Optimized Eyecatch-based Object Interception

Leyuan Wang*
Dept. of Computer Science

Abstract

We optimize the human interception model performing fast visually guided tasks. Based on the original paper that raised the novel visuomotor framework, we implement and extend the vision model of a human catching a ball achieving a faster and more accurate result. As proposed in that paper, eye movements plays a central role in coordinating movements of the head, hand and body and each interception behavior is composed of discrete piecewise linear submovements directed by uncertain visual estimates of target movement.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation;

Keywords: eye movements, human animation, object interception, visuo coordination, Kalman filter

1 Introduction

Based on the measurements and observations of the original paper [Yeo et al. 2012], eye and head movements which enable clear vision are an important component of human behaviors of catching a ball. Different from the usual animation of interception taking consider only body movements . The novel observations are described as follows. First, human needs to use vision primarily to quickly estimates the trajectory of the ball as soon as the ball is thrown. And there are two types of eye movements, catch-up saccades and smooth pursuit. Catch-up saccades are fast movements and it happens usually at the beginning when the quality of pursuit is not good. While smooth pursuit tracks the object continuously and it usually happens after the target has passed the apex of its trajectory. Second, the hand begins to move very early before an accurate estimation is made and hand movements can be expressed as a two-component model composed of an stereotypical initial open loop phase , followed by a close loop phase with individual corrections. Third, human movements are composed of discrete, short-duration submovements. Finally, theres close synchronization between the submovements of catch-up saccades and that of hand, head and body movements. It appears that saccades are well synchronized with peak hand velocity and starting times of decomposed hand submovements. This is the most important observation in the original paper for it suggests a hypothesis that the eye and hand share the same motor program triggered by sensory events in visually guided interceptive movements. Making use of this hypothesis makes it a lot easier to implement a natural looking coordination.

*e-mail:leywang@ucdavis.edu

2 Background & Preliminaries

Kalman Filter The Kalman filter can be presented as one of the most simple dynamic Bayesian networks. It recursively calculates estimates of the unknown values of states over time using noisy incoming measurements and a mathematical process model. The algorithm has two steps, predict and update. First, it estimates the current states with uncertainties. When the outcome of the next state is observed, the state estimates are updated using a weighted average. The Kalman filter the original paper uses is linear and highly simplified. Some extensions such as extended Kalman filter and unscented Kalman filter work on nonlinear systems. The authors in considered a linear system, but for more realistic effect, a more complex nonlinear system is needed.

In the extended Kalman filter (EKF), different from basic Kalman filter, the state transition and observation models can be nonlinear functions.

$$x_k = f(x_{k-1}, u_k) + w_k$$
$$z_k = h(x_k) + v_k$$

Function f is used to compute the predicted state from the previous estimate, and function h is used to compute the predicted measurement from the predicted state. When h and f are highly nonlinear, EKF has poor performance for the covariance is propagated through linearization of the underlying non-linear model. While the unscented Kalman filter (UKF) uses a deterministic sampling method called unscented transform to pick a minimal number of points around the mean which are then propagated through the nonlinear functions. And then the mean and covariance are recovered from those functions. The predict and update steps are shown in the following equations.

Predict

$$x^a_{k-1|k-1} = [\hat{x}^T_{k-1|k-1} E[w^T_k]]^T$$
$$P^a_{k-1|k-1} = \begin{bmatrix} P_{k-1|k-1} & 0 \\ 0 & Q_k \end{bmatrix}$$

Update

$$P_{z_k z_k} = \sum_{i=0}^{2L} W^i_c [\gamma^i_k - \hat{z}_k][\gamma^i_k - \hat{z}_k]^T$$
$$P_{x_k z_k} = \sum_{i=0}^{2L} W^i_c [\sigma^i_{k|k-1} - \hat{x}_{k|k-1}][\gamma^i_k - \hat{z}_k]^T$$
$$K_k = P_{x_k z_k} P^{-1}_{z_k z_k}$$
$$\hat{x}_{k|k} = P_{k|k-1} - K_k P_{z_k z_k} K^T_k$$

Compared with standard Kalman filter, UKF is able to handle not only non-affine state transition and observation functions, as well as some non-Gaussian noise models while it has the same computational complexity.

3 Visual Estimation

Vision Model It's commonly hypothesized that the brain is able to use the internal models of object dynamics to estimate and predict the state of the object overtime [Wolpert et al. 1995]. We use

unscented Kalman filter to predict the state of the ball with plausible models of noise introduced by visual sensing.

Figure 1 defines a 3D coordinate of the eye frame based on the original paper. The origin is set at the center of the globe, the X-axis is aligned with the visual axis representing the depth of vision, the Y- and Z-axis represent the space in front of the eye vertical to the depth axis. The perceived state of the ball is represented as a six dimensional vector x=$\begin{pmatrix} p \\ \dot{p} \end{pmatrix}$ with p the estimated position and $\dot{p}$ the perceived velocity. The gray ellipsoid around p in figure 1 is the error covariance of uncertainty of the true object location p which is represented as a multivariate normal distribution calculated by its deviation θ from X-axis and the axis of rotation ω.

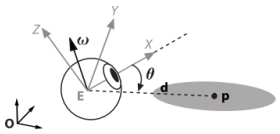

Figure 1: *Eye coordinate frame and the corresponding uncertainty of an object.*

As in the original paper, we represent the probability of the state of the ball seen by the eye as a multivariate normal distribution in the eye coordinate but with more complex model of noise.

$$\begin{bmatrix} p \\ \dot{p} \end{bmatrix} = N\left(\begin{bmatrix} \bar{p} \\ \dot{\bar{p}} \end{bmatrix}, A\Sigma^2 A^T\right) \tag{1}$$

$\bar{p}$ and $\dot{\bar{p}}$ are the true position and velocity of the ball. Σ is the error covariance and A is the transformation matrix from the target to the eye frame. The noise covariance matrix has six dimensions, respectively determined by spatial resolution (in Y- and Z-axis) and depth resolution(in X-axis). Similarly as the original paper, for spatial resolution, we use the standard 20/20 vision, and the small angle approximation to convert the error from retinal to eye coordinates. And since depth change is much worse than detecting the retinal location, and it's complex to measure, we also set the depth resolution to be much higher than spatial resolution.

Internal Model Using the vision model described above, we have a noisy observation of the target states. For simulation, instead of using a solid 20ms time step, we use 10ms during the beginning which is the catch-up saccades period where the eyes movements are fast, and 20ms to the smooth pursuit period where the eye movements are relatively predictable. This change makes the beginning stage of eye movements more realistic than the original paper.

For the internal model of the ball's dynamics, we represent it using quadratic polynomial function instead of a simple linear function. And we also assume that the brain has prior knowledge about gravity.

For the observer model, in order to more accurately estimate the observer, we use unscented Kalman filter described in section 2. The UKF is designed for state-estimation and nonlinear control application. It has better results than standard Kalman filter but has the same time complexity using Python. Besides, it can also adapt to special scenarios described in the original paper.

4 Movement Generation

Based on the target movement model described in previous section, we can now generate the gaze movement followed by synchronized movements of the head, hand and body.

To make this system efficient and tractable, we also choose to produce the movements kinematically same as the original paper. And as discussed in section 1, each movement is produced by blending short duration submovements together. And according to the original paper, each submovement is smooth, with a bell-shaped velocity profile which is chosen for it has fewer parameters and is thus relatively simple. The tangential velocity with unit displacement is defined in a time interval $t^0 < t < t^f$ as:

$$v(t^0, t, t^f) = \frac{30}{(t^f - t^0)^5}(t - t^0)^2(t - t^f)^2 \tag{2}$$

The velocity of a movement is correspondingly represented as a superposition of submovement velocity as follows:

$$\dot{u}(t) = \sum_{i=1}^{n} \dot{u}_i(t) = \sum_{i=1}^{n} b_i v(t_i^0, t_i^f, t) \tag{3}$$

where b_i is the basis vector representing the direction and magnitude of the submovement. Special attention needs to be paid to superimposed submovements: when the destination of the new submovement is given, b_i is not the difference between the destination and current position, but should be the difference between the new destination and the previous submovement destination. Submovement decompositions are done by nonlinear least squares optimization with pre-chosen number of submovements that minimizes the error between captured and composed trajectories. Animation of body movements are thus clear: for each update of visual information, corresponding submovements of each body part is determined and superimposed on the current motion.

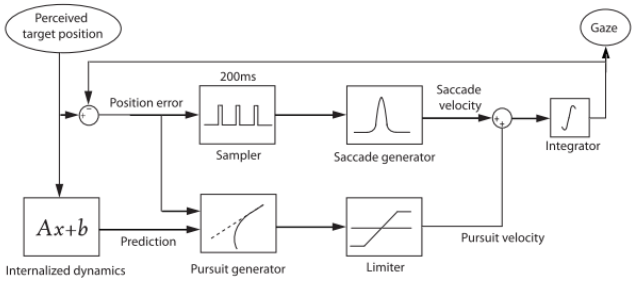

Figure 2: *Overall control structure of gaze.*

Gaze movements Given the estimated ball state in the previous section, we first need to determine gaze which is a primary coordinate in the whole system. Gaze is not significantly affected by body movement unless the eyes reach a biochemical limit in both humans and animals. Gaze can be divided into two parts, saccade and pursuit. Saccade is a very fast, poor-vision eye movement. Pursuit is a relatively slow eye movement that tends to track moving objects continuously.

Fig. 2 is the overall block diagram of gaze including saccade and pursuit. According to the original paper, the typical time required for two consecutive saccades is known to be around 200ms, and therefore we assume 200ms as the visuomotor update time that regulates the general movement including head and hand movement. The saccade amplitude is determined by current error between gaze and target angles. The velocity of saccade is chosen as in Eq.(2) while real saccades have a more asymmetric velocity profile. As for pursuit, we set its velocity to be updated every time step 20ms(Δt) with $v = (\bar{p} - p)/\Delta t$ where p is the current gaze and $\bar{p}$ is the position of the target at the next step predicted by the internal model.

Since the early estimation of the object is very noisy, if the saccade is triggered immediately when an object is detected, the initial eye movements will be very inaccurate and unrealistic. The brain avoids this problem by suppressing the initial saccade until the likelihood of position of the object exceeds some significance criterion. We relate this criterion by triggering the initial saccade only when the estimated error of the object position is reduced to a certain threshold.

Head movements Head movements are mainly for the assistant in eye tracking the target more easily. Thus the head movements are similar to eye movements except with a relatively slower speed and a longer lasting bell shaped velocity profile. The rotation of head follows the Donder's law [Crawford et al. 2003], which states that three-dimensional head rotation can be described by two parameters: longitudinal rotation and latitudinal rotation. The head rotation is $R = R_{h0}R_z(q_z)R_y(q_y)$. It's known that the same brain areas are involved in eye and head movement in both saccade and pursuit. Thus given a gaze shift, the corresponding saccade amplitude of the head is linearly related with a 20° dead zone. Once the head gaze change is determined, the peak velocity is also determined by its well known linear relationship to the amplitude with slope varying between $4s^{-1}$ to $8s^{-1}$. If we apply to Eq.(2), we get a constant submovement duration around 400ms. We define the submovement parameters in Eq.(3) for saccade velocity of the head gaze:

$$\xi = max(0, \theta - 0.349) \tag{4}$$

$$b_i = (e^{[w_i]\xi} - I)(g_{i-1} - p) \tag{5}$$

ξ is the required angular displacement of the head saccade and θ is the non-negative magnitude of the rotation. For detailed explanation of the model, we recommend reading the original paper.

Hand movements The interception can be modeled using a simple algorithm given the gaze behavior: when the eyes are continuously tracking the ball, we can simply move the hand towards the gaze. The hand will be driven sufficiently close to the ball when it reaches the body. The strategy is that before we trigger the final catch, we move our hand in the same way that we move hour eyes and head. Thus we only need to find when we make the final snatch, or the distance of the catch. As the data shown in the original paper in Fig. 3. From the 59 ball catching trials we can observe that each subject has a preferred interception distance from the head frame which may be affected by considerations such as manipulability of the arm or the effective compliance produced by muscle properties. For simplicity, we also choose a solid interception distance for catching. To determine the interception point, we

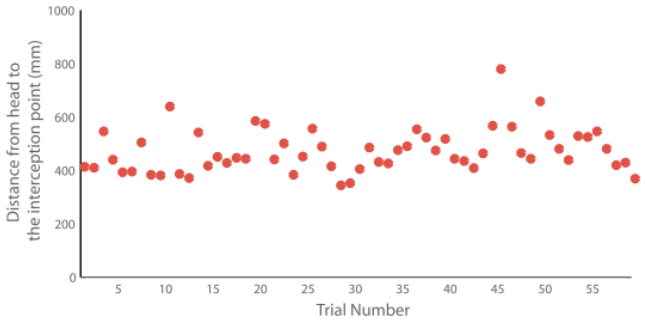

Figure 3: *Distance from head to the interception point.*

just need to draw a sphere with the preferred distance as the radius,

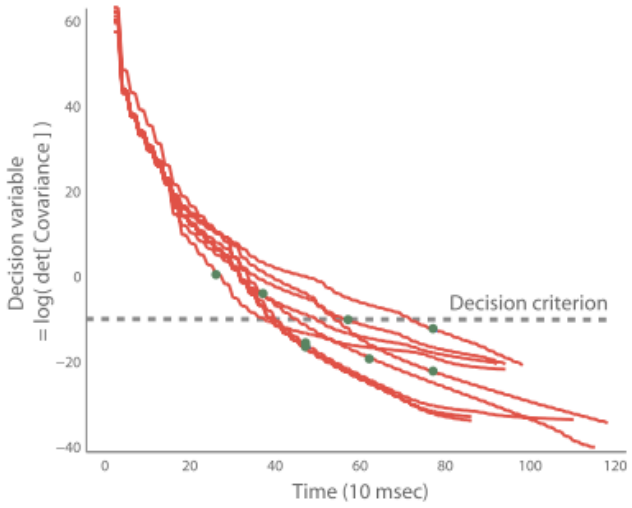

Figure 4: *Decision variables and ball apex.*

and make the sphere intercept with the ball trajectory which is relatively accurate. Besides the best position to catch, we also need to find the best time to catch for catching to late or too early can result in mistakes. As observed in the simulation result in the original paper, we can see that most decisions are made around the apex of the ball trajectory as shown in Fig. 4 where the red lines show the decision variable and green dots show the apexes of the ball trajectories. And the doted line is the decision criterion. And the authors also explained the reason intuitively: since the velocity of the ball reaches its minimum at the apex, and the gaze usually catches up with the ball by saccades and start to pursue, the quality of the position and velocity estimation is considerably improved around the apex. They also gave an example that skilled jugglers only look at the juggled balls at their apex.

At this point, we can divide the hand movements to two phases: reactive phase and proactive phase. In reactive phase, we set the destination 30° lower with respect to the gaze that corresponds to the captured configuration which we name as the ready pose, p_{ready}. The base vector in Eq.(2) is shown as follows:

$$b = R_e p_{ready} + p_e = \tilde{p}_m \tag{6}$$

where R_e and p_e are respectively the orientation and position of the eye frame, and $\tilde{p}_m$ is the destination of the previous submovement.

In proactive phase, the interception point and remaining time is determined if the predicted ball trajectory penetrates the sphere of preferred distance. Otherwise, we make the hand move to the closet point on the predicted trajectory with a maximum 2m/s speed and a physical limit of reach. As is described in the paper, we also use 90 degrees angle of attact between the ball and the palm and grasping happens in the last submovement when the decision of catch is made.

Body movements Body movements are solved still using the simple psudo-inverse of Jacobian to implement the IK model. The poisition of head is fixed on the torso, and the position of feet is also fixed to the ground. We solve the angles for the torso and arm joint angles according to the specified hand position and then solve for the torso and arm configuration. The average of these configurations is then blended with the previous timestep configuration and a weighted psudo-inverse of Jacobian is used to correct the resulting configuration so that the hand position matches the target.

The rotation of the hand is interpolated from its home orientation to a predetermined ready pose orientation with a set velocity and we make this velocity small at first half and faster and second half. And then we find the rotation that gives the nearest interpolation from the previous palm normal vector to estimated ball velocity vector and apply it to the hand.

5 Results

We use the $gerry$ model provided in the second assignment which has enough degree of freedom and fix the head position and feet position respectively to the torso and the ground. We implement the entire algorithm using Python (version 2.7.9) using $pykalman$ library and the software ran on a 3.2 GHz Intel 8-core E3-1225 v3 Xeon processors. We simulated the ball catching for several different trajectories and the character is able to catch the ball in almost all the generalized simple trials within limit. We didn't have time to test the special cases such as poor vision described in the original paper, but based on the principles of the unscented Kamlam filter, we believe this system will have realistic result for these scenarios.

6 Conclusion

We have implemented the system model proposed by Yeo et al. [Yeo et al. 2012] with several extensions. First, instead of using standard Kalman filter, we choose to use unscented Kalman filter which can give as more realistic simulating results using nonlinear models and maintaining same time complexity compared with standard Kalman filter. Second, we use a nonlinear model for the noise when estimating the ball trajectory instead of a simple linear Gaussian noise. Third, during implementation, we use python instead of matlab, but we believe these two software have similar performance library. Besides the above changes, there are also some minor changes during the process such as changing the speed of hand and time interval of gaze, etc which are all for the purpose of more realistic effect.

For the limitations from the original paper, we still haven't used a biochemical dynamic model, but choose to the same kinematic control as before for the purpose of efficiency. And the whole model is still highly simplified. But the system looks realistic and based on several principles and observation results, the system is practical and we believe it also generalizes well to different cases.

References

CRAWFORD, J., MARTINEZ-TRUJILLO, J., AND KLIER, E. 2003. Neural control of three-dimensional eye and head movements. *Current opinion in neurobiology 13*, 6, 655–662.

WOLPERT, D. M., GHAHRAMANI, Z., AND JORDAN, M. I. 1995. An Internal Model for Sensorimotor Integration. *Science 269*, 1880–1882.

YEO, S. H., LESMANA, M., NEOG, D. R., AND PAI, D. K. 2012. Eyecatch: Simulating visuomotor coordination for object interception. *ACM Trans. Graph. 31*, 4 (July), 42:1–42:10.