REPRESENTATION AND RECOGNITION OF HUMAN ACTIONS
A New Approach based on an Optimal Control Motor Model

Sumitra Ganesh and Ruzena Bajcsy
University of California, Berkeley, California, USA

Keywords: Human Activity Recognition, Optimal Control, Multiple Model Estimation, Particle Filters.

Abstract: We present a novel approach to the problem of representation and recognition of human actions, that uses an optimal control based model to connect the high-level goals of a human subject to the low-level movement trajectories captured by a computer vision system. These models quantify the high-level goals as a performance criterion or cost function which the human sensorimotor system optimizes by picking the control strategy that achieves the best possible performance. We show that the human body can be modeled as a hybrid linear system that can operate in one of several possible modes, where each mode corresponds to a particular high-level goal or cost function. The problem of action recognition, then is to infer the current mode of the system from observations of the movement trajectory. We demonstrate our approach on 3D visual data of human arm motion.

1 INTRODUCTION

The first fundamental problem in the analysis of human motion is that of representation. Several models of human motion have been proposed in literature. In (Bregler and Malik, 1997), linear dynamical systems are used to model simple motions and high-level complex motions are modeled using Hidden Markov Models, where each state corresponds to a particular dynamical system. Layered structures of Hidden Markov Models (Oliver et al., 2004)) and hierarchical Bayesian Networks ((Park and Aggarwal, 2004)) have been used to model multiple-levels of abstraction. The other broad approach has been to extract 3D spatio-temporal features or templates of movements using Principal Component Analysis (PCA) (Safonova et al., 2004), non-linear dimensionality reduction techniques (Fod et al., 2002) and other methods (Weinland et al., 2006).

In this paper, we propose a new approach to the problem of representation based on an optimal control model for human motion. The challenge lies in finding a mathematical model that can connect the high-level goals and intentions of a human subject to the low-level movement details captured by a computer vision system. Optimal control models of the human sensorimotor system do this in a natural manner. These models quantify the high-level goals as a performance criterion or cost function which the human sensorimotor system optimizes by picking the control strategy that achieves the best possible performance. Thus optimal control models of human motion place the high-level goals and the control strategy center stage, while the movement details arise naturally as a consequence of these goals. The different cost functions (corresponding to different simple goal-directed tasks), or equivalently the corresponding optimal control modules, are the basic building blocks in our representation. We view the human motor system as a hybrid system that switches between different control modules, in response to changing high-level goals. The problem of action recognition is to infer the hidden goal of the motion from observations of the movement trajectory. More complex actions can be modeled as a composition of these basic goals. For example, the action of lifting an object might be accomplished by the composition of two goal-directed motions - reaching for the object and then lifting it.

Optimal control models have been used in robotics and computer animation ((Nori and Frezza, 2005), (Li and Todorov, 2004)) for synthesis of motion, and in the field of computational neuroscience as a model for the human motor system. Optimal control models of the human sensorimotor system ((Todorov, 2004), (Harris and Wolpert, 1998), (Scott, 2004)) have been successful in explaining several empirical observations about human motion. Thus our model is both theoretically justified and physically meaning-

Ganesh S. and Bajcsy R. (2008).
REPRESENTATION AND RECOGNITION OF HUMAN ACTIONS - A New Approach based on an Optimal Control Motor Model.
In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pages 99-104
DOI: 10.5220/0001086700990104

ful. However, to the best of our knowledge, such a model has not been used to analyze human motion or recognize the higher-level goals of human motion.

Our approach is similar to (Bregler and Malik, 1997), in that we build our model using dynamical building blocks or *primitives*, rather than purely kinematic ones (e.g. (Fod et al., 2002)). But while the dynamical primitives in (Bregler and Malik, 1997) do not model the forces or control input involved in producing the motion, the control strategy plays a central role in our representation. In (Del Vecchio et al., 2003) Del Vecchio et al. used a switching linear dynamical system with simple control to study the 2D motion of a computer mouse being used to draw figures. However, our control model is richer and we test our hypotheses on 3D motions of the human arm.

The paper is organized as follows. In section 2 we describe our model of the human motor system, with particular emphasis on the control model. In section 3, we define the estimation problem and refer to the methods we use. In section 4, we describe the experimental setup and present results of the mode recognition on human arm data. We conclude by indicating directions of future work and possible applications of our work.

2 MODELS

In our model, the human motor system can be viewed as a hybrid system that switches between different control modules or modes as defined by the different cost functions (goals). We assume here that the cost functions corresponding to the different goals are known to us. Given noisy kinematic observations of a motion (e.g. the 3D hand trajectory for an arm motion), we wish to estimate the underlying mode sequence, or in other words, the sequence of basic underlying goals that motivated the motion. To define the problem more precisely we need to define models for the following :

1. Mode Evolution : Since each mode corresponds to a different goal, this model describes the probability of switching from one goal to another.
2. State Evolution : A biomechanical model is needed to define how the control and the current configuration/state of the body (joint angles and velocities) determine the body configuration at the next time instant.
3. Control : This model defines how the current state and mode are used to arrive at a control input for the biomechanical model.
4. Observation : The observation model defines the relation between the joint angles and the observations.

We define these mathematical models and then present the methods used for simultaneous state and mode estimation from observations.

2.1 State Evolution Model

Let $q(t)$ and $\tau(t)$ be $n \times 1$ vectors (for n degrees of freedom) that denote the joint angles and torques at time t, respectively. The human body can be approximately modeled as a structure of rigid links connected by joints. The equations of motion for such a model of the human body are of the form (Murray et al., 1994)

$$M(q)\ddot{q} + C(q,\dot{q})\dot{q} + N(q) = \tau \tag{1}$$

where the matrices $M(.), C(.)$ and $N(.)$ represent the configuration dependent inertia, coriolis and gravitational terms. Since $M(q)$ is always positive definite, this system can be feedback-linearized (Murray et al., 1994) by designing the control torque to be of the form

$$\tau = M(q)u + C(q,\dot{q})\dot{q} + N(q) \tag{2}$$

where u is a control sequence. The *equivalent* linearized system, from (1) and (2), is $\ddot{q} = u$. By *equivalent* what we mean is that optimal control methods can be used to determine the form of the control u required for the $\ddot{q} = u$ system to achieve the goal. This control u can then be transformed using (2) to obtain the torques that need to be applied to the nonlinear system in (1). This allows us to focus on the feedback linearized system as far as the mode estimation is concerned. Also note that the linearized system is independent of body parameters such as mass distribution and length, which vary from subject to subject.

For a sampling period of Δ, the time-discretized linear system of interest for the optimal control problem is

$$x_{k+1} = Ax_k + Bu_k + v_k \tag{3}$$

where x_k, the state at time $k\Delta$, is a $2n \times 1$ vector of the joint angles and velocities at that time instant, v_k is the process noise, u_k is the control and

$$A = \begin{bmatrix} \mathrm{I}_n & \Delta \mathrm{I}_n \\ 0_n & \Delta \mathrm{I}_n \end{bmatrix} \qquad B = \begin{bmatrix} \frac{\Delta^2}{2}\mathrm{I}_n \\ \Delta \mathrm{I}_n \end{bmatrix} \tag{4}$$

2.2 Optimal Control

In this section we describe how the optimal control law is determined for a general parametrized class of

cost functions, when they are optimized for the linearized state evolution model described in the previous section. The cost function we consider is of the form

$$\underbrace{\sum_{k=1}^{T}(Cx_k - r_k)^{\mathrm{T}} Q_k (Cx_k - r_k)}_{\text{accuracy}} + \underbrace{\sum_{k=1}^{T-1} u_k^{\mathrm{T}} R u_k}_{\text{energy}} + \underbrace{\rho T}_{\text{time}} \quad (5)$$

for a motion of duration T sampling instants. The parameters $C, \{r_1, ...r_T\}, Q_k \geq 0$ can be used to specify the goals or constraints, while the parameter $R > 0$ can be used to specify the penalty on energy consumption. The $(Cx_k - r_k)^T Q_k (Cx_k - r_k)$ term constrains a linear function of the state to be close to a reference value r_k. This term could be, for example, used to impose the goal of reaching a certain configuration or maintaining a certain pose by constraining the velocities to be close to zero. In the most extreme case, we can specify an exact trajectory to be followed for the entire duration. The last term imposes a penalty on the duration of the motion. Thus, in minimizing the cost function we attempt to achieve the goal with minimum error, in the minimum time, while consuming the least energy. The exact tradeoff between these conflicting demands is determined by the cost function parameters Q_k, R and ρ.

If the cost function only contained the first two terms (accuracy and energy), the resulting optimal control problem is called a Linear Quadratic (LQ) problem. In the LQ problem the duration of the motion is fixed. The interesting thing about the solution to this problem (see (Lewis and Syrmos, 1995) for details) is that not only is the form of the control known, but the optimal cost-to-go, i.e. the minimum total cost incurred from any time k until the fixed final time T, can be computed as a function of the current state x_k and the system parameters. Thus, given the current state we can compute the minimum cost-to-go for different values of the final time T. Denoting the first two terms of (5) as $J_{\mathrm{LQ}}()$, and the optimal cost-to-go function as $V()$, the minimum value of the cost function in (5) can be written as

$$\begin{aligned} J^* &= \min_{x_{2:T}, u_{1:T-1}, T} J_{\mathrm{LQ}}(x_{2:T}, u_{1:T-1}) + \rho T \quad (6) \\ &= \min_{T} \rho T + \min_{x_{2:T}, u_{1:T-1}} J_{\mathrm{LQ}}(x_{2:T}, u_{1:T-1}) \\ &= \min_{T} \rho T + V(x_1, T) \ . \end{aligned}$$

The optimal time T^* can be obtained as $T^* = \arg\min_T \rho T + V(x_1, T)$. The problem of minimizing (5) then reduces to the LQ problem of minimizing the first two terms for a fixed final time $T = T^*$.

Standard results from optimal control theory (Lewis and Syrmos, 1995), can then be used to determine the form of the control u_k.

$$u_k = -K_k^{\mathrm{fb}} x_k + K_k^{\mathrm{ff}} z_{k+1} \quad (7)$$

where the feedback and feedforward gain matrices, K_k^{fb} and K_k^{ff} respectively, and the auxiliary sequence z_{k+1} are determined by a backward recursion that is independent of the state sequence and only depends on the system and cost function parameters (See (Lewis and Syrmos, 1995) for details).

To summarize, the parameters of state evolution model (3) and the cost function (5) completely determine the form of the optimal control u_k as a function of the current state x_k. Thus, given the current mode and state, the control input to be applied to optimize the cost function corresponding to that mode (goal) can be determined.

2.3 Mode Evolution and Observation Models

The modes can be modeled as states of a markov chain. Let m_k, an integer value drawn from the set $\{1,..,N\}$, represent the current mode i.e. which of the N possible control modes is effective in determining the state transition from the $(k-1)$-th sampling instant to the k-th instant. The prior distribution of modes is given by $\pi_i = P(m_0 = i),\ i = 1,\ldots,N$ and the transition probabilities are given by

$$H_{ij} = P(m_k = j | m_{k-1} = i) \quad i, j = 1,\ldots,N \ . \quad (8)$$

The values of π_i and H_{ij} are assumed to be known.

The observations $\{y_1, ..., y_T\}$ available to us are of the form

$$y_k = g(x_k) + n_k \ , \quad (9)$$

where the function $g(.)$ and the distribution of the observation noise n_k are assumed to be known.

3 STATE AND MODE ESTIMATION

The recognition problem can be stated as follows : given observations $\{y_{1:T}\} \triangleq \{y_1, ..., y_T\}$ estimate the state (joint angles and velocities) trajectory $\{x_{1:T}\}$, the control sequence $\{u_{1:T-1}\}$ and the task or mode trajectory $\{m_{2:T}\}$. The problem requires simultaneous estimation of the continuous state and the mode of the system. The control sequence is not an independent sequence - it is determined by the state and the mode. Thus defined, the problem of action recognition is one of mode estimation in a hybrid system. Similar problems have been addressed in the tracking of a maneouvering targets (McGinnity and

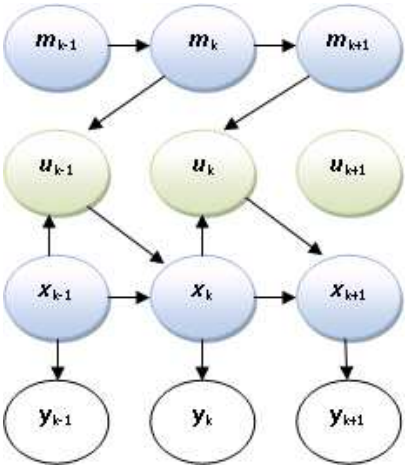

Figure 1: Graphical Model of State Evolution, Mode Evolution, Control and Observations. The control is an intermediate hidden variable that is completely specified by the mode and the current state. Thus it is sufficient to estimate the continuous state x_k and the mode m_k (hidden variables) from the observations y_k.

Irwin, 2000) and fault detection (de Freitas, 2002) in systems. The Interacting Multiple Model (IMM) algorithm (Blom and Bar-Shalom, 1988) and its variants (McGinnity and Irwin, 2000) have been the preferred method of solving this problem. We use a bootstrap method similar to that in (McGinnity and Irwin, 2000), with an auxiliary particle filter (Pitt and Shephard, 2001).

4 EXPERIMENTS AND RESULTS

We tested our ideas on 3D motion data sampled at 7 frames/sec, collected from a setup of 12 camera clusters. We used the algorithm proposed by Lien et al. in (Lien et al., 2007) for segmentation and tracking of the joints of the body. Our test motions consisted of motions of the arm for 2 subjects. The subjects were instructed in two tasks that involved lifting a 5 lb weight. The tasks are shown in figure 4. Each task consists of two goals, lifting and lowering, but the manner in which these are to be accomplished is different in the two tasks. Thus there are four distinct goals or modes in the data set as shown in figure 4. Five repetitions of each task were recorded for each subject were used for testing.

For estimating the mode and state, we only use the 3D position trajectory of the hand with respect to the shoulder as observations. The model of the arm and joint angles are shown in figure 2, and the reference coordinate system is shown in figure 3. The joint angle values are specified with respect to the reference

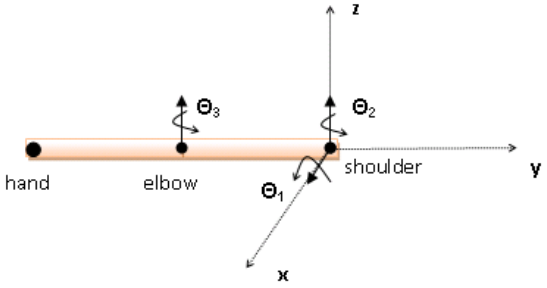

Figure 2: Model of the arm.

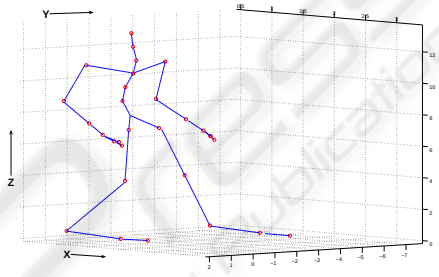

Figure 3: Coordinate System.

pose in figure 2, which corresponds to $\theta_1 = \theta_2 = \theta_3 = 0$ deg. The state of the system consists of the joint angles and velocities i.e. $x = [\theta_1\theta_2\theta_3\dot{\theta}_1\dot{\theta}_2\dot{\theta}_3]^T$. The observation function, that relates the joint angles to the observations of the hand position, is given by

$$g(x) = \begin{bmatrix} L_1\sin(\theta_2) + L_2\sin(\theta_2+\theta_3) \\ -L_1\cos(\theta_1)\cos(\theta_2) - L_2\cos(\theta_1)\cos(\theta_2+\theta_3) \\ -L_1\sin(\theta_1)\cos(\theta_2) - L_2\sin(\theta_1)\sin(\theta_2+\theta_3) \end{bmatrix} \quad (10)$$

where L_1 and L_2 denote the length of the upper and lower arm, which are obtained from the segmentation and tracking algorithm (Lien et al., 2007).

The cost functions for the four modes are constructed as follows. For modes 1 and 2, target poses (r_T) are specified in terms of constraints on θ_3 ($\theta_3 =$ 150 deg for mode 1 and 90 deg for mode 2) , the rotation of the elbow joint. For modes 3 and 4, the target poses are in terms of θ_1 ($\theta_1 = 0$ deg for mode 3 and 90 deg for mode 4), the rotation of the shoulder joint about the x axis. In all cases the final velocities are constrained to be zero.

While the constraints arise naturally from our experiment design and task specification, determining the relative weighting of accuracy, energy and time (Q, R, ρ) is not that simple. In this experiment we determine these parameters by comparing simulations with a training data set. We fixed $Q_T = 10^3 I_4$, and $R = 10^{-3} I_3$ for all modes. We found the value of ρ to

be quite different for the two subjects - 100 for subject 1 and 25 for subject 2. This indicates that while different subjects might have a common understanding of the task definition, they might have different preferences when it comes to the relative weighting of accuracy, energy and time. These varying internal preferences might explain the stylistic variations observed among subjects performing the same task. This matter requires further study. In our experiment we use different ρ values for the different subjects during mode estimation.

Since the data set is fairly small, we set all the modes to be equally likely apriori. The average time spent in any mode τ, as observed in the training data set, was used to set the transition probabilities as $H_{ii} = 1-(1/\tau)$, $i = 1,\ldots N$ and $H_{ij} = (1-H_{ii})/(N-1) \forall i \neq j$. The value of τ was fixed at 20 (sampling instants) for the results below, but the estimation performance was found to be not very sensitive to the value of τ.

The average accuracy of the mode estimation was 86 percent. The errors are almost entirely confined to the segmentation boundaries as can be seen in figures 5 and 6. At other times, the mode is usually correctly estimated with a high degree of confidence. Figures 7 and 8 compare the estimated joint angles with the ground truth obtained from the tracking algorithm (Lien et al., 2007).

5 CONCLUSIONS

In this paper, we have proposed a new approach to the problem of representation and recognition of human motion. Our experimental results clearly indicate the validity of our proposal. However, there are several issues that need to be addressed to solve the action recognition problem comprehensively, within this framework. Our experiments indicate that while different subjects might share a common goal for the motion, they might tend to tradeoff the competing concerns of accuracy, energy and time differently. We are currently working on extending the estimation algorithm to estimate the relative weights online, along with the state and the mode.

REFERENCES

Blom, H. A. P. and Bar-Shalom, Y. (1988). The interacting multiple model algorithm for systems with markovian switching coefficients. *Automatic Control, IEEE Transactions on*, 33(8):780–783.

Bregler, C. and Malik, J. (1997). Learning and recognizing human dynamics in video sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages pp 568–674.

de Freitas, N. (2002). Rao-blackwellised particle filtering for fault diagnosis. *Aerospace Conference Proceedings, 2002. IEEE*, 4.

Del Vecchio, D., Murray, R., and Perona, P. (2003). Decomposition of human motion into dynamics based primitives with ap- plication to drawing tasks. *Automatica*, 39(12):2085–2098.

Fod, A., Matarić, M., and Jenkins, O. (2002). Automated Derivation of Primitives for Movement Classification. *Autonomous Robots*, 12(1):39–54.

Harris, C. and Wolpert, D. (1998). Signal-dependent noise determines motor planning. *Nature*, 394(6695):780–4.

Lewis, F. and Syrmos, V. (1995). *Optimal Control*. Wiley-Interscience.

Li, W. and Todorov, E. (2004). Iterative linear-quadratic regulator design for nonlinear biological movement systems. *First International Conference on Informatics in Control, Automation and Robotics*, 1:222–229.

Lien, J.-M., Kurillo, G., and Bajcsy, R. (2007). Skeleton-based data compression for multi-camera tele-immersion system. In *Proceedings of the International Symposium on Visual Computing, Lake Tahoe, Nevada/California,Nov 2007, to appear.*

McGinnity, S. and Irwin, G. (2000). Multiple model bootstrap filter for maneuvering target tracking. *Aerospace and Electronic Systems, IEEE Transactions on*, 36(3):1006–1012.

Murray, R., Sastry, S., and Li, Z. (1994). *A Mathematical Introduction to Robotic Manipulation*. CRC Press.

Nori, F. and Frezza, R. (2005). Control of a manipulator with a minimum number of motion primitives. *Proceedings of the 2005 IEEE International Conference on Robotics and Automation, 2005.*, pages 2344–2349.

Oliver, N., Garg, A., and Horvitz, E. (2004). Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding*, 96(2):163–180.

Park, S. and Aggarwal, J. (2004). A hierarchical Bayesian network for event recognition of human actions and interactions. *Multimedia Systems*, 10(2):164–179.

Pitt, M. K. and Shephard, N. (2001). Auxiliary variable based particle filters. In book Sequential Monte Carlo Methods in Practice, Arnaud Doucet - Nando de Freitas - Neil Gordon (eds). Springer-Verlag, 2001.

Safonova, A., Hodgins, J., and Pollard, N. (2004). Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. *ACM Transactions on Graphics (TOG)*, 23(3):514–521.

Scott, S. (2004). Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience*, 5(7):532–546.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience*, 2004:907–915.

Weinland, D., Ronfard, R., and Boyer, E. (2006). Automatic Discovery of Action Taxonomies from Multiple Views. *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Volume 2*, pages 1639–1645.

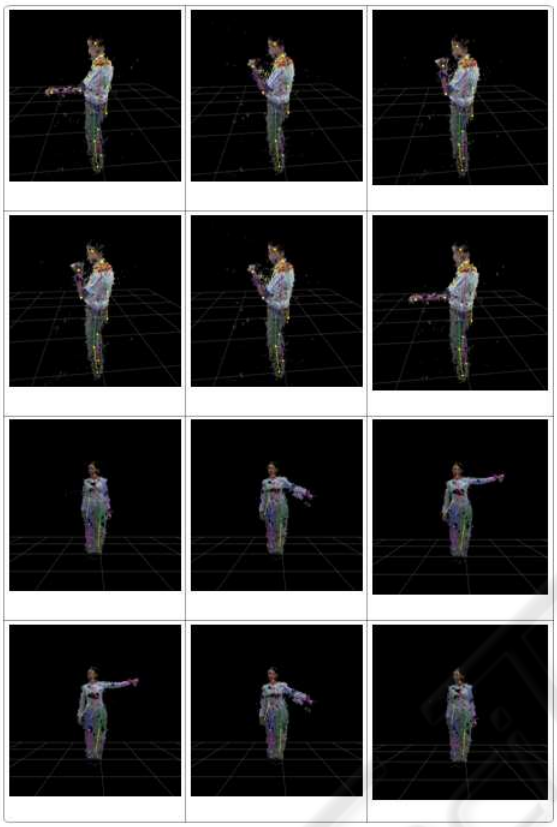

Figure 4: Modes in the Data. Row 1 : Task 1, Mode 1. Row 2 : Task 1, Mode 2. Row 3 : Task 2, Mode 3. Row 4 : Task 2, Mode 4. In Task 1, the subjects lift and lower the weights as indicated in rows 1 and 2, by only rotating the elbow joint. In Task 2, the subjects lift and lower the weights as indicated in rows 3 and 4, by only rotating the shoulder joint about the x-axis.

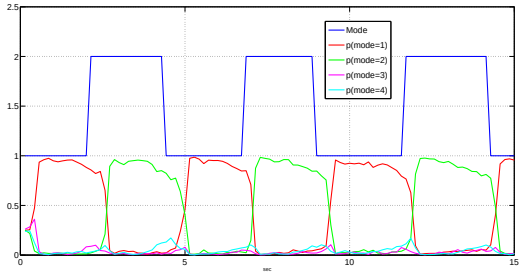

Figure 5: Mode Estimation : subject 2, task 1. In this task, the mode switches between 1 and 2, as indicated by the blue line. The other lines indicate the probability of each mode at each instant.

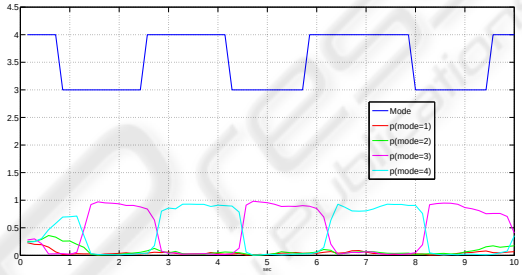

Figure 6: Mode Estimation : subject 1, task 2. In this task, the mode switches between 3 and 4, as indicated by the blue line. The other lines indicate the probability of each mode at each instant.

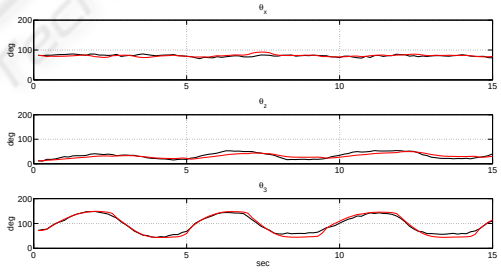

Figure 7: State Estimation : subject 2, task 1. Legend : black line is the ground truth from the tracking algorithm, red line is the estimate of the state.

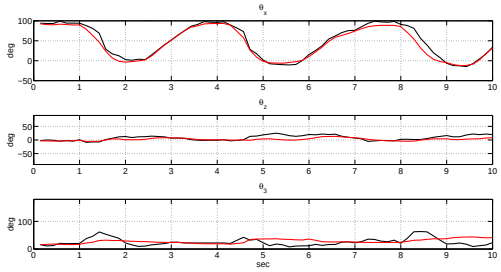

Figure 8: State Estimation : subject 1, task 2. Legend : black line is the ground truth from the tracking algorithm, red line is the estimate of the state.