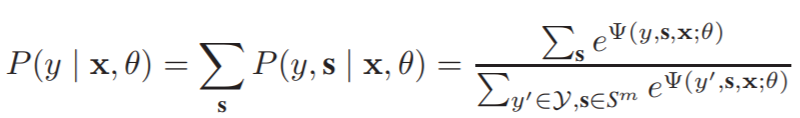
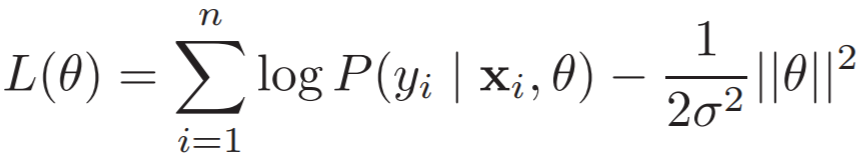
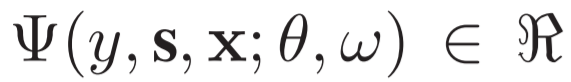
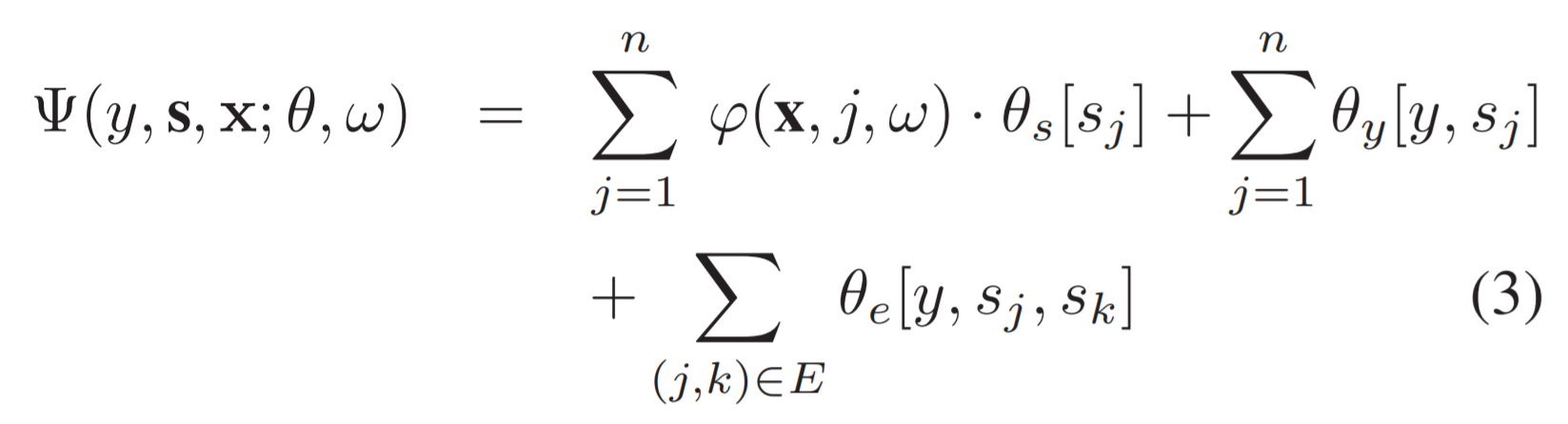
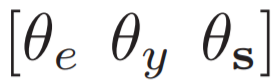
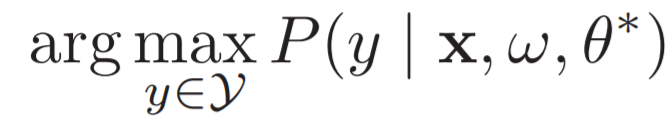
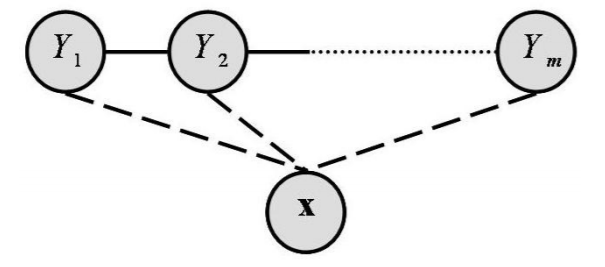
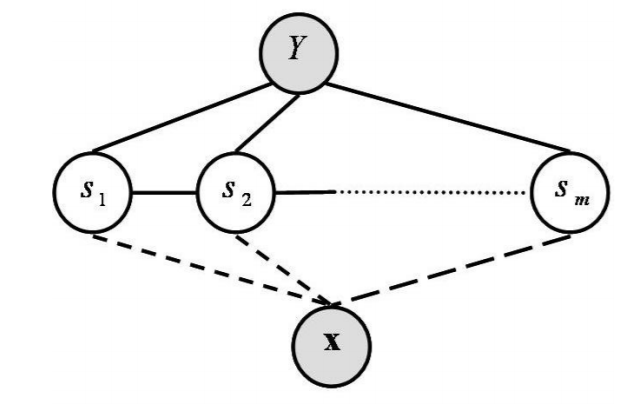
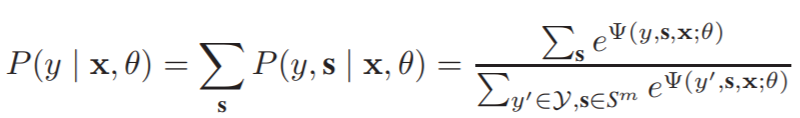
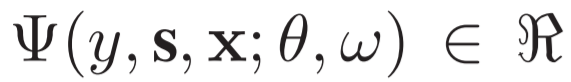
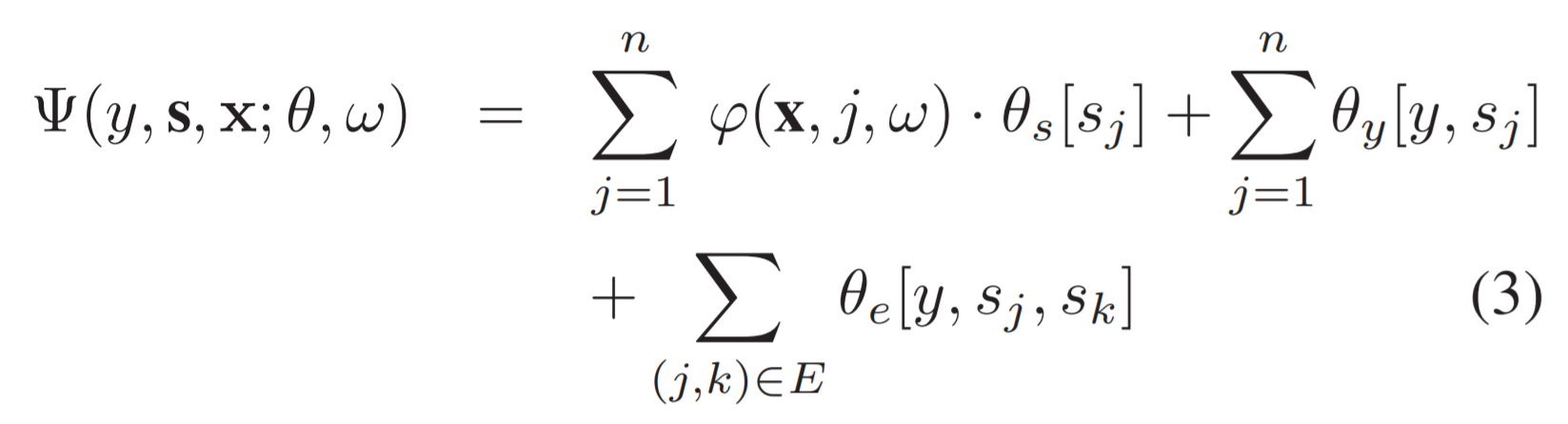
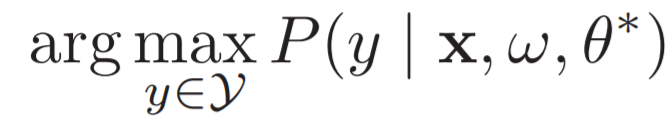
**Paper Review: “Hidden CRFs for Gesture Recognition**

* Gesture sequences often have a complex underlying structure
  + Models that can incorporate hidden structures have proven to be advantageous for recog tasks
* Most existing approaches to gesture recog with hidden states use HMM or variant to model gesture streams
* Significant limitation of these models is the requirement of conditional independence of observations
  + Additionally, hidden states in generative model are selected to maximize likelihood of generating all the examples of a given gesture class
  + Not necessarily optimal for discriminating class against other gestures
* Previous discrimination approaches to gesture sequence recognition have shown promising results but not incorporated hidden states nor addressed the problem of predicting the label of an entire sequence
* Paper discusses discriminative sequence model with hidden state structure and demonstrates utility in detection and multi-way classification
  + Evaluate on arm and hand gestures and compare performance to generative hidden state and discriminative fully observable models
* Automatic gesture recognition applicable for many applications in computer vision and pattern recognition
  + Gestures from head and arm often subtle, can happen at various time scales, and may exhibit long-range dependencies
* HMM and factored/coupled state models assume that observations are conditionally independent
  + Makes it difficult to accommodate long-range dependencies among observations or multiple overlapping features of observations
* CRFs use an exponential distribution to model entire sequence given observation sequence
  + Avoids independence assumption between observations
  + Allows non-local dependencies between state and observations
  + Markov assumption may still be enforced in state sequence, allowing inference to be performed efficiently using dynamic programming
* CRFs assign label for each observation (e.g. each time point in a sequence) and either capture hidden states nor directly provide a way to estimate conditional probability of class label of entire sequence
* Paper proposes model for gesture recognition which incorporates hidden state variables in a discriminative multi-class random field model and extends previous models for spatial CRFs into the temporal domain
* By allowing classification model with hidden states, no a-priori segmentation into substructures is needed and labels at individual observations are optimally combined to form class conditional estimation
* Hidden state CRF (HCRF) model can be used either as a gesture class detector (single class discriminatively trained against all other gestures) or multi-way gesture classifier (where discriminative models for multiple gestures are simultaneously trained)
* Latter approach has the potential to share useful hidden state structures across different classification tasks (allows higher recognition rates)
* Paper implements HCRF methods for arm and head gesture recognition
  + Compared performance against HMMs and fully-observable CRF techniques
* Advantages of MEMMs over HMM are that they can model arbitrary features of observation sequences and can therefore accommodate overlapping features
* Wish to learn a mapping of observations to class labels ‘y’ ϵ ‘Y’, where ‘x’ is a vector of ‘m’ local observations {x1, x2, …, xm} and each local observation ‘xj’ is represented by feature vector φ(xj) ϵ Rd
* HCRF models the conditional probability of class label given set of observations by:
* 
* Where ‘s’ = {s1, s2, …, sm}, each ‘si’ captures certain underlying structure of each class and ‘S’ is the set of all hidden states in the model
* If we assume ‘s’ is observed and that there is a single class label ‘y’ then the conditional probability of ‘s’ given ‘X’ becomes a regular CRF
* Potential function , parameterized by ‘ϴ’, measures compatibility between a label, a set of observations, and a configuration of the hidden states
* Use objective function in training parameters: 
* Where ‘n’ = total number of training sequences
* First term is the log likelihood, second is log of Gaussian prior with variance 
* Use gradient ascent to search for optimal parameter values: 
* Use Quasi-Netwon approximation technique from here
* HCRFs, being a discriminative model containing hidden states, are well-suited to problem of gesture recog
* Paper modifies original HCRF approach (where discriminative hidden state approach graphical model captured spatial dependencies between hidden object parts) to model sequences where the underlying graphical model captures temporal dependencies across frames and incorporates long-range dependencies
* To achieve the goal of distinguishing gesture classes, it learns a state distribution among different gesture classes in discriminative manner
  + Generative models can require a considerable number of observations for certain gesture classes
  + Generative models may also not learn a shared common structure among gesture classes, nor uncover the distinctive configuration that sets one gesture class uniquely against the others
* E.g. expect gesture classes with specific arm configurations to be easier to learn with discriminative model
* Also, would like a model that incorporates long-range dependencies (i.e. that the state at time ‘t’ can depend on observations that happened earlier or later in the sequence)
  + HCRF can learn discriminative state distribution and can be easily extended to incorporate long-range dependencies
* To incorporate long-range dependencies, modify the potential function ‘Ψ’ to include a window param ‘w’ that defines the amount of past and future history to be used when predicting state at time ‘t’
* Here,  defined as potential function parameterized by ϴ and ‘w’:
* 
  + Graph E is a chain where each node corresponds to a hidden state variable at time ‘t’
  +  is a vector that can include any feature of observation sequence within window ‘w’ (i.e. for window size of ‘w’, observation from ‘t-w’ to ‘t+w’ are used to compute features
  + Param vector ‘ϴ’ is made up of 3 components: : ϴs[sj] refers to params ϴs that correspond to state sj ϵ S, ϴy[y, sj] refers to params that correspond to class ‘y’ and state ‘sj’, and ϴe[y, sj, sk] refers to params that correspond to class ‘y’ and a pair of states ‘sj’ and ‘sk’
* Inner product (i.e. first of 3 terms in above equation) interpreted as a measure of compatibility between observation sequence and state at time ‘j’ and window size ‘w’
* Each param ϴy interpreted as a measure of the compatibility between hidden state ‘j’ and gesture ‘y’
* Each param ϴe measures the compatibility between pairs of consecutive states ‘j’ and ‘k’ and gesture ‘y’
* Given new test sequence ‘x’ and param vales ‘ϴ\*’ learned in training, take the label for sequence to be:
* 
* Since E is a chain, there are exact methods for inference and parameter estimation as both its objective function and its gradient can be written in terms of marginal distributions over state variables
  + Distributes can be computed via belief propagation
* Evaluation metric used for experiments here was percentage of sequences for which correct gesture label was predicted
* For head pose, used FFT on 3D angular velocities captured from pose tracking features as features for gesture recognition
* Total of 152 head nods, 11 head shakes, and 159 junk sequences were extracted based on ground truth labels obtained by interacting with a robot to obtain ground truths
  + Junk class had sequences that di not contain any nods or shakes
* Users performed six types of arm gestures in front of a stereo camera
  + For each image frame, 3D cylindrical body model was estimated using stereo-tracking algorithm
  + From these, both join angles and relative coordinates of the joints of the arms are used as observations for the experiments
  + 13 users performed these 6 gestures; average of 90 gestures per class were collected
*  
* CRF HCRF (multi-class)
* First, trained HMM w/ one model per class
  + 4 states with single Gaussian observation model
  + During evaluation, test sequences were passed through each model and the model with the highest likelihood was selected as the recognized gesture
* Second trained a single CRF chain model where every gesture class had a corresponding state
  + CRFs predict labels for each frame in sequence, not the entire sequence
  + During evaluation, found the Viterbi path under CRF and assigned sequence label based on most frequent occurring gesture label per frame
  + Ran additional experiments that incorporated long-range dependencies (different window sizes ‘w’)
* HCRF next used (one vs all)
  + For each gesture class, trained separate HCRF to discriminate from other classes
  + Each trained with 6 hidden states
  + For given test sequence, compared probs for single HCRF and highest scoring HCRF model is selected as the recognized gesture
* HCRF (multi-class) then used
  + Single HCRF trained using 12 hidden states
  + Test sequences were run with this model and gesture class with highest probability was selected as recognized gesture
  + Also conducted with different long-range dependencies as described above
* Number of hidden states for models set by minimizing the error on training data
* CRF models for arm and head dataset took ~200 iterations to train
  + HCRF models for the same took ~300-400 iterations to train
* Multi-class HCRF model performs better than HMM and CRF models at window size of ‘0’ (CRF slightly better than HMMs for head gesture and CRF even better with increased window sizes)
* HCRF multi-class significantly improved with larger window size, which indicates that incorporating long-range dependencies was useful
* For arm gesture recognition, CRF performed better than HMs at window size of 0, but at ‘w’=1 CRF performance was poorer, possibly due to overfitting when training the CRF model parameters
* For arm, both multi-class and one-vs-all HCRFs perform better than HMMs and CRFs
* Most significant improvement in performance was obtained when using a multi-class HCRF
  + Suggests it is important to jointly learn the best discriminative structure
* For multi-class in arm gesture, model found a unique distribution of hidden states for each gesture
  + Significant amount of state sharing among gesture classes
* Body poses that are visually more unique for a gesture class are assigned very distinct hidden states
  + Body poses common between different gesture classes are assigned the same states
* Accuracy on 3 arm gestures use multi-class increases from 86% to 97% to 98% by increasing w=0 to w=1 and w=2, respectively (i.e. more long-range dependencies incorporated)
  + Clear that incorporating some amount of contextual dependency is important
* HCRFs combine ability of CRFs for long-range dependencies w/ HMMs modelling of latent structures

**Significant Points and Takeaways from Paper**

* Most existing approaches to gesture recog with hidden states use HMM or variant to model gesture streams
* Significant limitation of these models is the requirement of conditional independence of observations
* Hidden states in generative model are selected to maximize likelihood of generating all the examples of a given gesture class, which is not necessarily optimal for discriminating class against other gestures
* Previous discrimination approaches to gesture sequence recognition have shown promising results but not incorporated hidden states nor addressed problem of predicting label of an entire sequence
* Using HMMs Makes it difficult to accommodate long-range dependencies among observations or multiple overlapping features of observations
* CRFs assign label for each observation (e.g. each time point in a sequence) and neither capture hidden states nor directly provide a way to estimate conditional probability of class label of entire sequence
* Paper proposes model for gesture recognition which incorporates hidden state variables in a discriminative multi-class random field model and extends previous models for spatial CRFs into the temporal domain
* HCRF models the conditional probability of class label given set of observations by:
* 
* Where ‘s’ = {s1, s2, …, sm}, each ‘si’ captures certain underlying structure of each class and ‘S’ is the set of all hidden states in the model
* Potential function , parameterized by ‘ϴ’, measures compatibility between a label, a set of observations, and a configuration of the hidden states
* To incorporate long-range dependencies, modify the potential function ‘Ψ’ to include a window param ‘w’ that defines the amount of past and future history to be used when predicting state at time ‘t’
* Here,  defined as potential function parameterized by ϴ and ‘w’:
* 
* Inner product (i.e. first of 3 terms in above equation) interpreted as a measure of compatibility between observation sequence and state at time ‘j’ and window size ‘w’
* Each param ϴy interpreted as a measure of the compatibility between hidden state ‘j’ and gesture ‘y’
* Each param ϴe measures the compatibility between pairs of consecutive states ‘j’ and ‘k’ and gesture ‘y’
* Given new test sequence ‘x’ and param vales ‘ϴ\*’ learned in training, take the label for sequence to be:
* 
* Number of hidden states for models set by minimizing the error on training data
* HCRF multi-class significantly improved with larger window size, which indicates that incorporating long-range dependencies was useful
* HCRFs combine ability of CRFs for long-range dependencies w/ HMMs modelling of latent structures