International Workshop on Applied Probabilities
8th International Workshop on Applied Probabilities (IWAP2016)

Using Tobit Kalman filtering in order to improve the motion recorded by Microsoft Kinect

K. Loumponias[1,2], N. Vretos[1], P. Daras[1], G. Tsaklidis[2]
[1]Centre for Research and Technology Hellas, Information Technologies Institute, Thessaloniki, Greece.
{loumponias, vretos, daras}@iti.gr
[2]Department of Mathematics, Aristotle University of Thessaloniki, Thessaloniki, Greece.
{kostikasl, tsaklidi}@math.auth.gr

ABSTRACT

In this paper, we analyze data from Microsoft Kinect $v2$ camera using Kalman Tobit and Kalman filters so as to minimize noise. The data concern three-dimensional spatial coordinates recording movements of a persons'A joints, which are subject to measurement errors. The noise variances of the process and the measurements are estimated using the maximum likelihood function. In order to include into the model restrictive conditions based on anthropometric data (e.g. the distances between various joints) we apply the Tobit Kalman Filter. Additionally, restrictions for the joints displacements per fame based on real data can be used in order to get better results. Finally simulations of skeleton before and after using Kalman filtering are presented.

Keywords: Microsoft Kinect Sensor, Human Skeleton Motion, Tobit Kalman Filter.

1. INTRODUCTION

Human skeleton tracking motion is a scientific field, which is studied by commercial RGB-D sensors last years, such as depth sensors (i.e., sensors using the depth coordinate as basic coordinate). The depth sensors are very useful in many applications, such as monitoring of daily activity recognition [1], and health tracking [2]. In this paper it is shown that the Microsoft Kinect $v2$ sensor is able to achieve human skeleton tracking performance in a low-cost manner for activity recognition.

However, Kinect sensor generates a low quality human skeleton tracking due to occlusion, self-occlusion and lack of accuracy in very fast movements. Especially when the human skeleton's joints are occluded, they often appear to be shifted in a no reasonable manner.

The method presented in this paper is based on appropriately smoothing the joints' spatial coordinates. In literature, in order to denoising , they have been

used various common stochastic filters, such as Kalman filter (KF) [3] and no stochastic model, such as Savitzky-Golay filter (SGF) [4] e.t.c.. Although, these filters do not use any restrictions on the noisy measurements. In this paper our aim is to develop a stochastic filter which will not allow the joints to move abnormally, and without affect the real movements. For this reason, we studied the joints' speeds by carrying out various experiments using groundtruth sensors. Then we applied the Tobit Kalman filter (TKF) [5] by taking into account the speeds restrictions.

Microsoft [6] proposed various filters for smoothing human skeleton motion data, but it does not refer how some filter's parameters should be chosen. In [7] a method based on the use of multiple Kinect sensors for human skeleton tracking is proposed. They achieve in determining the reliability of each 3D joint position by employing a data fusion method based on KF using multiple Kinect sensors. They take into account the measurement variance of noise for determining the contribution of an observation to the fused measurement. Additionally, they explain how to estimate the measurement variance for each one of the measurements. Finally, they present the average 3D position error of ten activities produced by their method, by a single Kinect and the average derived by multiple Kinect sensors, respectively. In all but one scenario (daily activities), their method gives better results than the standard KF.

Other scientists who are dealing with activity recognition via neural networks, use a simple SGF in order to smooth the data [8]. This method is based on the previous, as well as the current and the two following observations.

The rest of paper is organized as follows. In Section 2 TKF procedure along with the related likelihood function is provided. In Section 3 TKF approach for human skeleton tracking is established. Finally, in Section 4, conclusions are presented.

2. TOBIT KALMAN FILTER

In this section we describe briefly TKF, which provides a classification scheme for censored models [9]; these classes depend on the type of censoring and include also the cases of censoring [10], that depends on other variables. In the applications cases of censoring, the censored measurement model provides a measurement, either in knowing the exact value or in knowing that the value lies into an interval.

In the general case of scalar measurements, the Tobit model is referred to as the censored regression model determined by,

$$y_k^* = hx_k + v_k,$$
$$y_k = \begin{cases} y_k^*, & T_l < y_k^* < T_u \\ T_l, & y_k^* < T_l \\ T_u, & y_k^* > T_u, \end{cases} \tag{1}$$

where y_k^* is the latent variable, y_k is the censored measurement, h is an arbitrary scalar, T_l, T_u are the lower and upper thresholds-limits respectively and v_k is a Gaussian random variable with mean 0 and variance σ_v^2 . By (1) it is obvious that TKF determines a non-linear process.
In order to face the problem of censored measurements, we propose TKF defined in [5].

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{w}_k,$$
$$\mathbf{y}_k^* = \mathbf{H}\mathbf{x}_k + \mathbf{v}_k,$$

with

$$y_{k,i} = \begin{cases} y_{k,i}^*, & T_{l,i} < y_{k,i}^* < T_{u,i} \\ T_{l,i} & y_{k,i}^* < T_{u,i} \\ T_{u,i}, & y_{k,i}^* > T_{u,i}. \end{cases} \qquad i = 1, 2, ..., m \tag{2}$$

where k stands for the time step and $\mathbf{w}_k$ and $\mathbf{v}_k$ are random vector variables following $N(\mathbf{0}, \mathbf{Q}_k)$ and $N(\mathbf{0}, \mathbf{R}_k)$, respectively, where $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\mathbf{A}$ and $\mathbf{H}$ are the transition and the observation matrices, respectively, while $\mathbf{y}_k = (y_{k,i})_{i=1}^m, \mathbf{y}_k^* = (y_{k,i}^*)_{i=1}^m$ are the censored observations (measurements) and the latent observations, respectively and m designates the dimensionality of the process.
The probability function of the i^{th} component of the measurement given the state vector is

$$\begin{aligned} f(y_{k,i}|x_{k,i}) = \frac{1}{r_i}\phi\left(\frac{y_{k,i} - h_i x_{k,i}}{r_i}\right) u(y_{k,i} - T_{l,i}) u(T_{max,i} - y_{k,i}) \\ + \Phi\left(\frac{T_{l,i} - h_i x_{k,i}}{r_i}\right)\delta(T_{l,i} - y_{k,i}) \\ + \left(1 - \Phi\left(\frac{T_{max,i} - h_i x_{k,i}}{r_i}\right)\right)\delta(T_{max,i} - y_{k,i}), \end{aligned} \tag{3}$$

where ϕ and Φ are the probability and cumulative distribution function of standard normal distribution respectively, δ stands for the Kronecker delta function and u for the Heavyside function.
Next, we denote by $\mathbf{P}_{un,k}$, $\mathbf{P}_{l,k}$, $\mathbf{P}_{u,k}$ the probabilities of a measurement to be uncensored, or censored from below or censored from above, respectively, at time k. Then by (3) it is derived that

$$\mathbf{P}_{un,k} = diag \begin{bmatrix} \Phi\left(\frac{T_{u,1} - h_1 \hat{x}_{k,1}^-}{r_{k,1}}\right) - \Phi\left(\frac{T_{l,1} - h_1 \hat{x}_{k,1}^-}{r_{k,1}}\right) \\ ... \\ \Phi\left(\frac{T_{u,m} - h_m \hat{x}_{k,m}^-}{r_{k,m}}\right) - \Phi\left(\frac{T_{l,m} - h_m \hat{x}_{k,m}^-}{r_{k,m}}\right) \end{bmatrix}, \tag{4}$$

$$\mathbf{P}_{l,k} = diag\begin{bmatrix} \Phi\left(\frac{T_{l,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\right) \\ ... \\ \Phi\left(\frac{T_{l,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\right) \end{bmatrix}, \tag{5}$$

$$\mathbf{P}_{u,k} = diag\begin{bmatrix} 1-\Phi\left(\frac{T_{u,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\right) \\ ... \\ 1-\Phi\left(\frac{T_{u,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\right) \end{bmatrix}. \tag{6}$$

By taking into account the above matrices, the expected value of the measurement when censored and uncensored measurements are included given the a priori estimation of the state vector has the form:

$$\mathbf{E}(\mathbf{y}_k) = \mathbf{P}_{un,k}(\mathbf{H}\hat{\mathbf{x}}_k^- + \mathbf{R}^{\frac{1}{2}}l_k) + \mathbf{P}_{l,k}\mathbf{T}_l + \mathbf{P}_{u,k}\mathbf{T}_u \tag{7}$$

where $\mathbf{T}_u = (T_{u,i})_{i=1,..,m}, \mathbf{T}_l = (T_{l,i})_{i=1,..,m}$ and the parameter l_k at time k is the inverse Mill ratio [11],

$$l_k = \mathbf{P}_{un,k}^{-1}\begin{bmatrix} \phi\left(\frac{T_{u,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\right) - \phi\left(\frac{T_{l,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\right) \\ ... \\ \phi\left(\frac{T_{u,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\right) - \phi\left(\frac{T_{l,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\right) \end{bmatrix}.$$

The covariance matrix of the measurement is given by

$$\mathbf{R}_k^* = \mathbf{R}_k\Big(\mathbf{I} + \mathbf{P}_{un}^{-1}diag(c_k) - diag(l_k)^2\Big) \tag{8}$$

where the parameter c_k [11] is given by

$$c_k = diag\begin{bmatrix} \frac{T_{l,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\phi\left(\frac{T_{l,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\right) - \frac{T_{u,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\phi\left(\frac{T_{u,1}-h_1\hat{x}_{k,1}^-}{r_{k,1}}\right) \\ ... \\ \frac{T_{l,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\phi\left(\frac{T_{l,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\right) - \frac{T_{u,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\phi\left(\frac{T_{u,m}-h_m\hat{x}_{k,m}^-}{r_{k,m}}\right) \end{bmatrix}.$$

TKF process is defined as [12], [9]:

The Predict Stage:

$$\hat{\mathbf{x}}_k^- = \mathbf{A}\hat{\mathbf{x}}_{k-1}, \tag{9}$$

$$\mathbf{P}_k^- = \mathbf{A}\mathbf{P}_{k-1}\mathbf{A}^T + \mathbf{Q}_k. \tag{10}$$

The Update Stage:

$$\mathbf{R}_1 = \mathbf{P}_k^-\mathbf{H}^T\mathbf{P}_{un,k},$$

$$\mathbf{R}_2 = \mathbf{P}_{un,k}\mathbf{H}\mathbf{P}_k^-\mathbf{H}^T\mathbf{P}_{un,k} + \mathbf{R}_k^*,$$

$$\mathbf{K}_k = \mathbf{R}_1\mathbf{R}_2^{-1}, \tag{11}$$

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{K}_k(\mathbf{y}_k - \mathbf{E}(\mathbf{y}_k|\hat{\mathbf{x}}_k^-)), \tag{12}$$

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{P}_{un,k}\mathbf{H})\mathbf{P}_k^-. \tag{13}$$

The likelihood function for the i^{th} component of the censored measurement $\mathbf{y}_k$, is:

$$L_i(y_{1,i},...,y_{n,i}) = \prod_{y_{k,i}=T^i_{below}} \Phi(Tl_k^i) \prod_{y_{k,i}=T^i_{above}} (1-\Phi(Th_k^i))$$

$$\times \prod_{T^i_{below}< y_{k,i} < T^i_{above}} \frac{1}{\sqrt{h_i^2P_{k,i}^- + r_i^2}}\phi\left(\frac{y_{k,i} - h_i\hat{x}_{k,i}^-}{\sqrt{h_i^2P_{k,i}^- + r_i^2}}\right), \tag{14}$$

3. IMPLEMENTATIONS AND EXPERIMENTS

In this paper we use the Microsoft Kinect v2 sensor to record 3D point sequences of a human skeleton (in motion). In human skeleton motion tracking, the body is represented by a number of joints (25 in total), corresponding to different body parts such as head, neck, shoulders, etc. Each joint is represented by the vector of its Euclidean 3D space coordinates and our aim is to denoise the measurements for every joint in order to improve the representation of human movements. Thus, we denoise each one of the joints' coordinates separately; the input is the vector of the joints' coordinates, $\mathbf{y}_k^* = [y_{k,1}^*, y_{k,2}^*, y_{k,3}^*]$ and the output is the vector of the denoised states coordinates, $\mathbf{x}_k = [x_{k,1}, x_{k,2}, x_{k,3}]$. Thus, we define the initial observation and the transition matrices to be equal to the identity matrix.

Next, we have to estimate the covariance matrix for the process noise, $\mathbf{Q}_k$ at time step k. Firstly we assume that the covariance matrix of the measurement noise, $\mathbf{R}_k$, is constant and its entries are of the order $0.01\,m^2$. We chose to initialize the matrix $\mathbf{R}_k$ in that way under the assumption that the Kinect sensor exhibits significant errors on each axes. We have conducted various experiments by Kinect, showing that even if an individual is at rest and in front of the Kinect, the RMSE in the displacement estimation between measurement and groundtruth data is almost 0.02m while in the case where the human skeleton is occluded the RMSE is bigger (0.03m - 0.20m), thus a variance of 0.01 m^2 seems a good choice,

$$\mathbf{R} = 0.01\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}. \tag{15}$$

In order to correct the noise, we studied many recordings by state of the art, Vicon; we observed that the velocity of spatial coordinates x and z did not exceed

31 cm per two consecutive frames for every joint, while the coordinate y did not exceed 18 cm respectively. Thus we took these restrictions into account, in order to correct the data. So we constructed a TKF with limits $\mathbf{T}_l$ and $\mathbf{T}_u$ for the spatial coordinates $[x, y, z]$ as follows,

$$\mathbf{T}_{u,k} = (\hat{x}_{k-1} + 0.31, \hat{y}_{k-1} + 0.18, \hat{z}_{k-1} + 0.31),$$

$$\mathbf{T}_{l,k} = (\hat{x}_{k-1} - 0.31, \hat{y}_{k-1} - 0.18, \hat{z}_{k-1} - 0.31),$$

where $\mathbf{T}_{u,k}$ and $\mathbf{T}_{l,k}$ are the limits of TKF at time k which depend on the previous estimation of spatial coordinates. Thus, for the measurement $\mathbf{y}_k = [x_k, y_k, z_k]$ at time k we get

$$y_{k,i} = \begin{cases} y^*_{k,i}, & T^i_{l,k} < y^*_{k,i} < T^i_{u,k} \\ T^i_{l,k}, & y^*_{k,i} < T^i_{l,k} \\ T^i_{u,k}, & y^*_{k,i} > T^i_{u,k}. \end{cases} \qquad i = 1, 2, 3$$

The aforementioned TKF model can appropriately smooth big aberrant movements due to Kinect's errors. Apparently, if $T^i_{l,k} \to -\infty$ and $T^i_{u,k} \to \infty$ (i.e., the range of TKF's state values becomes too big) TKF becomes the standard KF.
Now, in order to create a general model for de-noising Kinect's measurements, (in which we will not estimate the matrix $\mathbf{Q}_k$ for every time-window, because this is time consuming) we can assume that $\mathbf{Q}_k$ is constant. Then, by the likelihood function (14), the entries of the matrix $\mathbf{Q}$ can be derived. Interestingly we noticed by various joints' movements, that the entries of $\mathbf{Q}$ appeared (were estimated) to be smaller than those of matrix $\mathbf{R}$, and generally they depend on the accuracy of the Kinect v2 sensor and the joints' speed. Concerning slow motions or the human skeleton at rest, the values are experimentally found to be smaller than $10^{-4}\, m^2$ and for faster motions they lie between $10^{-3}\, m^2$ and $10^{-2}\, m^2$. We have to notice that in some cases where the entries of $\mathbf{Q}$ were found to be quite large $(10^{-2}\, m^2)$, the human skeleton moved too quickly in an abnormal manner due to occlusions and self-occlusions. Thus, we assume that the covariance matrix of the noise process is

$$\mathbf{Q} = 0.001 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{16}$$

otherwise, if we assume smaller or bigger values, TKF will over-smooth or it will not denoise the Kinect's measurements, respectively. So, the above assumption seems to provide a good approximation in order to smooth the Kinect's v2 sensor measurements.
Beyond the above method (TKF), we use the standard KF, where the covariance matrix, $\mathbf{R}$, is defined as in (15), and the covariance matrix for the process noise, $\mathbf{Q}$, can be estimated by the likelihood function give in [13]. The results in human motion data capturing by Kinect showed that the entries of $\mathbf{Q}$ are almost the

same as in TKF, thus we assume that the matrix $\mathbf{Q}$ in KF is defined as in (16). In our first experiment, we evaluate our proposed method with respect to ground truth data. Thus, we monitor a man throwing a ball with his right hand, and we record this motion by a Kinect and the Vicon system at the same time. The number of Kinect's and Vicon's frames are 266 and 139, respectively. We note that Kinect time-stamp is almost 0.033 sec per frame while Vicon time-stamp is constantly 0.032sec. We interpolate Vicon data in order to deal with the time-stamp problem; after interpolation, the new Vicon data include 133 frames. Therefore, we temporally synchronize the two sensors to start together. To do so, we initially calculate the angles of knees and elbows obtained by Kinect and Vicon data and then, we calculate the RMSE between these angles for different delays. The results show that the minimum values of RMSE for every angle appeared for delays of 92-95 frames. The different delays between the angles in some cases are somewhat expected because Kinect records fast movements with delay. KF smooths the spatial coordinates without affecting the movement (see Fig. 1). TKF perform exactly the same smoothing in all joints as KF, while SGF does not perform a satisfactory smoothing in some points where the measurements have a significant error.

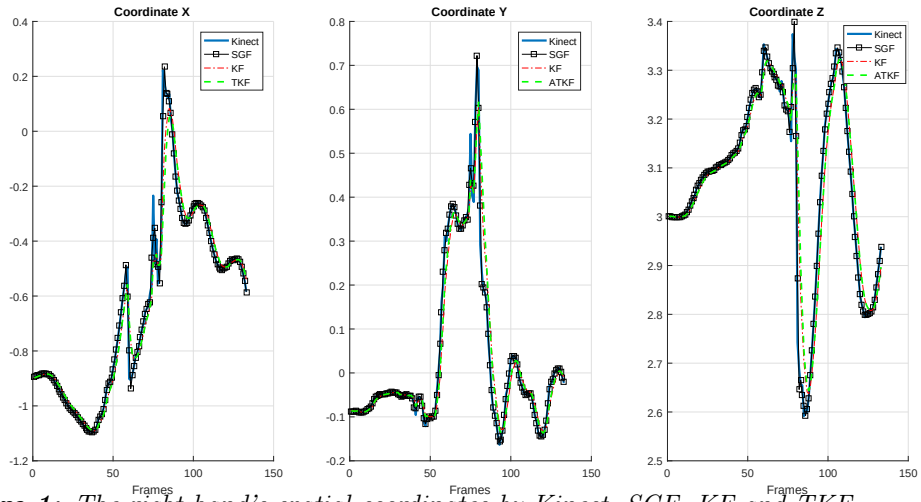

Figure 1: *The right hand's spatial coordinates by Kinect, SGF, KF and TKF.*

In Table 1 we observe the RMSEs for the angles as they arise for delays $t = 92, 93, 94, 95$ frames, respectively. In all cases, the RMSEs are big enough because of the occlusion of some joints during the recording. However, as can be seen, we get lower RMSEs in all cases via TKF.

In some others experiments, we record various human skeleton motions by a Kinect v2 sensor. In some of the recordings, the human skeleton seems to "fall

down" for one or two frames due to occlusions. Thus, we apply SGF, KF and TKF in order to correct this error. As can be seen in Fig. 2, the head's spatial coordinate y (of human skeleton) resulted by TKF, does not follow the error produced by the Kinect device. In constraint with TKF, SGF and KF follow the "fall" for 75cm and 50 cm respectively.

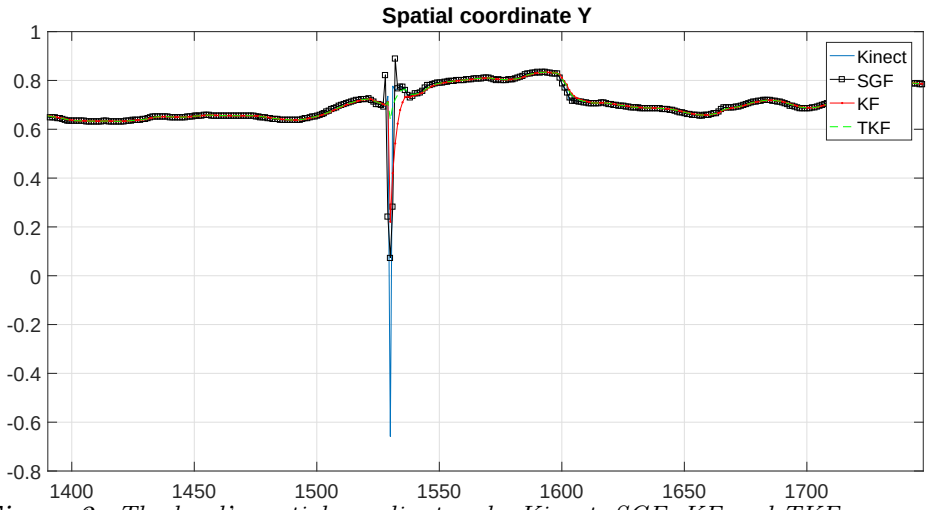

Figure 2: *The head's spatial coordinate y by Kinect, SGF, KF and TKF.*

4. CONCLUSION

The aim of this paper is to improve human skeleton tracking, using a Kinect v2 sensor, which generates error in recordings due to occlusion, self-occlusion e.t.c.. Thus, we propose to use TKF for human skeleton motion tracking in real time. In this approach we defined the limits $\mathbf{T}_{u,k}$ and $\mathbf{T}_{l,k}$ in a reasonable manner for every time k. For that purpose we considered human skeleton motion data, with various joints' movements, which were obtained by means of the groundtruth sensor, Vicon.

The covariance matrix of the noise process $\mathbf{Q}$, using TKF procedure was estimated via maximum likelihood estimation. Between the three filters, i.e., SGF, the standard KF and TKF, the last one was more accurate performing a better human skeleton tracking. Furthermore, in some frames when the human skeleton seemed to "fall down" due to occlusion, the method of proposed TKF, corrected better the error in recordings than the standard KF and SGF.

Acknowledgements: This work was supported by the European Project (Horizon2020) ICT4Life under the GA 690090.

References

[1] Chunguang Zhang and Lifang Zhang. Activity recognition in smart homes based on second-order hidden markov model. *International Journal of Smart Home*, 7(6):237–244, 2013.

[2] Brook Galna, Gillian Barry, Dan Jackson, Dadirayi Mhiripiri, Patrick Olivier, and Lynn Rochester. Accuracy of the microsoft kinect sensor for measuring movement in people with parkinson's disease. *Gait & posture*, 39(4):1062–1068, 2014.

[3] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[4] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[5] Bethany Allik. *The tobit kalman filter: An estimator for censored data*. PhD thesis, University of Delaware, 2014.

[6] Microsoft Kinect. Skeletal joint smoothing white paper, 2016.

[7] Sungphill Moon, Youngbin Park, Dong Wook Ko, and Il Hong Suh. Multiple kinect sensor fusion for human skeleton tracking using kalman filtering. *International Journal of Advanced Robotic Systems*, 13, 2016.

[8] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1110–1118, 2015.

[9] Bethany Allik, Cory Miller, Michael J Piovoso, and Ryan Zurakowski. The tobit kalman filter: An estimator for censored measurements. *IEEE Transactions on Control Systems Technology*, 24(1):365–371, 2016.

[10] James Tobin. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36, 1958.

[11] Donald R Barr and E Todd Sherrill. Mean and variance of truncated normal distributions. *The American Statistician*, 53(4):357–361, 1999.

[12] Bethany Allik, Cory Miller, Michael J Piovoso, and Ryan Zurakowski. Estimation of saturated data using the tobit kalman filter. In *2014 American Control Conference*, pages 4151–4156. IEEE, 2014.

[13] James Douglas Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.

Angles	Kinect	SGF	KF	TKF
Right Elbow	39.31	37.44	35.57	**35.02**
Left Elbow	31.58	30.65	25.92	**25.10**
Right Knee	16.70	16.79	15.90	**14.75**
Left Knee	26.25	25.81	24.80	**25.00**

Angles	Kinect	SGF	KF	TKF
Right Elbow	38.76	36.86	34.94	**34.37**
Left Elbow	32.18	31.27	26.27	**25.34**
Right Knee	17.03	17.12	15.74	**14.57**
Left Knee	26.38	26.01	24.36	**24.46**

Angles	Kinect	SGF	KF	TKF
Right Elbow	38.43	36.63	34.45	**33.85**
Left Elbow	32.99	32.09	26.27	**25.88**
Right Knee	17.77	17.79	15.86	**14.74**
Left Knee	26.67	26.46	24.19	**24.16**

Angles	Kinect	SGF	KF	TKF
Right Elbow	38.39	36.64	34.25	**33.66**
Left Elbow	33.96	33.06	27.61	**25.70**
Right Knee	18.78	18.78	16.21	**15.19**
Left Knee	27.14	27.02	24.27	**24.12**

Table 1: *RMSEs for the angles by Kinect , SGF, KF and TKF for time delay 92, 93, 94 and 95.*