Respiratory Motion Compensation with Relevance Vector Machines

Robert Dürichen[1,2], Tobias Wissel[1,2], Floris Ernst[1], and Achim Schweikard[1]

[1] Institute of Robotics and Cognitive Systems, University of Lübeck, Germany
[2] Graduate School for Computing in Medicine and Life Sciences, Lübeck, Germany
{duerichen,wissel,ernst,schweikard}@rob.uni-luebeck.de

Abstract. In modern robotic radiation therapy, tumor movements due to respiration can be compensated. The accuracy of these methods can be increased by time series prediction of external optical surrogates. An algorithm based on relevance vector machines (RVM) is introduced. We evaluate RVM with linear and nonlinear basis functions on a real patient data set containing 304 motion traces and compare it with a wavelet based least mean square algorithm (wLMS), the best algorithm for this data set so far. Linear RVM outperforms wLMS significantly and increases the prediction accuracy for 80.3 % of the data. We show that real time prediction is possible in case of linear RVM and discuss how the predicted variance can be used to construct promising hybrid algorithms, which further reduce the prediction error.

Keywords: respiratory motion compensation, relevance vector machine, radiotherapy, bayesian learning.

1 Introduction

A challenging problem in modern stereotatic body radiation therapy (SBRT) is the precise irradiation of moving tumors in chest and abdomen while sparing critical surrounding structures. These movements are mainly caused by breathing and by cardiac motion, and can have an amplitude of up to 5 cm in extreme cases [1]. This motion can be compensated by new technical developments like multileaf collimators, robotic patient couches, Vero or the CyberKnife® Synchrony system [2]. All mentioned techniques must compensate for various sources of latencies. In the case of the CyberKnife Synchrony system, the latency is 115 ms and is mainly due to mechanical limitations, image acquisition and processing time. Other systems, e.g. robotic couches, can have time delays of up to $300 - 400$ ms. This systematic error can be reduced by time series prediction.

In recent years, several novel approaches have been investigated including neural networks, wavelets and support vector regression (SVR) algorithms [3,4]. In [5], Ernst *et al.* presented one of the most comprehensive studies so far. The authors compared six prediction algorithms on a set of 304 motion traces, which are available online. On average, the highest prediction accuracy was achieved using a wavelet based least mean square (wLMS) algorithm [6], followed by multi

K. Mori et al. (Eds.): MICCAI 2013, Part II, LNCS 8150, pp. 108–115, 2013.

linear step method (MULIN) and SVR algorithm, which exhibit a slightly worse performance.

Here, we present a prediction algorithm based on relevance vector machines (RVM). To the authors best knowledge, this is the first study using a probabilistic approach for respiratory motion compensation in SBRT. RVM has several advantages compared to non-probabilistic algorithms. First, part of the algorithm's internal parameters, the so called hyperparameters, can be estimated by maximizing the marginal likelihood without further cross-validation. Second, due to explicitly chosen prior distributions on the hyperparameters, prediction errors caused by overfitting can be reduced automatically. Third, the framework of RVM is capable of incorporate linear and nonlinear basis functions. Fourth, every predicted point is the mean of a predicted distribution. The variance of the predicted distribution can be used as an indicator of the prediction accuracy.

2 Methods

It is assumed that the external optical surrogates are equidistantly sampled with sampling rate f_s. Let t be the index for the current time step and y_t the measured signal amplitude at time step t. Let δ be the prediction horizon and $y_{t+\delta}$ the predicted point. The result of the prediction algorithm is $\hat{y}_{t+\delta}$ depending on the previous M points - the number of features. According to the current version of the CyberKnife, the prediction horizon was set to $\delta = 3$ at a sample rate of $f_s = 26\,\mathrm{Hz}$.

2.1 Motion Data Set

For our experiments, we use the data presented by [5], which is available online (http://signals.rob.uni-lubeck.de). It consists of 304 motion traces from 31 patients with durations from $80 - 150\,\mathrm{min}$ recorded at Georgetown University Hospital. Ernst *et al.* analyzed the motion traces with six prediction algorithms, namely: support vector regression (SVR), extended Kalman Filter, MULIN, recursive least square (RLS), normalized least mean square (nLMS) and wLMS. The latter showed the best performance among the algorithms. The basic idea of wLMS is to decompose a signal using an à trous wavelet into a superposition of $J + 1$ scales. On each scale a least mean square prediction can be performed [6]. To be consistent with [5], the algorithms are evaluated with respect to the relative root mean square (RMS$_{rel}$), which is defined as:

$$RMS_{rel} = \frac{\sqrt{\sum\limits_{i=1+\delta}^{N} (y_i - \hat{y}_i)^2 / N}}{\sqrt{\sum\limits_{i=1+\delta}^{N} (y_i - y_{i-\delta})^2 / N}} \tag{1}$$

RMS$_{rel}$ compares the squared prediction error relative to the squared prediction error in case of no prediction. For RMS$_{rel} < 100\,\%$, the algorithm would improve the prediction. To compare the predicted variances with each other, the motion traces have been scaled to $y_t \in [0, 1]$, which is not necessary in real applications.

Even though the data set only consists of 1D motion traces, the analysis could easily be extended to 3D signals by applying the prediction algorithms to the x-, y- and z-component of a signal separately.

2.2 Relevance Vector Machine (RVM)

In [7], Tipping proposed a sparse Bayesian learning framework for regression and classification tasks. If we assume autoregressive (AR) properties of the signal, a point $y_{t+\delta}$ can be predicted by:

$$y_{t+\delta} = \hat{y}_{t+\delta} + \varepsilon_{t+\delta} = \mathbf{w}^T \mathbf{y}_t + \varepsilon_{t+\delta} \tag{2}$$

where $\mathbf{y}_t \in \mathbb{R}^M$ is a vector of M measured points $\mathbf{y}_t = [y_t, ...y_{t-M+1}]^T$ and $\varepsilon_{t+\delta}$ the measurement noise at time $t + \delta$. The aim of the prediction algorithm is to learn the optimal weight vector $\mathbf{w} \in \mathbb{R}^M$. To account for nonlinear relationships between $\hat{y}_{t+\delta}$ and $\mathbf{y}_t$, the model can be extended using a basis function $\boldsymbol{\phi}(\mathbf{y_t})$:

$$y_{t+\delta} = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{y}_t) + \varepsilon_{t+\delta} \tag{3}$$

For robust training of $\mathbf{w}$, N training pairs can be considered simultaneously. Thus, the vector $\boldsymbol{\phi}(\mathbf{y}_t)$ is extended to a design matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N \times M}$, where $\boldsymbol{\Phi}_{t-\delta} = [\boldsymbol{\phi}(\mathbf{y}_{t-\delta}), \boldsymbol{\phi}(\mathbf{y}_{t-\delta-1}), ..., \boldsymbol{\phi}(\mathbf{y}_{t-\delta-N+1})]^T$:

$$\mathbf{y}_t = \boldsymbol{\Phi}_{t-\delta} \mathbf{w} + \boldsymbol{\varepsilon}_t \tag{4}$$

The noise vector $\boldsymbol{\varepsilon}_t$ can be assumed to be normally distributed with zero mean and a variance of σ^2 ($\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \sigma^2)$). Therefore, the probability $P(\mathbf{y}_t | \mathbf{y}_{t-\delta}, \mathbf{w}, \sigma^2)$ of the vector $\mathbf{y}_t$ given $\mathbf{w}$, σ^2 and the trainings data matrix $\mathbf{y}_{t-\delta}$ is also a normal distribution and can be expressed as:

$$P(\mathbf{y}_t | \mathbf{y}_{t-\delta}, \mathbf{w}, \sigma^2) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y}_t - \boldsymbol{\Phi}_{t-\delta}\mathbf{w}\|\right) \tag{5}$$

To constrain the algorithm to smooth and less complex functions and to avoid overfitting, a prior probability distribution on $\mathbf{w}$ can be defined. By choosing a zero-mean Gaussian prior

$$P(\mathbf{w}_t | \boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(0, \alpha_i^{-1}), \tag{6}$$

the algorithm will prefer weights which are zero or close to it and will penalize large weights within the vector. All nonzero weights are the so called relevance vectors (RV). The parameters $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_M]^T$ and σ^2 are called hyperparameters. After choosing an initial value, the unknown hyperparameters can be estimated iteratively by maximizing the posterior probability $P(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2 | \mathbf{y}_t)$ which

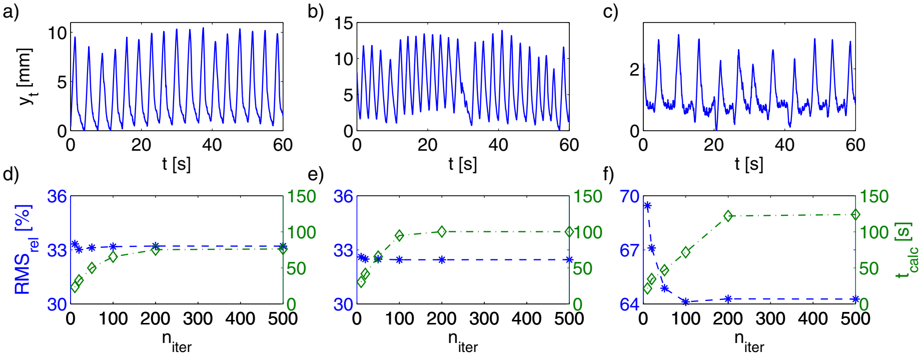

Fig. 1. Amplitude-time plot of the selected motion traces: (a) regular, (b) irregular and (c) irregular/noisy; (d-f) optimized relative RMS error ($*$) and compuation time ($\diamond$) depending on number of iterations n_{iter} for the selected motion traces (a-c)

is the probability over all the unknown parameters, given the data. Here, n_{iter} defines the number of iterations to compute the optimized hyperparameters $\boldsymbol{\alpha}^*$ and σ^{2*}. Using Bayes' rule, the posterior distribution can be defined as:

$$P(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2 | \mathbf{y}_t) = \frac{P(\mathbf{y}_t | \mathbf{w}, \boldsymbol{\alpha}, \sigma^2) P(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2)}{P(\mathbf{y}_t)} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{7}$$

which is a normal distribution with a mean $\boldsymbol{\mu}$ and a covariance $\boldsymbol{\Sigma}$ of

$$\boldsymbol{\mu} = \sigma^{-2} \boldsymbol{\Sigma} \boldsymbol{\Phi}_{t-\delta}^T \mathbf{y}_t, \qquad \boldsymbol{\Sigma} = \left(\mathbf{A} + \sigma^{-\mathbf{2}} \boldsymbol{\Phi}_{t-\delta}^T \boldsymbol{\Phi}_{t-\delta})^{-1} \right) \text{ with } \mathbf{A} = diag(\boldsymbol{\alpha}). \tag{8}$$

The derivation of (8), the optimization of the hyperparameters is beyond the scope of this paper and can be found in [7]. The optimized hyperparameters $\boldsymbol{\alpha}^*$ and σ^{2*} can be used to make a prediction with a new test set $\mathbf{y}_t$

$$\hat{y}_{t+\delta} = \boldsymbol{\mu}^T \boldsymbol{\phi}(\mathbf{y}_t), \qquad \hat{\sigma}_{t+\delta}^2 = \sigma^{2*} + \boldsymbol{\phi}(\mathbf{y}_t)^T \boldsymbol{\Sigma} \boldsymbol{\phi}(\mathbf{y}_t). \tag{9}$$

The predicted variance $\hat{\sigma}_{t+\delta}^2$ is the sum of the variances caused by the measurement noise and the uncertainty in the prediction of $\mathbf{w}_t$.

The algorithm was implemented in Matlab using the SBS toolbox on an office computer (i7@3.4GHz, 16GB RAM).

3 Results

The first experiment focuses on the influence of the RVM specific parameters and computation time. Figure 1.a-c shows three motion fragments. Figure 1.a represents a regular, Fig.1.b. a "slightly" irregular and Fig.1.c a noisy and irregular breathing signal. Each fragment has a duration of 60 s. The performance of RVM was optimized using a grid search with respect to the number of features M, the number of training pairs N and maximum number of

Table 1. Optimized relative RMS error and computation time of linear RVM for a regular, irregular and irregular/noisy signal for different number of iteration n_{iter}; optimized by grid search for the number of features M and the number of training pairs N and for comparison the results of wLMS

signal	**RVM**					**wLMS**	
	n_{iter}	RMS_{rel}	t_{comp}	M	N	RMS_{rel}	t_{comp}
regular	20	33.02 %	32.89 s	150	850	39.67 %	6.86 s
	50	33.12 %	50.05 s	120	850		
	100	33.18 %	65.44 s	120	850		
irregular	20	32.49 %	41.89 s	70	1290	36.49 %	6.88 s
	50	32.48 %	65.66 s	50	1450		
	100	32.45 %	94.79 s	50	1400		
irregular/noisy	20	67.11 %	34.58 s	150	990	69.46 %	6.86 s
	50	64.87 %	46.89 s	240	820		
	100	64.12 %	71.7 s	240	850		

iterations to determine the hyperparameter n_{iter}. The parameters have been evaluated in an interval of $I_M = \{10, 20, ..., 250\}$, $I_N = \{10, 20, ..., 1500\}$ and $I_{n_{iter}} = \{10, 20, 50, 100, 200, 500\}$. The start values of the hyperparameters have been set to zero, after initial results have shown a negligible influence of the choice of individual start values. Only a linear regression function has been considered ($\boldsymbol{\phi}(\mathbf{y}_t)$=$\mathbf{y}_t$).

In Fig. 1.d-f, we can see the influence of n_{iter} on the relative RMS error and the computation time t_{comp} for optimal M and N, respectively. While in all cases the computation times increases for increasing n_{iter} until a maximum, RMS$_{rel}$ is almost independent of n_{iter} for the first two signals. In contrast, the RMS$_{rel}$ decreases by 5.3 % for the irregular and noisy breathing signal. Exemplarily, Tab. 1 shows the optimal M and N for $I_{n_{iter}} = \{20, 50, 100\}$ and a comparison to RMS$_{rel}$ and t_{calc} of wLMS. The mean number of relevance vectors $\bar{n}_{RV}$ are $\bar{n}_{RV} = 10.6$ for the regular breathing signal, $\bar{n}_{RV} = 13.3$ for irregular and $\bar{n}_{RV} = 17.8$ for irregular and noisy.

Based on these results, a second experiment was performed for all 304 motion traces of the data set and we compared the results to the wLMS algorithm. Beside a linear function $\boldsymbol{\phi_{lin}}$, two nonlinear functions have been investigated, to incorporate quadratic and cubic relationships between $\mathbf{y}_t$ and $y_{t+\delta}$. The linear and nonlinear functions are defined as:

$$\boldsymbol{\phi_{lin}}(\mathbf{y}_t) = \mathbf{y}_t, \quad \boldsymbol{\phi_{quad}}(\mathbf{y}_t) = [\mathbf{y}_t,\, \mathbf{y}_t^2]^T, \quad \boldsymbol{\phi_{cub}}(\mathbf{y}_t) = [\mathbf{y}_t,\, \mathbf{y}_t^2,\, \mathbf{y}_t^3]^T \qquad (10)$$

Here, $\mathbf{y}_t^2$ and $\mathbf{y}_t^3$ denotes the second and third power of each vector element $\mathbf{y}_t$. In contrast to $\boldsymbol{\phi_{lin}}$, the number of weights $\mathbf{w}$ are doubled or tripled in case of $\boldsymbol{\phi_{quad}}$ or $\boldsymbol{\phi_{cub}}$, consequently increasing the number of hyperparameters and the computation time. The parameters have been fixed to $M = 100$, $N = 1000$ and $n_{iter} = 100$. The results are illustrated in Fig. 2.a, where the cumulative relative RMS error was plotted in a histogram for the entire data set. The histogram shows the percentage of data for which the relative RMS error was equal or

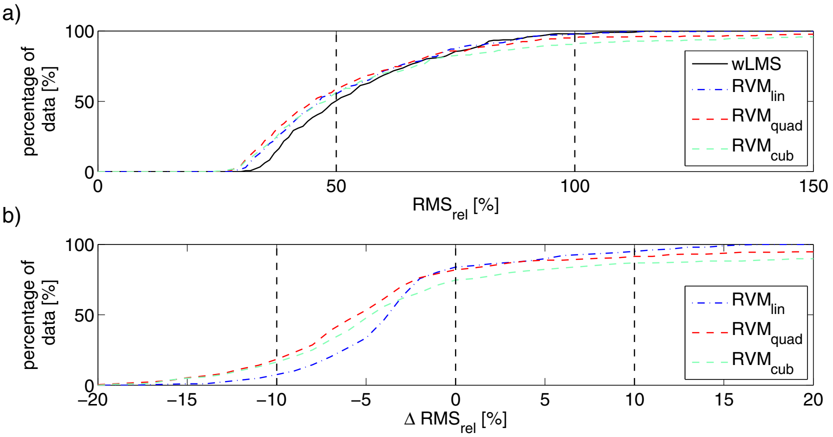

Fig. 2. a) Cumulative relative RMS error histogram for 304 motion traces evaluated with wLMS, RVM$_{lin}$, RVM$_{quad}$ and RVM$_{cub}$; b) Cumulative histogram for the relative RMS error difference between RVM$_{lin}$, RVM$_{quad}$ and RVM$_{cub}$ and wLMS for 304 motion traces

smaller than a certain value. For example the percentage of the motion traces, for which a relative RMS error of 50 % or less was achieved (marked by the left vertical line) is: 50.65 % for wLMS, 55.26 % for RVM$_{lin}$, 56.25 % for RVM$_{cub}$ and 59.21 % for RMV$_{quad}$. These results show that RVM can predict a higher percentage of the data with a relative RMS error of 50 % and less compared to wLMS. Analyzing the results further, Fig. 2.b shows the cumulative relative RMS error difference ΔRMS$_{rel}$, defined as the difference of the relative RMS error between RVM based algorithm and wLMS. The auxiliary line at 0 % separates the percentage of data for which RVM based algorithms are better or equally good as wLMS. This is the case for 84.58 % of the data for RVM$_{lin}$, for 81.44 % for RVM$_{cub}$ and for 74.32 % for RVM$_{cub}$.

Figure 3.a shows a fraction of a motion trace, which RVM$_{lin}$ predicted with a 16 % higher relative RMS error compared to wLMS. Due to the probabilistic approach of RVM, each prediction is the mean of a normal distribution. The pre-

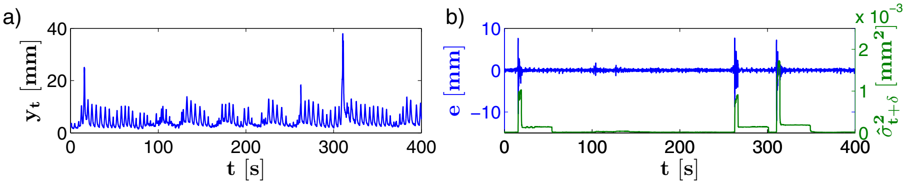

Fig. 3. Amplitude-time plot (a) and prediction error e_t and predicted variance $\hat{\sigma}^2_{t+\delta}$ plot (b) for RVM$_{lin}$ of one motion trace

Table 2. Relative RMS error of wLMS, RVM$_{lin}$, RVM$_{quad}$, RVM$_{cub}$ and hybrid algorithms using both wLMS and RVM with a threshold variance $\sigma^2_{thres} = 10^{-4}\,\text{mm}^2$; the best results for each motion trace are highlighted bold

motion trace	wLMS	RVM$_{lin}$	RVM$_{quad}$	RVM$_{cub}$	wLMS/ RVM$_{lin}$	wLMS/ RVM$_{quad}$	wLMS/ RVM$_{cub}$
1	58.5 %	74.2 %	65.7 %	67 %	67.1 %	60.2 %	**57.7** %
2	60 %	77.9 %	70.3 %	71.6 %	69.7 %	62.7 %	**58.9** %
3	**82.8** %	99.4 %	96.1 %	91.2 %	95.5 %	90.6 %	87.2 %
4	55.2 %	71.2 %	205.3 %	750.4 %	53.6 %	**51.5** %	51.8 %
5	114.4 %	137.6 %	210.5 %	762.3 %	113 %	111.9 %	**111.6** %

dicted variance $\hat{\sigma}^2_{t+\delta}$ (9) can be used as an indicator for the prediction accuracy. Evaluating the prediction error e_t and predicted variance $\hat{\sigma}^2_{t+\delta}$ reveals a strong correlation between these signals (Fig. 3.b). To further increase the prediction accuracy of RVM and reduce outliers, we propose a hybrid algorithm based on a threshold variance σ^2_{thres}. In Tab. 2, the relative RMS errors of wLMS, RVM and threshold based hybrid algorithms of wLMS and RVM are shown for five motion traces. All motion traces have been predicted with a high RMS$_{rel}$ by RVM. The threshold was set to $\sigma^2_{thres} = 10^{-4}\,\text{mm}^2$.

4 Discussion

The optimal N is between 820 and 1400 in the first experiment. Considering $f_s = 26\,\text{Hz}$, the training pairs cover multiple breathing periods. This finding is in line with other motion prediction algorithms as e.g. the wLMS algorithm, which has a learning factor μ to incorporate older training pairs. The optimal M seems to be very high with up to 240 features for an AR process. However, due to the prior distribution over $\mathbf{w}$, only a few weights are nonzero. The mean numbers of relevance vectors $\bar{n}_{RV}$ indicate that with increasing "complexity" of the signal, the algorithm requires more features per training pair to be able to perform an accurate prediction. The evaluation of the number of iterations showed that with increasing n_{iter}, the computation time t_{comp} increases. To be able to perform real time prediction with the currently implemented version, the algorithm would have to be limited to approximately $n_{iter} = 50$.

Evaluating all 304 motion traces showed that, in general, wLMS is outperformed by RVM$_{lin}$. Performing a t-test with $p = 0.001$ on the results confirmed that RVM$_{lin}$ significantly increases the prediction accuracy compared to wLMS. In case of nonlinear RVMs, no significant (RVM$_{quad}$) or a less significant (RVM$_{cub}$ with $p = 0.005$) improvement could be found. This is caused by strong outliers in case of RVM$_{quad}$ and RVM$_{cub}$. As visible in Fig. 2.b, nonlinear RVMs can predict a higher percentage of the data with a relative RMS error difference $\Delta\text{RMS}_{rel} \leq -10\,\%$ than RVM$_{lin}$ (marked by the left vertical line). On the other side, the percentage of the data with $\Delta\text{RMS}_{rel} \geq 10\%$ is lower compared to RVM$_{lin}$ (marked by the right vertical line), concluding that the prediction accuracy is very poor for a subset of the data with nonlinear RVM. However, as indicated in Tab. 2, a hybrid algorithm can effectively increase the prediction accuracy for this subset. Partly, RMS_{rel} of the hybrid algorithm is even lower than RMS_{rel} of wLMS. It has to be further investigated which threshold is optimal and if hybrid algorithms can decrease RMS_{rel} for the complete data set.

It should be pointed out that for these results all algorithm specific parameters have been fixed and no patient specific training was necessary. Our results suggest that a further decrease of RMS_{rel} is possible by e.g. estimating M and N using cross-validation on the first minute of each motion trace. Additionally, so far only linear, quadratic and cubic basis functions have been evaluated. The evaluation could be extended using, e.g., a polynomial or Gaussian kernel.

5 Conclusion

To the authors best knowledge this was the first attempt using a Bayesian inference approach to compensate for respiratory motion in radiation therapy. By using RVM_{lin}, the prediction accuracy was significantly increased. Further improvements are expected due to the flexible framework of RVM, which allows to incorporate nonlinear basis functions and the usage of RVM in a hybrid algorithm approach.

References

1. Giraud, P., De Rycke, Y., Helfre, S., Voican, D., Guo, L., Rosenwald, J.-C., Keraudy, K., Housset, M., Touboul, E., Cosset, J.-M.: Conformal radiotherapy (CRT) planning for lung cancer: analysis of intrathoracic organ motion during extreme phases of breathing. International Journal of Radiation Oncology 51(4), 1081–1092 (2001)
2. Schweikard, A., Glosser, G., Bodduluri, M., Murphy, M.J., Adler, J.R.: Robotic Motion Compensation for Respiratory Movement during Radiosurgery. Journal of Computer Aided Surgery 5(4), 263–277 (2000)
3. Krauss, A., Nill, S., Oelfke, U.: The comparative performance of four respiratory motion predictors for real-time tumour tracking. Physics in Medicine and Biology 56(16), 5303–5317 (2011)
4. Riaz, N., Shanker, P., Wiersma, R., Gudmundsson, O., Weihua, M., Widrow, B., Xing, L.: Predicting respiratory tumor motion with multi-dimensional adaptive filters and support vector regression. Physics in Medicine and Biology 54, 5735–5748 (2009)
5. Ernst, F., Dürichen, R., Schlaefer, A., Schweikard, A.: Evaluating and comparing algorithms for respiratory motion prediction. Physics in Medicine and Biology 58(11), 3911–3929 (2013)
6. Ernst, F., Schlaefer, A., Schweikard, A.: Prediction of Respiratory Motion with Wavelet-Based Multiscale Autoregression. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part II. LNCS, vol. 4792, pp. 668–675. Springer, Heidelberg (2007)
7. Tipping, M.E.: Sparse Bayesian learning and the relevance vector machine. The Journal of Machine Learning Research 1, 211–244 (2001)