Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Human Intention Estimation using Fusion of Pupil and Hand Motion

Daniel Trombetta * **Ghananeel S. Rotithor** ** **Iman Salehi** ***
Ashwin P. Dani ****

* *Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269 USA (e-mail: daniel.trombetta@uconn.edu)*
** *Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269 USA (e-mail: ghananeel.rotithor@uconn.edu)*
*** *Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269 USA (e-mail: iman.salehi@uconn.edu)*
**** *Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269 USA (e-mail: ashwin.dani@uconn.edu)*

Abstract: This paper addresses the problem of human intention inference in the context of human-robot collaboration by fusing information from both hand motion obtained using skeletal tracking and eye gaze obtained using pupil tracking to predict a human's current intention. Intention is modeled as a motion profile that converges to a goal location. A Kalman filter is used on eye gaze data to obtain gaze point estimates. The gaze estimates are transformed into a reference frame common to the hand data. An IMM filter that tracks hand motion is designed which takes advantage of the gaze filter's model probabilities by fusing them with its own. The fusion is performed with user chosen parameters that determine the degree to which each filter's predictions are weighed over time. An experiment is designed to show the utility of the proposed algorithm in a setting in which multiple reaching tasks are completed in an unknown order. The results show that the proposed algorithm can accurately predict the human's intention before the tasks are completed.

1. INTRODUCTION

When two humans interact, they infer one another's intent in order to safely and effectively collaborate Baldwin and Baird (2001); Simon (1982). When humans and robots collaborate, inference of the human's intention improves the overall performance of the task Liu et al. (2016); Li and Ge (2014); Warrier and Devasia (2016). So far in the human-robot collaboration context, the fusion of pupil tracking and hand motion data to estimate the human intention is not studied. This paper presents a methodology for early estimation of human's hand reaching intention by fusing information from pupil tracking and hand motion. Over the past decade, there has been an increased interest in the measurement and estimation of 3-dimensional (3D) human eye gaze. Studies have suggested a human's gaze is directly related to their intended actions Yarbus (1967). In Flanagan and Johansson (2003), it is demonstrated that adults predict action goals by fixating on the end location of an action even before it is reached. In Kleinke (1986), it is shown that gaze communicates attention. It follows naturally that gaze information is a likely candidate to improve current human intention estimation schemes. Various different modalities of information about the human, e.g., characteristics of the objects in the workspace (Koppula et al., 2013), human movement (Mainprice et al., 2015),

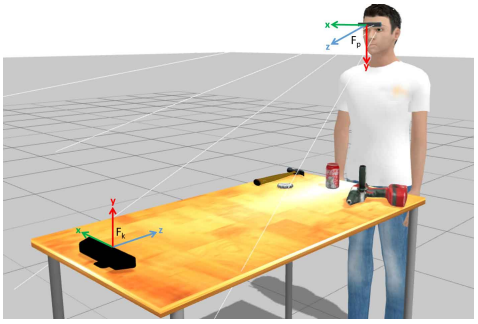

Fig. 1. A simulated test scenario generated in Gazebo. Hand motion measurements are acquired by a Kinect sensor, and gaze data is collected by Pupil Glasses.

or physiological information, such as electromyography (Razin et al., 2017), heart rate and skin response (Kulic and Croft, 2007), have been used to infer the intention. In Sakita et al. (2004), gaze is used as an intention measure to provide instruction to a robot assistant. In Strabala et al. (2013), the intention to handover an object is predicted by using key features extracted from the vision and the pose (position + orientation) data. In Lang et al. (2017), a Gaussian Process (GP) is used to predict hand trajectories during an object handover task. Intention inference as a goal-reaching motion profile estimation for collaboratively carrying heavy objects is presented in (Ravichandar and Dani, 2015a). In Ravichandar et al. (2018), gaze information estimated from RGB camera images using a convolution neural network (CNN) model is utilized to ini-

⋆ This work was in part supported by a Space Technology Research Institutes grant (number 80NSSC19K1076) from NASA Space Technology Research Grants Program, and in part by Office of Naval Research Award No. N00014-20-1-2040

tialize model probabilities for hand motion tracking using an Interacting Multiple Model (IMM) filter Bar-Shalom et al. (2001). This method, however, does not utilize gaze information after initialization.

The algorithm presented in this paper continuously monitors the human's gaze using pupil data and hand motion during a task. A method of acquiring the homogeneous transformation between the reference frame of the gaze measurements and the reference frame of the hand measurements is presented. As shown in Figure 1, F_P represents the reference frame of the gaze measurements, and F_K represents the reference frame of the hand measurements. The Hyperface convolutional neural network (CNN) presented in Ranjan et al. (2017) is used to predict orientations of the human head in RGB images such that the filtered gaze data recorded in F_P can be transformed into a reference frame common to the hand position data F_K. This information is leveraged before the mixing stage of each iteration of the IMM filter. The fusion equation contains two free parameters that can be chosen by the user. The first parameter controls the degree to which the model probabilities from each filter are weighed as time progresses. The second controls the rate at which the shift occurs. This enables the algorithm to determine the desired reaching motion intention when tasks are performed sequentially as the gaze is likely to fixate on the target before the hand gets close to the target. The shift in weight over time prevents the algorithm from failing in the likely event that the gaze will shift to the next goal in a sequence before the previous reaching motion is complete. Section 2 describes a generic case wherein the proposed algorithm is viable. In Section 3, a description is given on converting points in F_P to F_K as well as a summary of the Hyperface CNN. In Section 4, models for the gaze and hand data are given. The novel algorithm for human intention estimation using fused gaze and motion model likelihoods is proposed in Section 5. Section 6 describes the experiment performed and results. Figure 1 shows a simulated recreation of the described set-up.

2. PROBLEM FORMULATION

Consider a situation in which a human and a robot are working collaboratively to complete an objective comprised of a sequence of subtasks such as a manufacturing assembly or a surgical task. The sequence is not necessarily required to be completed in a specific order, i.e., the order of the tasks can be interchanged to achieve the same desired result. Each task is associated with a model which consists of a motion profile that terminates at a goal location. The goal locations are the positions of the task-relevant objects in the workspace. The only information known to the robot at the onset of the objective is the goal location of each model. The human partner may progress through the sequence in any order which they see fit. The robot does not have knowledge of the sequence a priori but must be able to infer which task is currently being completed. During the operation, the robot partner collects measurements of the human partner's current hand motion and gaze point location. Using this information, the robot must infer which model the human is operating under, or moreover, which task the human is currently performing. A Microsoft Kinect Sensor is used to track the 3-dimentional

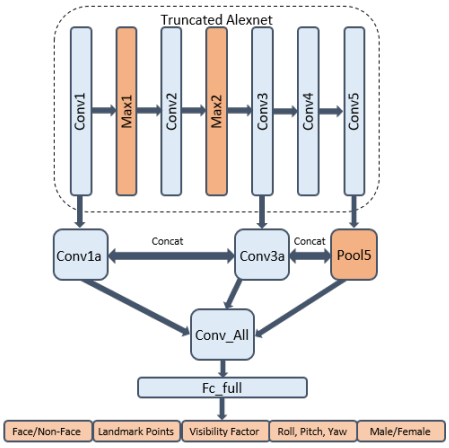

Fig. 2. Overview of the Hyperface CNN architecture.

(3D) position of the human's skeleton in the Kinect frame, F_K, and Pupil Glasses by Pupil Labs Kassner et al. (2014) worn by the human are used to acquire 3D estimates of the human's gaze locations in the Pupil glasses frame, F_P. The transformation of the gaze points into F_K can be obtained using the 3D location of the human's head as tracked by the Kinect sensor in conjunction with the Hyperface CNN, which can detect faces in RGB images and predict the roll, pitch, and yaw orientations.

3. EYE GAZE DATA PROCESSING

Given a single 3D measurement of a gaze point $A_P \in \mathbb{R}^3$ in the Pupil glasses frame F_P, it is required to transform A_P into frame F_K in order to calculate corresponding model probabilities. The motivation behind this transformation is two fold. Firstly, the positions of the goal locations are defined in F_K. More importantly, it is desirable to have the gaze point measurements with respect to a static frame. F_P, however, is dynamic with respect to the goal locations because the glasses are being worn by a mobile human. Let the origin of F_P be approximated by the 3D coordinates of the human head, $X_{head} \in \mathbb{R}^3$, measured by the Kinect sensor in F_K. Thus, the translation component of the homogeneous transformation matrix can be represented by X_{head}. The rotation component can be obtained using the Hyperface CNN. Hyperface is a CNN architecture that takes an RGB image of any size as its input and returns whether or not the image contains one or more faces, places landmarks on relevant points of the faces, provides a measure of the visibility of the landmarks, gives estimates of the roll, pitch, and yaw of each face in the image in its own reference frame F_H, and predicts the gender associated with each face. Hyperface is trained on annotated images provided by the Annotated Facial Landmarks in the Wild (AFLW) dataset Koestinger et al. (2011). The general structure of the Hyperface CNN architecture is shown in Figure 2. The RBG images of size 640×480 collected by the Kinect sensor are used as input for the Hyperface CNN in order to obtain solely the roll, pitch, and yaw estimates of the human's head in F_H. In order to utilize this information, the point A_P must first be transformed into the Hyperface frame F_H using $A_H = R_P^H A_P$, where P_H denotes a point in the the Hyperface frame F_H and $R_P^H \in \mathbb{R}^{3\times3}$ denotes the rotation

matrix from frame F_H to F_P. The complete transformation can then be given as

$$A_K = R_H^K R_P^H A_P + X_{head} \tag{1}$$

$$T_P^K = \begin{bmatrix} R_H^K R_P^H & X_{head} \\ 0_{1\times 3} & 1 \end{bmatrix} \tag{2}$$

where $R_H^K \in \mathbb{R}^{3\times 3}$ denotes the rotation matrix from F_K to F_H and is obtained using the prediction of the human head orientation from Hyperface, and $T_P^K \in \mathbb{R}^{4\times 4}$ is the homogeneous transformation that maps the point $\left[(A_p)^T, 1\right]^T$ to the point $\left[(A_K)^T, 1\right]^T$.

4. MOTION MODELS

In this section, human hand motion and human eye-gaze motion models are described in detail.

4.1 Human Hand Motion

At any given time, the human is assumed to be operating according to one of N models. Let $G = [g_1, g_2, \ldots, g_N]$ represent the vector of all N goal locations. Then, the i^{th} model M_i is associated with a single goal location g_i. Each model is characterized by the motion of the human hand as well as the evolution of the gaze point. The human hand motion associated with the i^{th} model is given by

$$\begin{bmatrix} x_H(k+1) \\ \dot{x}_H(k+1) \\ \ddot{x}_H(k+1) \end{bmatrix} = \begin{bmatrix} \mathrm{diag}_3(1) & \mathrm{diag}_3(T_s) & \mathrm{diag}_3(\frac{1}{2}T_s^2) \\ 0 & \mathrm{diag}_3(1) & \mathrm{diag}_3(T_s) \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_H(k) \\ \dot{x}_H(k) \\ \ddot{x}_H(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ f_i(x_H(k), \dot{x}_H(k), \ddot{x}_H(k)) \end{bmatrix} + \begin{bmatrix} W_1 \\ W_2 \\ W_3 \end{bmatrix} w_1(k) \tag{3}$$

where $x_H \in \mathbb{R}^3$ is the 3-dimentional (3D) position of the human hand, T_s is the sampling time, the operator $\mathrm{diag}_\eta(\rho)$ denotes a square matrix of dimension $\eta \times \eta$ with the value ρ along the central diagonal, $f_i : \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ is a continuously differentiable function associated with the i^{th} model, $W_1 = \mathrm{diag}_3(\frac{1}{6}T_s^3)$, $W_2 = \mathrm{diag}_3(\frac{1}{2}T_s^2)$, $W_3 = \mathrm{diag}_3(T_s)$, and $w_1 \sim \mathcal{N}(0, Q_1)$ is a Gaussian distributed process noise with zero mean and known covariance $Q_1 \in \mathbb{R}^{3\times 3}$ that represents the model uncertainty in acceleration update. Each function f_i is approximated by a neural network whose parameters are learned from data collected during the training phase. The training is performed subject to a contraction metric which garuntees, even with minimal training data, that predictions made by each f_i are stable and exponentially converge to the i^{th} goal location. For a more detailed look at the training method, the reader is referred to Ravichandar and Dani (2015b). The noisy measurements of the human partner's hand positions are modeled as

$$z_H(k) = x_H(k) + \nu_1(k) \tag{4}$$

where $\nu_1(k) \in \mathbb{R}^{3\times 3}$ is a Gaussian distributed measurement noise with zero mean and known covariance $R_1 \in \mathbb{R}^{3\times 3}$.

4.2 Human Eye-Gaze Motion

Unlike the hand motion, which has a distinct motion model for each goal location, there is a single model for the gaze as the behavior of the eye motion is expected to be similar regardless of goal location. The evolution of the human's gaze point is modeled as

$$\begin{bmatrix} x_E(k+1) \\ \dot{x}_E(k+1) \end{bmatrix} = \begin{bmatrix} \mathrm{diag}_3(1) & \mathrm{diag}_3(T_s) \\ 0 & \mathrm{diag}_3(1) \end{bmatrix} \begin{bmatrix} x_E(k) \\ \dot{x}_E(k) \end{bmatrix} + \begin{bmatrix} W_4 \\ W_5 \end{bmatrix} w_2(k) \tag{5}$$

where $x_E \in \mathbb{R}^3$ is the 3D position of the human's gaze, $W_4 = \mathrm{diag}_3(\frac{1}{2}T_s^2), W_5 = \mathrm{diag}_3(T_s)$, and $w_2 \sim \mathcal{N}(0, Q_2)$ is a Gaussian distributed process noise with zero mean and known covariance $Q_2 \in \mathbb{R}^{3\times 3}$ that represents model uncertainties in velocity update. The measurement model is given as

$$z_E(k) = x_E(k) + \nu_2(k) \tag{6}$$

where $\nu_2(k) \in \mathbb{R}^{3\times 3}$ is a Gaussian distributed measurement noise with zero mean and known covariance $R_2 \in \mathbb{R}^{3\times 3}$.

5. ESTIMATION OF HUMAN INTENTION

Consider the N models, $M_1, M_2, \ldots, M_N$, with goal locations $g_1, g_2, \ldots g_N$ and the measurement models defined in (4) and (6). Let $X_H = [x_H^T, \dot{x}_H^T, \ddot{x}_H^T]^T$, $X_E = [x_E^T, \dot{x}_E^T]^T$ denote the human hand and eye-gaze state vectors, respectively, and $Z_H^{1:k} = [z_H(1), z_H(2), \ldots, z_H(k)]$, $Z_E^{1:k} = [z_E(1), z_E(2), \ldots, z_E(k)]$ denote a set of k measurements of the human hand and the eye-gaze, respectively. The objective is to fuse the information obtained from the measurements of the gaze point and the human hand in order to infer which model the human is currently operating under and effectively compute the state estimate $\hat{X}_H(k|k)$. Note that the true model that the human is operating under is not known and the human could switch among the N models at any time. The formulation is separated into two subsections. First, a Kalman filter (KF) that estimates the current gaze point is presented. The gaze point estimates are used to calculate probabilities that the human is operating according to each model. The second is an IMM filter for human hand motion which uses N extended Kalman filters (EKFs) running in parallel to filter the hand motion. At the beginning of each iteration of the IMM filter, the model posterior probabilities produced in the previous iteration are fused with those from the eye-gaze filter to generate more informative model probabilities.

5.1 Eye-gaze Filter

Using the dynamics in (5) and the noisy measurements in (6), a Kalman filter is designed to obtain estimates of the gaze point and the corresponding covariances. Once two measurements are available, the filter is initialized using the two-point differencing method. Each iteration, $\hat{x}_E$ and S, the eye-gaze filter's state estimate and innovation covariance, respectively, are obtained. The probability that the current estimated gaze point is associated with the model having a goal location g_j can be represented as

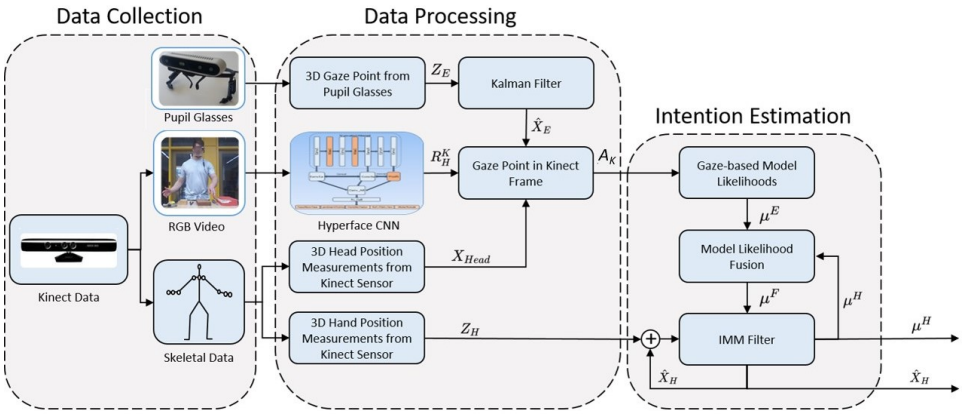

Fig. 3. A block diagram to summarize the data aquisition, processing, intention estimation, and hand motion prediction for the proposed algorithm.

$$\mu_j^E(k) = P(\hat{x}_E(k)|g_j)$$
$$\mu_j^E(k) = \frac{1}{\sqrt{|2\pi S|}} e^{-\frac{1}{2}(\hat{x}_E - g_j)^T S^{-1}(\hat{x}_E - g_j)} \quad (7)$$

5.2 Hand Motion Filter With Gaze Fusion

Once the first model probability $\mu_j^E(0)$ is made available by the eye-gaze filter, the prior probabilities for the IMM $\mu_j^H(0)$ can be initialized as $\mu_j^H(0) = \mu_j^E(0)$. In subsequent iterations of the filter, the model probabilities from each filter $\mu_j^H(k)$ and $\mu_j^E(k)$ are fused according to

$$\mu_j^F(k) = \alpha e^{-\beta T_t} \mu_j^E(k) + (1 - \alpha e^{-\beta T_t}) \mu_j^H(k) \quad (8)$$

where $T_t \in [0, \infty)$ denotes the time since the most recent model switch, $\alpha \in [0, 1]$ is a user defined parameter that determines the degree to which the weight shifts between the model probabilities from the two filters over time, and $\beta \in (0, \infty)$ is a user defined parameter which controls how quickly the weight shift occurs. Dynamically weighing the model probabilities can account for situations wherein one source is expected to provide more reliable insight. For example, at the onset of a task, the human partner may look directly at the goal location even though their hand may be far away. However, as time goes on, the human may begin to look at the next goal location in a sequence before their hand reaches the previous goal. In this case, one would want $\mu_j^E(k)$ to hold a higher weight at the beginning of a task because it provides better insight to the current objective. However, toward the end of the objective, $\mu_j^H(k)$ should hold more weight because the gaze point has shifted although the current goal location has not changed.

Interaction/Mixing: At the beginning of each iteration, the initial conditions (state estimate $\hat{x}_H^{0j}(k-1|k-1)$ and covariance $\hat{P}_H^{0j}(k-1|k-1)$), where superscript 0 denotes initial condition, j denotes the number of the filter, at time k, are adjusted by mixing the filter outputs from the previous iteration (time instant $k-1$) in the following way

$$\hat{x}_H^{0j}(k-1|k-1) = \sum_{i=1}^{N} \hat{x}_H^i(k-1|k-1) \times \mu_{i|j}^F(k-1|k-1), j = 1, .., N \quad (9)$$

$$\hat{P}_H^{0j}(k-1|k-1) = \sum_{i=1}^{N} \mu_{i|j}^F(k-1|k-1) \hat{P}_H^i(k-1|k-1) + [\hat{x}_H^i(k-1|k-1) - \hat{x}_H^{0j}(k-1|k-1)] [\hat{x}_H^i(k-1|k-1) - \hat{x}_H^{0j}(k-1|k-1)]^T, \quad j = 1, .., N \quad (10)$$

where $\hat{x}_H^i(k-1|k-1)$, $\hat{P}_H^i(k-1|k-1)$ are the state estimate and its covariance respectively corresponding to model M_j at time $k-1$ and the mixing probabilities $\mu_{i|j}^F(k-1|k-1)$ are given by

$$\mu_{i|j}^F(k-1|k-1) = \frac{\Pi_{ij} \mu_i^F(k-1)}{\bar{c}_j}, \; i, j = 1, 2, ..., N \quad (11)$$

where $\Pi_{ij} = p(M(k) = M_j | M(k-1) = M_i)$ is the model transition or jump probability and $\mu_i^F(k-1) = p(M_i | Z_H^{1:k-1}, Z_E^{1:k-1})$ is the fused probability of i^{th} model M_i being the right model at time $k-1$ and $\bar{c}_j = \sum_{i=1}^{N} \Pi_{ij} \mu_i^F(k-1)$ are the normalizing constants.

Model Matched Filtering: Once the initial conditions $\hat{x}_H^{0j}(k-1|k-1)$ and $\hat{P}_H^{0j}(k-1|k-1)$ are available for each filter, the state estimate and its covariance for each model are computed using the EKFs matched to the models. Along with the state estimates and the corresponding covariances, the likelihood functions $\Lambda_j(k)$ are computed using the mixed initial condition (9) and the corresponding covariance (10). The likelihood, a Gaussian distribution with the predicted measurement as the mean and the covariance equal to the innovation covariance, is given by

$$\Lambda_j(k) = p(z_H(k) | M_j(k), Z_H^{1:k-1})$$
$$\Lambda_j(k) = \mathcal{N}(z_H(k); \hat{z}_H^j(k|k-1; \hat{x}_H^{0j}(k-1|k-1)), S_H^j(k; \hat{P}_H^{0j}(k-1|k-1))), \; j = 1, .., N \quad (12)$$

where $S_H^j(k; P_H^{0j}(k-1|k-1))$ is the innovation covariance and $\hat{z}_H^j(k|k-1; \hat{x}_H^{0j}(k-1|k-1))$ is the j^{th} filter's predicted measurement at time t.

Model Probability Update: After the likelihood functions of the models $\Lambda_j(k)$ are available, the model posterior probabilities $\mu_j^H(k)$ are calculated as follows

$$\mu_j^H(k) = P(g_j|Z_H^{1:k}) = P(M_j(k)|Z_H^{1:k})$$
$$\mu_j^H(k) = p(z_H(k)|M_j(k), Z_H^{1:k-1})P(M_j(k)|Z_H^{1:k-1})$$
$$\mu_j^H(k) = \frac{\Lambda_j(k)\bar{c}_j}{\sum_{i=1}^{N}\Lambda_i(k)\bar{c}_i}, \qquad j = 1, 2, ..., N \tag{13}$$

and the goal location estimate $\hat{g}(t)$ is given by

$$\hat{g}(k) = \arg\max_{g\in G} \ \mu_j^H(k) \tag{14}$$

The optimization problem in (14) is solved by choosing the location $g_i \in G$ corresponding to the model M_i with the highest model probability $\mu_i^H(k)$ at time k. Figure 3 summarizes the gaze and motion fusion algorithm in the form of a block diagram.

Fig. 4. Five frames from the RGB video collected by the Kinect sensor each overlaid with a bounding box around the current predicted goal object.

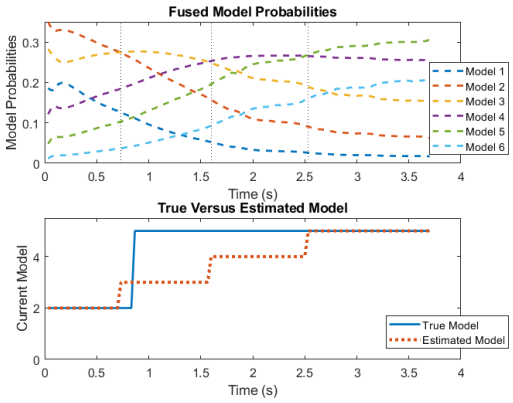

Fig. 5. The top plot shows vertical dotted lines when the current most likely model has changed. The lower plot compares the true model with the predicted model.

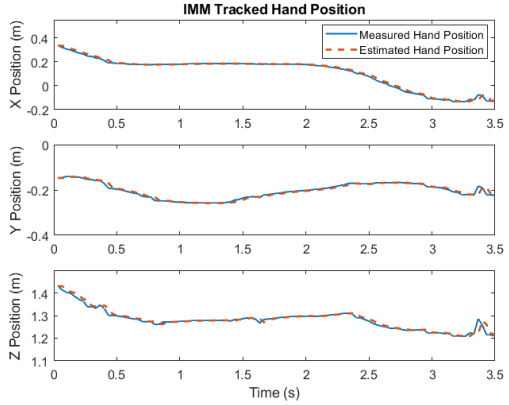

Fig. 6. Hand position tracked by the IMM filter using fused model probabilities.

6. EXPERIMENTAL RESULTS

In order to validate the utility of the proposed data fusion method, an experiment is designed in which a human partner must complete a set of tasks in any order. Using available measurements of the human hand motion and eye gaze location, the robot must determine the sequence of tasks being completed on the fly. This experimental structure is analogous to real life collaborative tasks such as two workers collaboratively hammering nails into a board at multiple locations, carrying a heavy object from one location to one of many possible destinations, or manufacturing a product that may have leniency in the order in which it is assembled, i.e. an electrical circuit. For this experiment, six goal objects, $N = 6$, are used: a hammer, a screwdriver, pliers, two wood blocks of different sizes, and a cardboard box. The objects are placed arbitrarily within the field of view of a Kinect sensor at known locations in F_K. The human first grabs one of the tools at random, and relocates it atop any one of the three boxes as they see fit. The human's 3D gaze point is determined using noisy measurements provided by Pupil glasses worn by the human partner and the filter described in (5). Noisy measurements of the human hand motion are made available by the Kinect motion sensor's skeletal tracking feature according to (4). The gaze point can be represented in the Kinect frame using (1). The objective is to show that the proposed algorithm can predict which tool the human is reaching for before it is grasped, and determine where it will be placed before it reaches that point. The results of this experiment are shown in the following section. Figure 4 shows frames from the RGB video acquired by the Kinect at relevant time instances, namely, the first and last frames along with each frame in which a model switch was predicted. A bounding box has been overlaid on each image around the object which the algorithm believes is the goal location at the current time instance. The parameter α from (8) was chosen to be 0.6 and β was chosen to be 1 meaning that whenever a model change is predicted, the fused model probabilities are weighted 60% on μ^E and only 40% on μ^H. The more time spent operating under the same model, the more weight that is shifted to μ^H. As a direct result of this, the prediction of the goal location at the onset of the experiment is correct even though the human hand is not yet near the target object. As the subject reaches for the first goal location, i.e., the screwdriver, their gaze begins to move towards the next goal location. This causes the model prediction to change slightly before the true model changes and once again gives μ^E a higher weight. The incorrect intermediate predictions seen in frames 22 and 46 are due to the gaze point traveling over the objects between the previous goal and the current goal. Frame 76 shows that the correct model prediction is made 35 frames before the screwdriver is actually placed on the wood board in frame 111. Figure 5 shows the performance of the algorithm over time. The top plot shows the evolution of each of the six model probabilities associated with each object in the experiment. The vertical dotted lines denote the times when the algorithm predicts that the current model has changed. The bottom graph shows a comparison between the true intention and the estimated intention.

The hand position tracked by the IMM filter using the fused model probabilities can be seen in Figure 6.

7. CONCLUSION

A novel framework is presented that fuses information from skeletal tracking measurements of a human hand and pupil position measurements in order to predict which of a finite and known set of actions is the human's true action intention. In an initial training phase, the human demonstrates each of the possible actions and hand positions are recorded using skeletal tracking with a Microsoft Kinect Sensor. Hand position data is then used to train multiple single layer NN's (one per action) subject to contraction constraints such that the predictions made by the NN are stable. In the execution phase, the Kinect Sensor is used to collect skeletal tracking data of the human's hand and head, and Pupil Lab's Pupil Glasses, worn by the human, are used to acquire 3D gaze point estimates from pupil position data. The gaze tracking data is transformed into a frame common to the hand tracking data by using the Hyperface CNN to predict roll, pitch, and yaw orientations associated with a face detected in an RGB image and the human head tracked by the Kinect Sensor to predict the homogenous transformation matrix between the two frames. Each iteration, a Kalman filter is used on the transformed gaze data in order determine model likelihoods conditioned on gaze measurements. The first likelihood aqcuired from the gaze filter is used to initialize the hand tracking IMM. In subsequent interations, the likelihoods from the Kalman filter and IMM are fused before the Mixing/Interacting stage of the IMM. The IMM predicts both the human hand position conditioned on both the gaze and hand measurements as well as the current action intention from the trained models. This algorithm is shown to accurately predict the correct action intention before the completion of each task in a sequence.

REFERENCES

Baldwin, D.A. and Baird, J.A. (2001). Discerning intentions in dynamic human action. *Trends in cognitive sciences*, 5(4), 171–178.

Bar-Shalom, Y., Li, X.R., and Kirubarajan, T. (2001). *Estimation with Applications to Tracking and Navigation*. John Wiley and Sons.

Flanagan, J.R. and Johansson, R.S. (2003). Action plans used in action observation. *Nature*, 424(6950), 769–771.

Kassner, M., Patera, W., and Bulling, A. (2014). Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 1151–1160.

Kleinke, C.L. (1986). Gaze and eye contact: a research review. *Psychological bulletin*, 100(1), 78.

Koestinger, M., Wohlhart, P., Roth, P.M., and Bischof, H. (2011). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *IEEE international conference on computer vision workshop*, 2144–2151.

Koppula, H.S., Gupta, R., and Saxena, A. (2013). Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8), 951–970.

Kulic, D. and Croft, E.A. (2007). Affective state estimation for human–robot interaction. *IEEE Transactions on Robotics*, 23(5), 991–1000.

Lang, M., Endo, S., Dunkley, O., and Hirche, S. (2017). Object handover prediction using gaussian processes clustered with trajectory classification. *arXiv preprint arXiv:1707.02745*.

Li, Y. and Ge, S. (2014). Human-robot collaboration based on motion intention estimation. *IEEE/ASME Transactions on Mechatronics*, 19(3), 1007–1014.

Liu, C., Hamrick, J.B., Fisac, J.F., Dragan, A.D., Hedrick, J.K., Sastry, S.S., and Griffiths, T.L. (2016). Goal inference improves objective and perceived performance in human-robot collaboration. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*, 940–948.

Mainprice, J., Hayne, R., and Berenson, D. (2015). Predicting human reaching motion in collaborative tasks using inverse optimal control and iterative re-planning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 885–892.

Ranjan, R., Patel, V.M., and Chellappa, R. (2017). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1), 121–135.

Ravichandar, H. and Dani, A.P. (2015a). Human intention inference through interacting multiple model filtering. In *IEEE Conference on Multisensor Fusion and Integration*, 220–225.

Ravichandar, H. and Dani, A.P. (2015b). Learning contracting nonlinear dynamics from human demonstration for robot motion planning. In *ASME Dynamic Systems and Control Conference (DSCC)*.

Ravichandar, H.C., Kumar, A., and Dani, A. (2018). Gaze and motion information fusion for human intention inference. *International Journal of Intelligent Robotics and Applications*, 2(2), 136–148.

Razin, Y.S., Pluckter, K., Ueda, J., and Feigh, K. (2017). Predicting task intent from surface electromyography using layered hidden markov models. *IEEE Robotics and Automation Letters*, 2(2), 1180–1185.

Sakita, K., Ogawara, K., Murakami, S., Kawamura, K., and Ikeuchi, K. (2004). Flexible cooperation between human and robot by interpreting human intention from gaze information. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 1, 846–851.

Simon, M.A. (1982). *Understanding human action: Social explanation and the vision of social science*. SUNY Press.

Strabala, K.W., Lee, M.K., Dragan, A.D., Forlizzi, J.L., Srinivasa, S., Cakmak, M., and Micelli, V. (2013). Towards seamless human-robot handovers. *Journal of Human-Robot Interaction*, 2(1), 112–132.

Warrier, R.B. and Devasia, S. (2016). Inferring intent for novice human-in-the-loop iterative learning control. *IEEE Transactions on Control Systems Technology*, 25(5), 1698–1710.

Yarbus, A.L. (1967). *Eye Movements and Vision*. Springer.