Bayes Filter for Gesture Recognition

Definitions

- Hidden State Space
  - $\mathcal{X}$: the set of m objects in the scene that can be referenced and a state for no item being referenced
  - An example would be: $x =$ Object 1 is being referenced
  - Each object in the scene has an associated set of three dimensional points (x_p) and set of descriptor keywords (x_d)
- Observations
  - $\mathcal{Z}$: the set of four-tuple $\{l, r, h, s\}$
  - l represents the observed origin (l_o) and vector (l_v) for the left arm
  - r represents the observed origin (r_o) and vector (r_v) for the right arm
  - h represents the observed origin (h_o) and vector (h_v) for head
  - s represents the observed speech from the user, consisting of a list of words
- Transition Function
  - $\mathcal{T}$: a function such that $\mathcal{T}(x_a, x_b)$ is equivalent to the probability that x_a transitions to x_b
  - **ST: Model with Poisson?**
- $\mathcal{N}$: a function that returns the probability of a sample under a Gaussian distribution given a mean and variance
  - Applied as $\mathcal{N}(\mu, \sigma, \text{sample})$
- Φ: a function that, given an origin and two points, returns the angle between the two points
  - Applied as $\Phi(\text{origin}, p_1, p_2)$
- $\mathcal{I}$: an indicator function applied as $\mathcal{I}($word, corpus that returns 1 if the word is in the corpus and 0 otherwise

- *len*: the number of items in a list

Equations

- Time Update
  - An equation used to determine the probability that $\mathcal{X}_t = x$ given only previous belief states
  - $P(\mathcal{X}_t = x|\mathcal{X}_{t-1}....\mathcal{X}_0) = \sum_{x' \in \mathcal{X}} \mathcal{T}(x, x') * bel(\mathcal{X}_{t-1} = x')$
  - This function computes the new probability for a state by multiplying the probability for each past state times transition probability and then summing all these new probabilities.
- Measurement Update
  - An equation used to determine the belief that $\mathcal{X}_t = x$ given observation $\mathcal{Z}_{t-1}$
  - $P(\mathcal{X}_t = x|\mathcal{Z}_{t-1}) = P(X_t = x|l_{t-1}) * P(X_t = x|r_{t-1}) * P(X_t = x|h_{t-1}) * P(X_t = x|s_{t-1})$
  - $P(\mathcal{X}_t = x|l_{t-1}) = [\prod_{p \in x_p} \mathcal{N}(\mu_l = 0, \sigma_l, \Phi(l_o, l_v, p))]^{(\frac{len(x_p)}{\sum_{x' \in \mathcal{X}} len(x'_p)})}$
  - $P(\mathcal{X}_t = x|r_{t-1}) = [\prod_{p \in x_p} \mathcal{N}(\mu_r = 0, \sigma_r, \Phi(r_o, r_v, p))]^{(\frac{len(x_p)}{\sum_{x' \in \mathcal{X}} len(x'_p)})}$
  - $P(\mathcal{X}_t = x|h_{t-1}) = [\prod_{p \in x_p} \mathcal{N}(\mu_h = 0, \sigma_h, \Phi(h_o, h_v, p))]^{(\frac{len(x_p)}{\sum_{x' \in \mathcal{X}} len(x'_p)})}$
  - These three equations are all very similar. They basically compute the probability of each point being seen given the vector of the gesture and multiply this out for all points. Finally, it is raised to a power equal to the ratio of points in the given object to the total number of points to normalize so that small objects don't have much larger probabilities.
  - $P(\mathcal{X}_t = x|s_{t-1}) = \frac{\sum_{w \in s_{t-1}} \mathcal{I}(w, x_d)}{\sum_{x \in \mathcal{X}} \sum_{w \in s_{t-1}} \mathcal{I}(w, x'_d)}$

  - This simply computes the ratio of words in the spoken phrase that are in the descriptors of the object to the number of the words that match any object

- Belief Update
  - An equation that produces the probability that $\mathcal{X}_t = x$ given all past observations and belief states, namely the product of the measurement and time updates
  - $bel(X_t = x) = P(\mathcal{X}_t = x|\mathcal{Z}_{t-1}) * P(\mathcal{X}_t = x|\mathcal{X}_{t-1}....\mathcal{X}_0)$
  - $bel(\mathcal{X}_0 = x) = \frac{1}{m+1}$ (uniform initialization of belief)