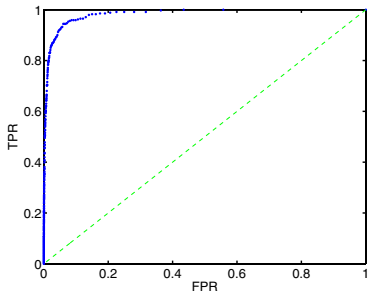
miss-detection rate vs. delivery efficiency for $>$ 2mm error



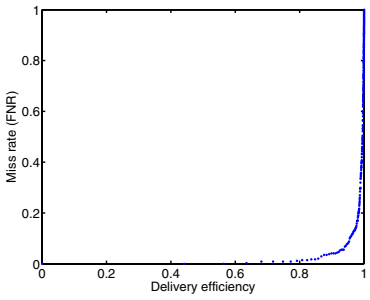
ROC for detecting $>$ 3mm error



miss-detection rate vs. delivery efficiency for $>$ 3mm error



ROC for detecting $>$ 4mm error



miss-detection rate vs. delivery efficiency for $>$ 4mm error

Figure 2. Detection performance for 570 ms prediction length. Left column: ROC curve for detecting the prediction error exceeding 1 mm. Blue dots represent the performance in terms of FPR and TPR at various discrimination thresholds η, and the diagonal dashed green line illustrates the detection performance with a random decision rule. Right column: miss rate versus delivery efficiency.

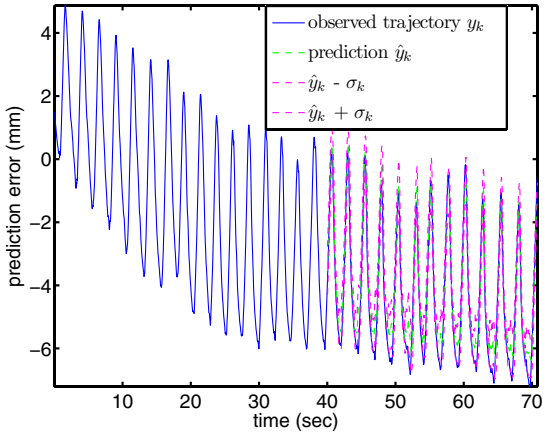


Figure 3. Observed trajectory versus real-time prediction for 160 ms. The root mean squared prediction error was 0.50 mm. An uncertainty band was generated by adding to and subtracting from the prediction value the estimated standard deviation $\hat{\sigma}_k$.

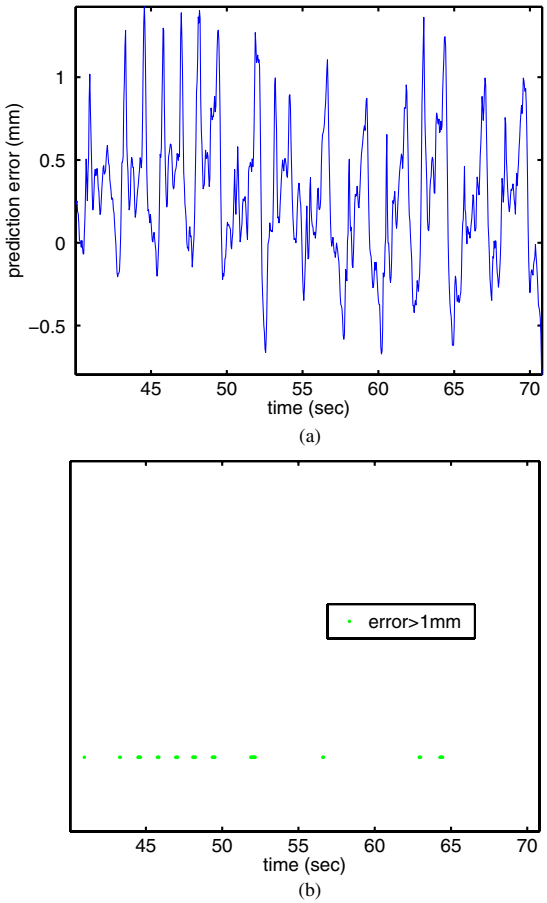


Figure 4. Retrospective identification of regions with large prediction errors. (a) Prediction error and (b) large prediction error region.

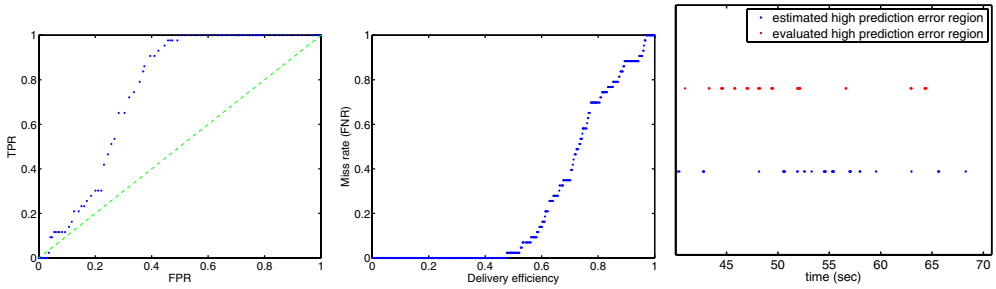


Figure 5. Detection performance. Left column: ROC curve for detecting the prediction error exceeding 1 mm. Blue dots represent the performance in terms of FPR and TPR at various discrimination thresholds η, and the diagonal dashed green line illustrates the detection performance with a random decision rule. Middle column: miss-detection rate versus delivery efficiency. Right column: an exemplary detection trace. Blue dots depict the prospectively estimated occurrence of the large prediction error, and red dots show the 'ground-truth'.

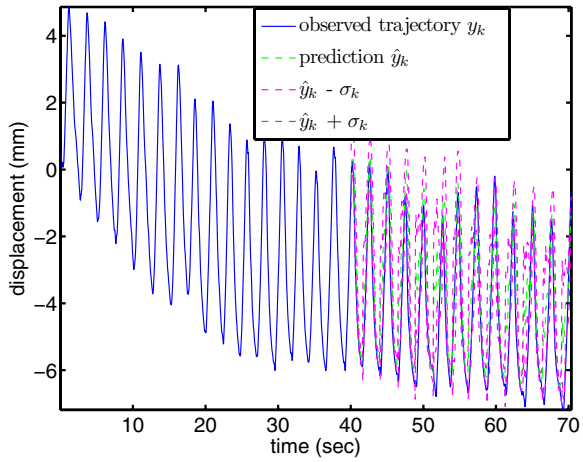


Figure 6. Observed trajectory versus real-time prediction for 570 ms. The root mean squared prediction error was 1.19 mm. An uncertainty band was generated by adding to and subtracting from the prediction value the estimated standard deviation $\hat{\sigma}_k$.

curve and one exemplary detection time series corresponding to a single point on the ROC curve.

For 570 ms prediction, the largest prediction error exceeds 3 mm (figure 7), and we perform detection for errors exceeding $1, 2$ and 3 mm. Figure 8 reports the ROC curve, the clinical tradeoff curve and an exemplary detection trace for each error definition, respectively.

The second trajectory presents the biggest challenge among all Synchrony traces under test—it exhibits irregularity in trend, oscillatory magnitude, as well as phase perturbation. The RMSEs for 160 ms and 570 ms prediction are 1.85 mm and 3.34 mm, respectively (cf figures 9 and 12), with the maximum prediction errors exceeding 5 mm (figures 10 and 13).

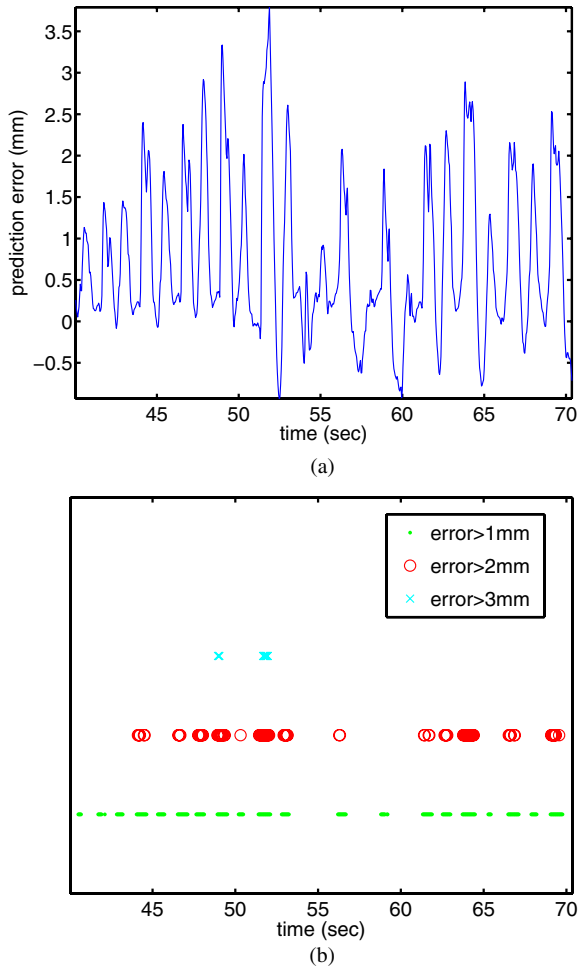
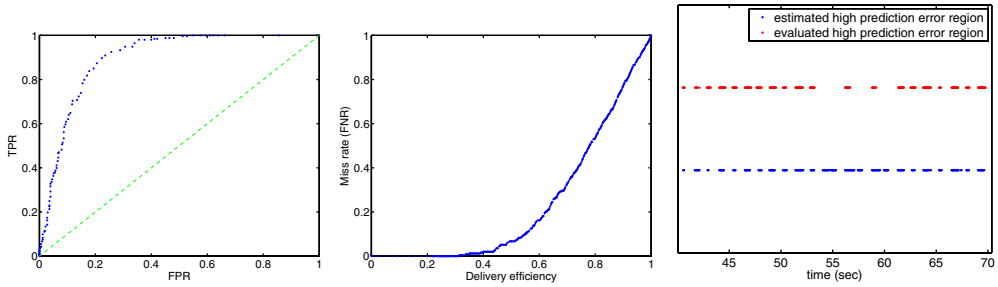


Figure 7. Retrospective identification of regions with large prediction errors. (a) Prediction error and (b) regions of the large prediction error.

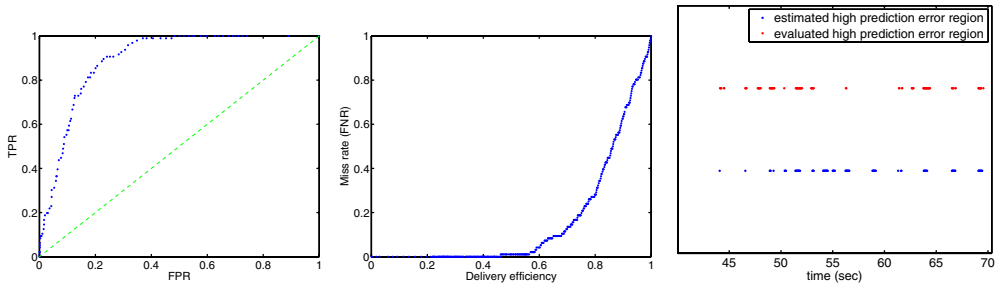
The same hypothesis testing procedure has been performed and the results are reported in figures 11 and 14.

4. Discussions and concluding remarks

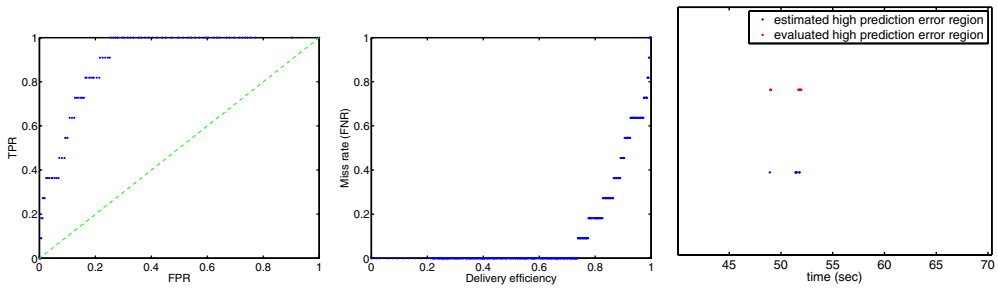
Despite the intensive studies on prediction methods for respiratory motion, rigorous research on prospectively estimating prediction performance remains elusive. This is because physiological/physical motion often deviates from strictly defined mechanical models, and the noise hardly follows any simple statistical distribution. In this study, we take advantage of the empirical pdf, developed with the KDE method, and interpret the variance estimate as a measure of 'uncertainty', a notion tightly related to prediction confidence. The problem of prospectively detecting large prediction errors greater than a tolerance is cast



ROC, miss-detection vs. efficiency, and exemplary detection trace for $>$ 1mm error



ROC, miss-detection vs. efficiency, and exemplary detection trace for $>$ 2mm error



ROC, miss-detection vs. efficiency, and exemplary detection trace for $>$ 3mm error

Figure 8. ROC curves and exemplary detection traces for prediction errors exceeding 1, 2 and 3 mm, respectively.

in a binary hypothesis testing setup. The null hypothesis is posed as a Gaussian distribution concentrated around the prediction estimate and the alternative hypothesis invokes a non-informative uniform distribution. The LRT is used to determine whether sufficient evidence exists to reject the assumption that the prediction value is a representative realization of a Gaussian random variable defined by the null hypothesis. We use the MOM to estimate the parameters for the null hypothesis and reduce the corresponding LRT to a simple comparison rule between the estimated variance of the test response variable and a discrimination threshold. Thanks to the KDE of the empirical estimate, which represents the pdf of the prediction as a Gaussian mixture; the prediction variance can be computed efficiently without explicit integration. This promises feasibility for real-time implementation of the detection strategy.

The ROC is a powerful way of characterizing detector performance, when the truth is available. In our experiments, the notion of 'ground-truth' needs to be interpreted appropriately

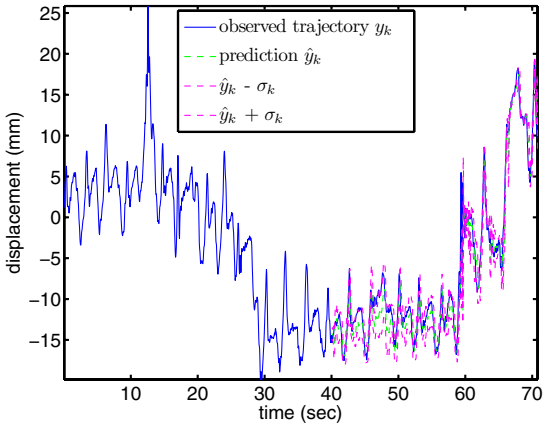


Figure 9. Observed trajectory versus real-time prediction for 160 ms. The root mean squared prediction error is 1.85 mm. An uncertainty band is generated by adding to and subtracting from the prediction value the estimated standard deviation $\hat{\sigma}_k$.

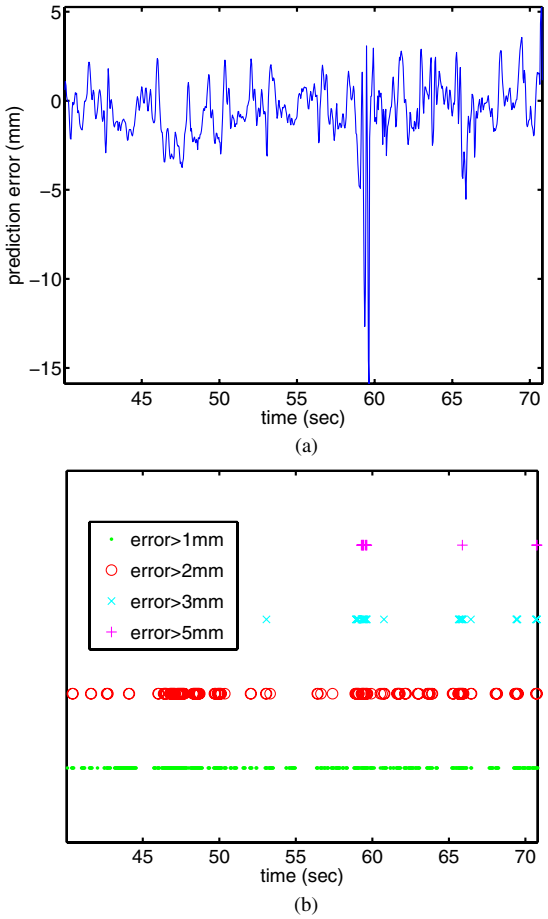
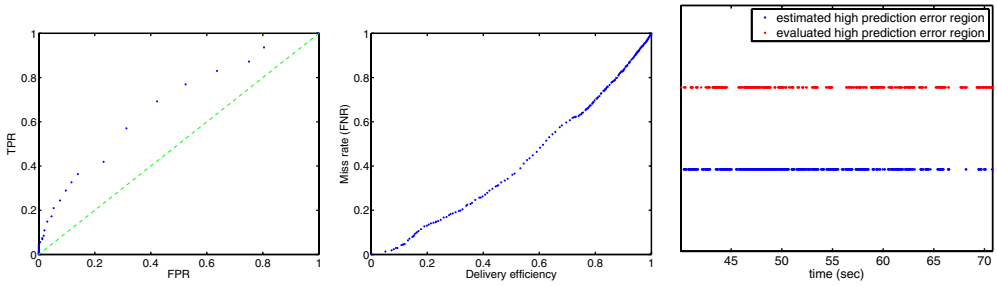
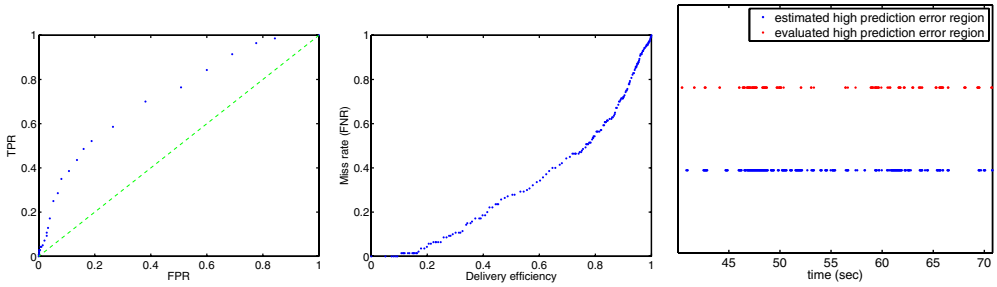


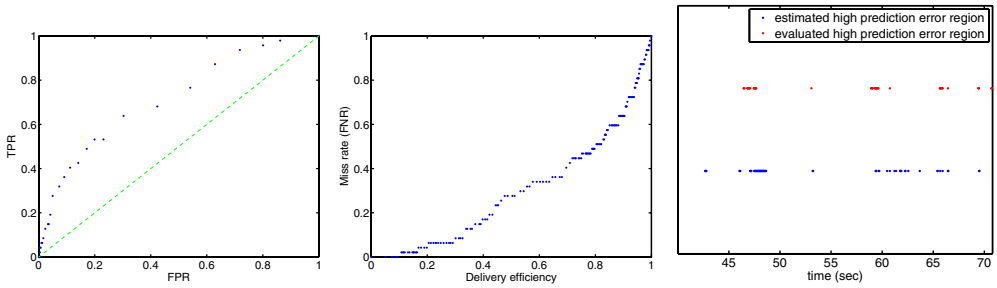
Figure 10. Retrospective identification of regions with large prediction errors. (a) Prediction error and (b) regions of the large prediction error.



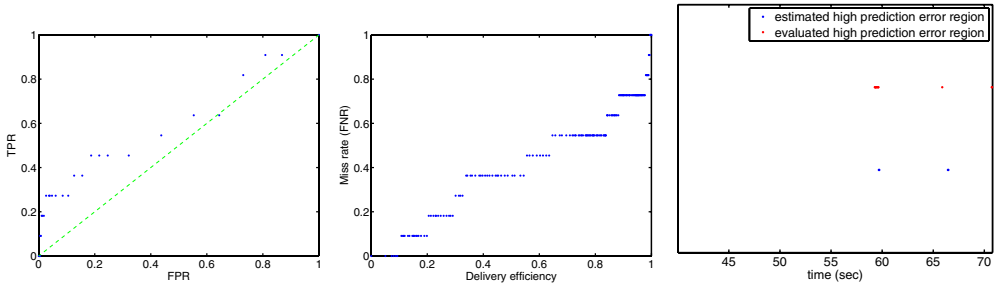
ROC, miss-detection vs. efficiency, and exemplary detection trace for $>$ 1mm error



ROC, miss-detection vs. efficiency, and exemplary detection trace for $>$ 2mm error



ROC, miss-detection vs. efficiency, and exemplary detection trace for $>$ 3mm error



ROC, miss-detection vs. efficiency, and exemplary detection trace for $>$ 5mm error

Figure 11. ROC curves and exemplary detection traces for prediction errors exceeding 1, 2, 3, and 5 mm, respectively.

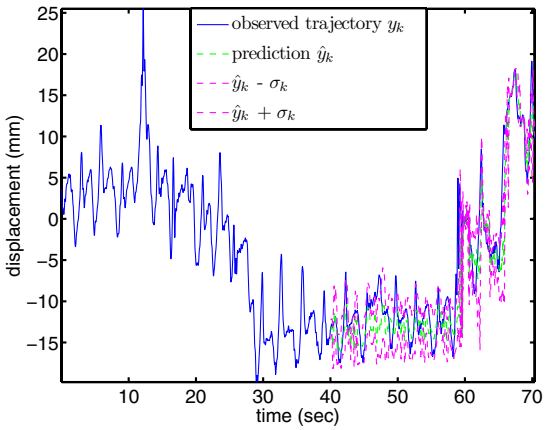


Figure 12. Observed trajectory versus real-time prediction for 570 ms. The root mean squared prediction error is 3.34 mm. An uncertainty band is generated by adding to and subtracting from the prediction value the estimated standard deviation $\hat{\sigma}_k$.

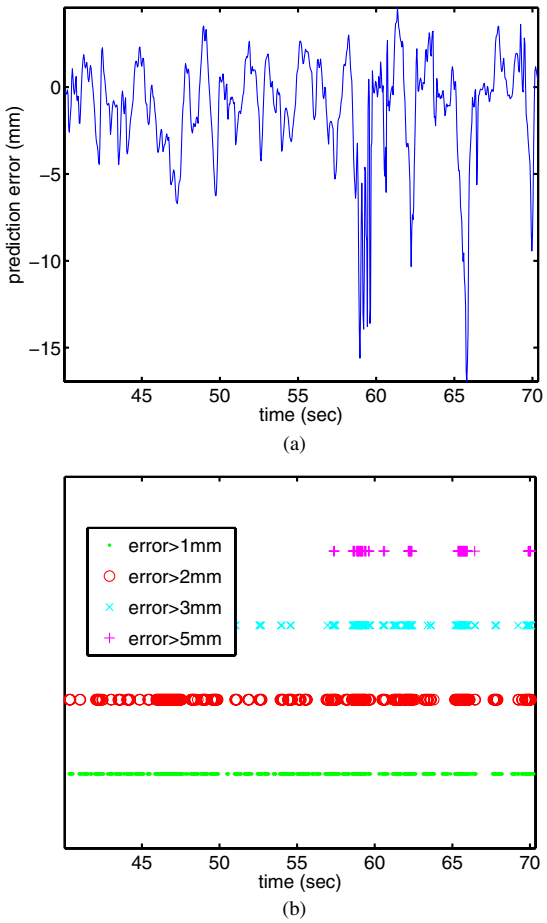
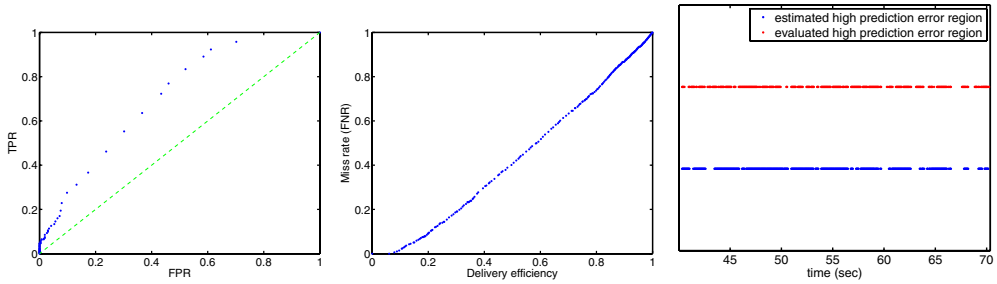
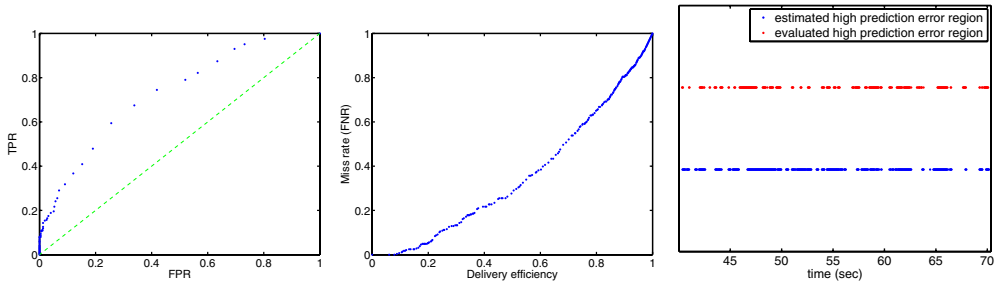


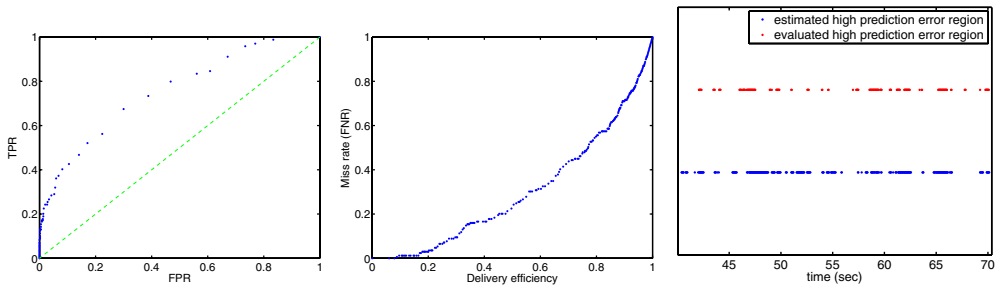
Figure 13. Retrospective identification of regions with large prediction errors. (a) Prediction error and (b) regions of the large prediction error.



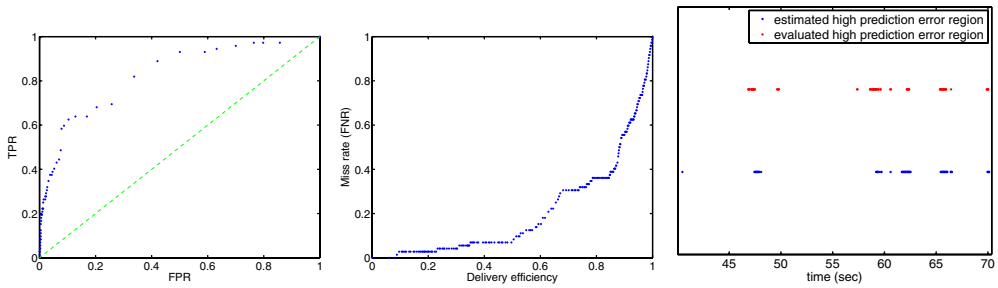
ROC, miss-detection vs. efficiency, and exemplary detection trace for $>$ 1mm error



ROC, miss-detection vs. efficiency, and exemplary detection trace for $>$ 2mm error



ROC, miss-detection vs. efficiency, and exemplary detection trace for $>$ 3mm error



ROC, miss-detection vs. efficiency, and exemplary detection trace for $>$ 5mm error

Figure 14. ROC curves and exemplary detection traces for prediction errors exceeding 1, 2, 3 and 5 mm, respectively.

and taken with caution. Firstly, all observations were acquired in the presence of noise. Furthermore, observations acquired with a Synchrony system are also subject to modeling error in internal/external inference. More importantly, an instantaneous observation is a single sample of the possible outcome at the time of acquisition, given the stochastic nature of

the underlying motion. Therefore, the observation is representative of the *true* tumor location only to a certain degree. In other words, a large discrepancy between the prediction and the retrospective observation value for a single trace may not necessarily imply high likelihood of a large prediction error for another realization of the same underlying trace. This suggests that in general ROC curves on the left columns of figures 5, 8, etc are a more informative performance measure than the straight point-wise detection trace comparison on the right columns of the same figures.

It is also noticeable that the ROC curves approach the upper-left corner of the unit square $[0, 1] \times [0, 1]$ better for more regular trajectories. This is expected, as irregularity poses more challenge for estimating the pdf of the response variable and affects all of its moments, including the first moment as the prediction estimate as well as the second moment for the variance estimate. The clinical tradeoff curves in terms of the miss detection rate versus delivery efficiency share the same underlying information as those of the ROC and expresses such indications explicitly.

The discrimination threshold η in the LRT determines the operating point on the ROC curve and should be chosen based on the specific application. For example, if one wants to detect a potentially large prediction error in order to decide whether to trigger beam pause, then rejections of the null hypothesis lead to sacrifice of delivery efficiency, which is a significant cost. Therefore, it may be preferable to operate near the lower-left corner of the ROC curve, and select a relatively big η, to avoid too much beam pauses. On the other hand, if the detection mechanism is merely used to decide whether to obtain more observations or to update models as in Synchrony systems, then the cost of a 'false positive' is relatively small compared to a 'false negative', and it would be desirable to operate at the upper-right corner of the ROC curve, with a small η value. The aforementioned analysis is based on a single ROC, which corresponds to detection with respect to a specific error tolerance ε. However, the detection scheme itself does not rely on it, which gives rise to an alternative interpretation of the discrimination threshold η: a large η corresponds to a relatively relaxed detection task of finding $|e_k| > \varepsilon$ for a large ε; while a small η corresponds to a more strict detection of finding all $|e_k| > \varepsilon$ for a small ε. With the LRT being a thresholding decision rule on $\hat{\sigma}_y$, it is obvious that the detected large-prediction-error sets $C_\eta = \{k : \text{point } k \text{ claimed to have a large prediction error}\}$ form an ordered sequence as η decreases according to the set inclusion relation.

In practice, one could start with a training segment, estimate the prediction variance $\hat{\sigma}_y$ and then incrementally 'swipe' $\hat{\sigma}_y$ until the appropriate discrimination threshold value η is found.

To summarize, this study presents the first rigorous study on prospectively detecting large prediction errors in real time. The intermediate estimate of the prediction variance provides confidence-interval type information on the prediction output. The hypothesis testing approach for detection offers flexibility in choosing (1) the definition of the 'large prediction error' and (2) the tradeoff between type I and type II detection errors. Development and derivations based on the KDE method offers an efficient real-time implementation of the proposed approach. The result of this study can be used to automatically trigger the adjustment of motion management schemes, such as gating and tracking, in a prospective fashion.

Acknowledgments

This study is partially supported by NIH/NCI R01 93626 and AAPM Seed Funding Initiative. The author is grateful to Dr Paul Keall for his continuous encouragement and support.

Appendix. Derivation of the second moment

Substituting the Gaussian mixture pdf into the definition of the second moment yields

$$\begin{aligned}\hat{\sigma}_{\boldsymbol{y}}^2 &= E[(Y_k - E(Y_k|\boldsymbol{x}_k))^2|\boldsymbol{x}_k] \\ &= \frac{1}{\tilde{C}}\int \sum_i (\boldsymbol{y}-\hat{\boldsymbol{y}}_k)^2 w_i \exp\left[-\|\boldsymbol{y}-\boldsymbol{y}_i\|^2/2\sigma_{\mathrm{ker},y}^2\right] \mathrm{d}\boldsymbol{y} \\ &= \frac{1}{\tilde{C}}\sum_i w_i \int (\boldsymbol{y}-\hat{\boldsymbol{y}}_k)^2 \exp\left[-\|\boldsymbol{y}-\boldsymbol{y}_i\|^2/2\sigma_{\mathrm{ker},y}^2\right] \mathrm{d}\boldsymbol{y}, \end{aligned} \tag{A.1}$$

with the normalization factor $\tilde{C} = \sqrt{2\pi}\sigma_{\mathrm{ker},y}\sum_i w_i$.

Note that each integral can be related to the variance of a Gaussian component in the mixture pdf as follows. Let g_i be the pdf for $\mathcal{N}(\boldsymbol{y}_i, \sigma_{\mathrm{ker},y}^2)$,

$$g_i(\boldsymbol{y}) = \frac{1}{C}\exp\left[-\|\boldsymbol{y}-\boldsymbol{y}_i\|^2/2\sigma_{\mathrm{ker},y}^2\right],$$

where $C = \sqrt{2\pi}\sigma_{\mathrm{ker},y}$.

Then each integral component in (5) can be rewritten as

$$\begin{aligned}\int (\boldsymbol{y}-\hat{\boldsymbol{y}}_k)^2 \exp\left[-\|\boldsymbol{y}-\boldsymbol{y}_i\|^2/2\sigma_{\mathrm{ker},y}^2\right] \mathrm{d}\boldsymbol{y} &= C\int (\boldsymbol{y}-\hat{\boldsymbol{y}}_k)^2 g_i(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\ &= C\int (\boldsymbol{y}-\boldsymbol{y}_i+\boldsymbol{y}_i-\hat{\boldsymbol{y}}_k)^2 g_i(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\ &= C\int (\boldsymbol{y}-\boldsymbol{y}_i)^2 g_i(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} + C\int (\boldsymbol{y}_i-\hat{\boldsymbol{y}}_k)^2 g_i(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\ &= C\left\{\sigma_{\mathrm{ker},y}^2 + (\boldsymbol{y}_i-\hat{\boldsymbol{y}}_k)^2\right\}. \end{aligned} \tag{A.2}$$

Let E_i denote the expectation with respect to the pdf g_i. The cross term for integrating $(\boldsymbol{y}-\boldsymbol{y}_i)(y_i-\hat{\boldsymbol{y}}_{\mathrm{mean}})$ is dropped from line 3 to line 4 because $E_i[Y-\boldsymbol{y}_i] = 0$. The last line follows from the facts that the covariance of the ith Gaussian component is $E_i[(Y-\boldsymbol{y}_i)^2] = \sigma_{\mathrm{ker},y}^2$ and that the pdf g_i integrates to unity in the second term.

Substituting (A.2) into (A.1), we obtain

$$\begin{aligned}\hat{\sigma}_{\boldsymbol{y}}^2 &= \frac{1}{\sum_i w_i}\sum_i w_i\left\{\sigma_{\mathrm{ker},y}^2 + (\boldsymbol{y}_i-\hat{\boldsymbol{y}}_k)^2\right\} \\ &= \sigma_{\mathrm{ker},y}^2 + \sum_i \frac{w_i}{\sum_i w_i}(\boldsymbol{y}_i-\hat{\boldsymbol{y}}_k)^2. \end{aligned} \tag{A.3}$$

References

Duda R O, Hart P E and Stork D G 2001 *Pattern Classification* (New York: Wiley)

Ernst F, Schlaefer A and Schweikard A 2007 Prediction of respiratory motion with wavelet-based multiscale autoregression *Med. Image Comput. Comput.-Assist. Intervention* **10** 668–75

Hansen L P 1982 Large sample properties of generalized method of moment estimators *Econometrica* **50** 1029–54

Isaksson M, Jalden J and Murphy M J 2005 On using an adaptive neural network to predict lung tumor motion during respiration for radiotherapy applications *Med. Phys.* **32** 3801–9

Kakar M, Nystrom H, Aarup L R, Nottrup T J and Olsen D R 2005 Respiratory motion prediction by using the adaptive neuro fuzzy inference system (ANFIS) *Phys. Med. Biol.* **50** 4721–8

Keall P J, Cattell H, Pokhrel D, Dieterich S, Wong K H, Murphy M J, Vedam S S, Wijesooriya K and Mohan R 2006 Geometric accuracy of a real-time target tracking system with dynamic multileaf collimator tracking system *Int. J. Radiat. Oncol. Biol. Phys.* **65** 1579–84

Keall P J, Kini V R, Vedam S S and Mohan R 2002 Potential radiotherapy improvements with respiratory gating *Australas. Phys. Eng. Sci. Med.* **25** 1–6

McCall K C and Jeraj R 2007 Dual-component model of respiratory motion based on the periodic autoregressive moving average (periodic ARMA) method *Phys. Med. Biol.* **52** 3455–66

Mood A M, Graybill F A and Boes D C 1974 *Introduction to the Theory of Statistics* 3rd edn (New York: McGraw-Hill)

Murphy M J and Dieterich S 2006 Comparative performance of linear and nonlinear neural networks to predict irregular breathing *Phys. Med. Biol.* **51** 5903–14

Neyman J and Pearson E S 1933 On the problem of the most efficient tests of statistical hypotheses *Phil. Trans. R. Soc.* A **231** 289–337

Nuyttens J J, Prevost J B, Praag J, Hoogeman M, Van Klaveren R J, Levendag P C and Pattynama P M 2006 Lung tumor tracking during stereotactic radiotherapy treatment with the cyberknife: marker placement and early results *Acta. Oncol.* **45** 961–5

Poulsen P, Cho B, Ruan D, Sawant A and Keall P 2010 Dynamic MLC tracking of respiratory target motion based on a single kilovoltage imager during arc radiotherapy *Int. J. Radiat. Oncol. Biol. Phys.* **76** 914–23

Putra D, Haas O C, Mills J A and Burnham K J 2008 A multiple model approach to respiratory motion prediction for real-time IGRT *Phys. Med. Biol.* **53** 1651–63

Ruan D 2010 Kernel density estimation based real-time prediction for respiratory motion *Phys. Med. Biol.* **55** 1311–26

Ruan D, Fessler J A and Balter J M 2007 Real-time prediction of respiratory motion based on nonparametric local regression methods *Phys. Med. Biol.* **52** 7137–52

Suh Y, Dieterich S, Cho B and Keall P J 2008 An analysis of thoracic and abdominal tumour motion for stereotactic body radiotherapy patients *Phys. Med. Biol.* **53** 3623–40

Vedam S S, Keall P J, Docef A, Todor D A, Kini V R and Mohan R 2004 Predicting respiratory motion for four-dimensional radiotherapy *Med. Phys.* **31** 2274–83